

Appropriate Evaluation of Diagnostic Utility of Machine Learning Algorithm Generated Images

Young Joon (Fred) Kwon
Danielle Toussie
Lea Azour
Jose Concepcion
Corey Eber
G Anthony Reina
Ping Tak Peter Tang
Amish H Doshi
Eric K Oermann
Anthony B Costa

FRED.KWON@ICAHN.MSSM.EDU
DANIELLE.TOUSSIE@MOUNTSINAI.ORG
LEA.AZOUR@NYULANGONE.ORG
JOSE.CONCEPCION@MOUNTSINAI.ORG
COREY.EBER@MOUNTSINAI.ORG
G.ANTHONY.REINA@INTEL.COM
PINGTAKTANG@GMAIL.COM
AMISH.DOSHI@MOUNTSINAI.ORG
ERIC.OERMANN@NYULANGONE.ORG
ANTHONY.COSTA@MSSM.EDU

Editors: Emily Alsentzer[⊗], Matthew B. A. McDermott[⊗], Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy[†], Stephanie L. Hyland[‡]

Abstract

Generative machine learning (ML) methods can reduce time, cost, and radiation associated with medical image acquisition, compression, or generation techniques. While quantitative metrics are commonly used in the evaluation of ML generated images, it is unknown how well these quantitative metrics relate to the diagnostic utility of images. Here, fellowship-trained radiologists provided diagnoses and qualitative evaluations on chest radiographs reconstructed from the current standard JPEG2000 or variational autoencoder (VAE) techniques. Cohen’s kappa coefficient measured the agreement of diagnoses based on different reconstructions. Methods that produced similar Fréchet inception distance (FID) showed similar diagnostic performances. Thus in place of time-intensive expert radiologist verification, an appropriate target FID – an objective quantitative metric – can evaluate the clinical utility of ML generated medical images.

Keywords: Generative Models, Data Compression, Clinical Validation

1. Introduction

Medical imaging is vital for the diagnosis and management of a wide range of diseases. However, due to multiple considerations such as time, cost, and radiation dose required for image acquisition, certain imaging modalities may not be available for certain patients. Machine learning (ML) techniques have been proposed as a means of addressing these shortcomings and generating medical images ([Lundervold and Lundervold, 2019](#)). A famous example of this currently in use is compressed sensing magnetic resonance imaging (CS-MRI), which decreases the time of a scan, reduces burden on the patient, and can minimize motion artifacts and undesired contrast washout ([Yang et al., 2016, 2018](#)). Generative adversarial networks (GANs) have generated synthetic computed tomography (CT) images from MRI images, eliminating unnecessary

radiation exposure to the patient and simplifying radiation treatment planning (Lei et al., 2019; Nie et al., 2018). Autoencoders compress medical data into dimension reduced latent vectors and can ease the transfer and storage requirements of large hospital systems that acquire thousands of images every day (Theis et al., 2017). The decoder of autoencoders generate an image output that restores the original image from the compressed latent vectors.

Successful correlation of quantitative metrics to diagnostic performance can reduce the time to translation of machine learning techniques and guide more faithful reconstruction of target modalities. Establishing an objective evaluation metric for diagnostic utility can alleviate the time and volume limitations of expert radiologist evaluations, which often form the bottleneck of validation experiments. While expert radiologist evaluations are still vital and cannot be replaced, quantitative metrics can decrease the time to develop generative ML algorithms to the clinics by speeding up the prototyping process.

In this paper, we compute diagnostic performance based on various quantitative metrics. We used a multi-level, vector-quantized variational autoencoder (VQ-VAE-2) approach to create compressed chest radiographs (CXR) of variable compression performance and image qualities, as evaluated by quantitative metrics, in relation to the original source image (Razavi et al., 2019). We compare this method to the current standard of image compression, JPEG2000, and assess the diagnostic utility of generated images in a simulated radiology workflow (Liu et al., 2017).

2. Related Works

Quantitative metrics have been developed to objectively evaluate the similarity of simulated or generated images to real-world im-

ages. Peak signal to noise ratio (PSNR), derived from the root mean square error (RMSE), and structural similarity index measure (SSIM) have been used to estimate qualitative metrics of diagnostic quality, but not diagnostic performance, of medical images from new methods to those from the original standard of care method (Mason et al., 2020). A more recent metric, Fréchet inception distance (FID), can provide an objective measure of how closely the semantic structure of generated images resemble those from target acquisition modalities (Heusel et al., 2017; Lei et al., 2019; Nie et al., 2018). While these quantitative metrics can guide development of ML algorithms, review of generated images by expert radiologists is necessary for successful and safe integration of these techniques to the clinical workflow. Mason et al. have compared objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR Images, but radiologists graded the subjective image quality instead of making any diagnoses from the MR images (Mason et al., 2020). To our knowledge, no studies have been performed to directly correlate radiologist's performance on actual classes of diagnoses on various ML generated images to these quantitative metrics.

3. Materials and Methods

3.1. Datasets

We used the CheXpert (<https://stanfordmlgroup.github.io/competitions/chexpert/>) dataset to train our model and MIMIC-CXR (<https://doi.org/10.13026/C2JT1Q>) dataset to externally test our model (Irvin et al., 2019; Johnson et al., 2019). CheXpert and MIMIC-CXR have a large number of radiographs (224,316 and 377,100, respectively) and contain both frontal and lateral views. For the training set, the

frontal view radiographs from CheXpert were center cropped and downsampled to a 256x256 pixel resolution ($n=191,027$). All preprocessed images were stored in an HDF5 dataset to improve input/output (I/O) performance in the data loader (Folk et al., 2011). For the test set, the frontal view radiographs of the previously unseen MIMIC-CXR dataset were center cropped and sampled at a higher 1024x1024 pixel resolution ($n=1759$).

3.2. Compression Architecture

We used the JPEG2000 compression algorithm configured to the standard currently used in a hospital system based in New York City. We compared this to our models created from the two-level vector quantized, variational autoencoder (VQ-VAE-2) architecture (Figure 1) (Razavi et al., 2019). 2D convolutions were of filter size 4, stride 2, and padding 1 (i.e., each dimension was isometrically reduced by 50% each time). We created two VQ-VAE models of varying levels of compression , standard compression and high compression, by changing the number of 2D convolutions at each level of encoding (Figure 2). The number of channels of the encoder output was set to 64 (i.e. for each integer index/key, there are 64 values) for the image reconstruction task. The loss function for VQ-VAE-2 consists of reconstruction loss, codebook loss, and commitment loss, further explained in Razavi et al. (2019)'s paper.

3.3. Expert Evaluation of Radiographs

We recruited three board-certified radiologists. Two had completed their thoracic radiology fellowship training, and one was in the final year of thoracic radiology fellowship.

We first selected 20 radiographs and reconstructed them using the JPEG2000 and

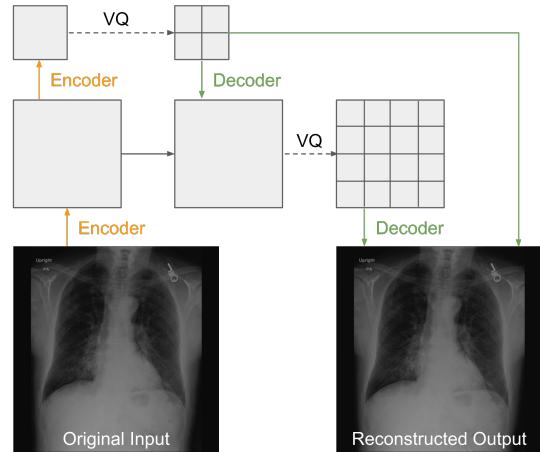


Figure 1: Multi-level vector-quantized variational autoencoder (VQ-VAE).
VQ = vector quantization layer.

two VQ-VAE models. Radiologists were provided with all three reconstructions of the same radiograph, blinded to the reconstruction method. For each set of 3 reconstructions for a given radiograph, radiologists identified and grouped reconstructions that they thought were indistinguishable from one another (i.e. they would make the same diagnosis based on each reconstructions; Figure 3).

We selected 115 radiographs for evaluation by subspecialty trained thoracic radiologists. Five radiographs were duplicated to test for the internal consistency of radiologists. Each of the resulting 120 radiographs was compressed and reconstructed using JPEG2000, VQ-VAE-2 standard compression, and VQ-VAE-2 high compression. Each of the three radiologists was randomly assigned one of the three reconstructions for every radiograph. Then we measured inter-rater agreement between radiologists, each of whom provided a diagnosis based on the radiograph but reconstructed differently from either JPEG2000 or

Original	JPEG2000 (Current Standard)	VQ-VAE-2 Standard Compression (2 + 1 Convolutions)	VQ-VAE-2 High Compression (3 + 2 Convolutions)
Compression Ratio	15.4	38.4	180.7
Fréchet Inception Distance	2.13	2.28	9.08
Peak Signal to Noise Ratio	46.05	45.39	44.05
Structural Similarity Index	0.968	0.965	0.947

Figure 2: Sample images and quantitative metrics of tested compression and reconstruction methods. Both VQ-VAE models retain key diagnostic features such as the spinous processes within the trachea and the costophrenic angle as well as the letters “L 27” (mirrored), though high compression shows loss of spatial resolution that is vital for certain tasks such as identification of pneumothorax. Standard compression VQ-VAE and the current JPEG2000 standard used in clinics today are both qualitatively and quantitatively nearly identical.

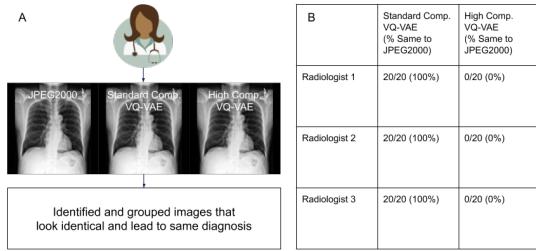


Figure 3: Qualitative review of the different reconstructions of the same radiograph.

standard or high compression VQ-VAE models Figure 4.

The radiologists were then asked to assess for the presence of the following pathological findings: 1) Lung Lesion, 2) Lung Opacity, 3) Pleural Effusion, 4) Pneumothorax, 5) Pulmonary or Interstitial Edema, 6) Fractures, 7) Medical Device, and 8) Normal Findings.

For each pathological class, radiologists provided a label of a definitive presence of, a definitive absence of, or an inability to determine the pathology. Again for each pathological class, radiologists graded the radiograph on the following Likert scale: (5) excellent diagnostic quality, (4) good diagnostic quality, (3) fair diagnostic quality, (2) poor diagnostic quality, (1) non-diagnostic.

3.4. Measurement of Internal Consistency of Radiologists

Five radiographs were duplicated to test for the internal consistency of radiologists. The original radiographs were included in the first half of the dataset, and duplicate radiographs were included in the second half of the dataset. The radiologists did not know that there were duplicate radiographs in the second half of the dataset. We evaluated the consistency of both the assessment

of presence or absence of eight pathological classes and their associated Likert scores based on a pair of first and second observations. Based on five radiographs, eight pathological classes, and three radiologists, we evaluated a total of 120 diagnostic assessments and associated Likert score pairs.

3.5. Statistical Analysis

Cohen’s kappa coefficient measured the inter-rater agreement of the diagnostic tasks between different compression methods (i.e. JPEG2000 vs standard compression VQ-VAE, JPEG2000 vs high compression VQ-VAE, or standard compression VQ-VAE vs high compression VQ-VAE). For each comparison between two reconstruction modalities, the Cohen’s kappa coefficient is the average value calculated from all individual diagnoses. The kappa coefficient weighed disagreements between definitive presence and definitive absence twice as much as the disagreement between definitive absence or presence and inability to determine. To calculate 95% confidence intervals for the kappa coefficient, we used bootstrapping experiments as previously described (DiCiccio and Efron; Hall et al., 2004; Boyd et al., 2013). We used an analysis of variance (ANOVA) to compute confirm or reject the null hypothesis that the distributions of Likert scores were the same among JPEG2000, standard compression VQ-VAE, and high compression VQ-VAE with significance set at $p = 0.05$. We then used Tukey’s Honestly Significant Difference (HSD) to find which Likert score populations were unequal and computed its significance level.

4. Results

4.1. Image Reconstruction

Two machine learning, autoencoder-based image reconstruction models had different

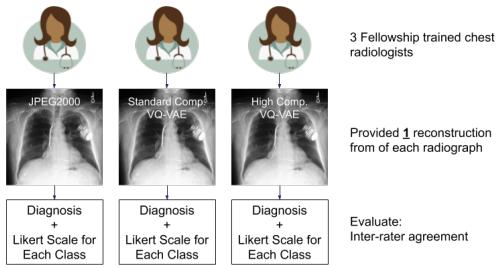


Figure 4: Clinical trial design to evaluate ML techniques of image reconstruction.

observed compression ratio and reconstruction quality. Standard compression VQ-VAE had a similar Fréchet inception distance (FID) as that of the current standard JPEG2000 algorithm (2.13 vs 2.28) but 2.5 times the compression ratio (38.4 vs 15.4). High compression VQ-VAE had additional convolutional layers that increased both the compression ratio (180.7) and the FID (9.08). Upon visual assessment by radiologists JPEG2000 and standard compression VQ-VAE looked identical, but high compression VQ-VAE had visible decreases in the spatial resolution (Figure 2). Sample images of each reconstruction method is provided in Appendix A. Given that the radiologists indicated that the images from JPEG2000 and standard compression VQ-VAE are indistinguishable, we proceeded to a double blinded simulated clinical trial to confirm the ability to deploy standard compression VQ-VAE to clinical practice.

4.2. Assessment of Internal Consistency of Radiologists

Each radiologist had five duplicate radiographs; each radiograph had eight pathological classes; three radiologists were recruited for our study. We evaluated a total of 120

pairs (5 radiographs x 8 classes x 3 radiologists) of diagnostic assessments and Likert scores for evaluation of internal consistency. All three radiologists demonstrated 100% internal consistency and gave the same diagnostic assessment for all pathological classes for both the first and second observations. Of the 120 Likert score pairs from the first and second observation of the duplicate radiographs, only 6 values differed, and never more than a Likert score by 1 (Appendix B).

4.3. Agreement of Radiologist Diagnostic Assessment

Radiologists, when provided all three reconstructions from the compression methods tested, concluded that they will have made the same diagnostic assessment for all pairs of JPEG2000 and standard compression VQ-VAE (Figure 3).

The average weighted Cohen's kappa coefficient of diagnoses from JPEG2000 and standard compression VQ-VAE was 0.608 (95% confidence interval: [0.569, 0.647]; Table 1). The weighted Cohen's kappa coefficients between JPEG2000 and high compression VQ-VAE was 0.525 (95% confidence interval: [0.486, 0.564]), similar to the weighted Cohen's kappa coefficients between high compression VQ-VAE and low compression VQ-VAE was 0.517 (95% confidence interval: [0.448, 0.556]).

4.4. Agreement of Radiologist Evaluation of Diagnostic Quality

The average Likert score for each pathological class for the three tested compression algorithms are presented in Table 2. A one-way ANOVA showed that not all Likert score populations from all three compression methods were the same ($p < 0.001$). In the subgroup analysis, JPEG2000 and standard compression VQ-VAE had the same distribution of Likert scores (Tukey HSD test $p >$

Cohen’s Kappa Coefficient	Standard Compression VQ-VAE	High Compression VQ-VAE
JPEG2000 (Current Standard)	0.608 [0.569, 0.647]	0.525 [0.486, 0.564]
Standard Compression VQ-VAE	—	0.517 [0.448, 0.556]

Table 1: Measurement of inter-rater agreement of diagnostic assessment between radiologists, each of whom provided diagnosis based on the radiograph but reconstructed differently from JPEG2000 or standard or high compression VQ-VAE. The inter-rater agreement between the radiologists who provided diagnoses on JPEG2000 and Standard Compression VQ-VAE, despite the different reconstruction technique, was within the substantial agreement range and similar to previously reported inter-rater agreements for diagnoses from chest radiographs. The interval indicates 95% confidence interval.

0.99), whereas JPEG2000 and high compression VQ-VAE had different distributions of Likert scores (Tukey HSD test $p < 0.001$). As expected, medical devices with high contrast or pleural effusions that are easy to diagnose on CXR had high Likert scores. Pneumothorax and fractures, pathological classes that require high spatial resolution to make a definitive diagnosis, showed the greatest decrease in Likert score and percent agreement of diagnosis from standard compression to high compression VQ-VAE. Lung lesion likely showed the lowest percent diagnosis agreement due to variable interpretation by radiologists (e.g. inclusion vs exclusion of benign, non-cancerous mass).

5. Discussion

Previously, quantitative metrics and qualitative evaluations have directed machine learning techniques that generate medical images. However, few studies have assessed inter-rater consistency in clinical diagnoses rendered based on the ML-generated images. We presented three clinical subspecialty trained radiologists with ML generated images of variable compression performance to demonstrate non-inferiority of a new ML

technique with a similar quantitative metric. We acquired quantitative metrics and correlated them to the inter-rater agreement for the diagnosis classes to inform target metrics for future medical image generative models and minimize the need for time intensive, volume limited review by expert radiologists.

We first confirm previous studies that radiologists have remarkable intra-rater consistency (Bonnyman et al., 2012; Neuman et al., 2012; Smith et al., 2013; Thelle et al., 2015). All three radiologists gave the same diagnostic assessment for duplicate radiographs (120 pairs of diagnostic decisions in total; Appendix B). The Likert score that evaluated qualitative diagnostic quality only differed on 6 of 120 pairs of measurements, and never by more than 1 score.

Although the inter-rater consistency was lower than the intra-rater consistency, the agreement between JPEG2000 and standard compression VQ-VAE was “substantial” based on the originally published guideline of the Cohen’s kappa coefficient (Cohen, 1960). This level of substantial agreement is on par with the previous published agreements between board certified radiologists and within the acceptable level of difference between clinicians (Rajpurkar et al.,

Pathology	JPEG2000 (Current Standard)	Standard Comp. VQ-VAE	High Comp. VQ-VAE
No Finding	4.867	4.891	4.008
Lung Lesion	4.900	4.941	3.875
Edema	4.858	4.808	3.733
Lung Opacity	4.942	4.958	4.025
Pleural Effusion	4.892	4.933	4.267
Pneumothorax	4.783	4.791	3.150
Fractures	4.958	4.867	3.533
Medical Device	4.992	5.000	4.775
Total	4.899	4.899	3.921
p-value (vs JPEG2000)	—	p > 0.99	p < 0.001

Table 2: Likert scale for each requested diagnostic task and the percent agreement between the diagnoses from JPEG2000 and either standard or high compression VQ-VAE. Likert scale is nearly indistinguishable between JPEG2000 and standard compression VQ-VAE. P-value from Tukey Honestly Significant Difference (HSD) test after ANOVA showed that not all Likert scores are the same (p < 0.001)

2018). Inter-rater consistency was not as reliable as intra-rater consistency, but 2D chest radiographs (CXRs) is a less precise modality than computed tomography (CT) and thus can lead to greater variability of what each radiologist requires as a threshold for diagnosis of certain findings.

Overall, the Cohen’s kappa coefficient was substantial and within the previously reported range of inter-rater variability when comparing diagnostic assessments from compression-reconstruction modalities of nearly identical FID scores (JPEG2000 vs standard compression VQ-VAE). High compression VQ-VAE had increased the number of convolutions, which increased compression but decreased the spatial resolution of the reconstructed images. The decreased spatial resolution is reflected by the increased FID score compared to JPEG2000 or standard compression, and the statistically significant drop in the diagnostic performance as measured by the decrease in the Cohen’s kappa coefficient was nearly identical when

comparing JPEG2000 or standard compression to high compression VQ-VAE. This not only confirmed that increased FID score correlates to poorer diagnostic performance, but also showed that JPEG2000 and standard compression created qualitatively indistinguishable reconstructions with similar diagnostic utility.

Various quantitative metrics (e.g. PSNR, SSIM, and FID) were nearly identical between the current JPEG2000 standard and standard compression VQ-VAE. The difference in magnitude of quantitative metrics was greatest in FID (2.1-2.2 for JPEG2000 and standard compression, 9.1 for high compression VQ-VAE) compared to other metrics that are widely used PSNR (45-46 for JPEG2000 and standard compression vs 44 for high compression VQ-VAE) and SSIM (0.965-0.968 for JPEG2000 and standard compression vs 0.947 for high compression VQ-VAE). High compression VQ-VAE still had high PSNR and SSIM values despite the demonstrated decrease in the diagnos-

tic quality, and thus PSNR and SSIM values should be used with caution. Especially for generative models that attempt to simulate an image that exists (e.g. original vs compressed radiographs, fast acquired vs traditionally acquired MRI, synthetic vs actual CT scans), computing the FID between generated images and target images is more appropriate to quantify the dependability of the generative task.

We also demonstrate that the target FID score may change depending on the diagnostic task. Diagnostic tasks that are sensitive to spatial resolution, such as determination of pneumothorax, demonstrated a decrease in the Likert score when FID score increased (Table 2). On the other hand, the increase in FID from 2 to 9 did not decrease the Likert score for this task significantly for determination of objects of high contrast (e.g. medical devices such as central lines; Table 2). That is, even though the radiographs may qualitatively look different (Appendix A), the diagnostic performance and the associated Likert scale assessment of diagnostic quality may not change for certain tasks. Therefore both quantitative metric and the diagnostic task of interest should be considered, as some pathological and anatomical patterns are more robust to image distortions than others.

Our study did not study how different methods of image distortions (e.g. injected black and pepper noise, blur, black rectangles, swirls, etc.) that generate the same increase in the FID score specifically affect the diagnostic performance (Heusel et al., 2017). We also lack the number of datasets of varying FID to quantitatively correlate the metric to their associated diagnostic performance. Nonetheless, as long as there exists a standard (e.g. currently used compression algorithm or existing CT scans for MRI to CT generation), ML methods can likely demonstrate non-inferiority to the current standard

based on similar FID scores. By comparing ML generated images to the existing standard during training, even non experts may identify obvious distortions that result in the same FID score or identify the clinical relevance of the distortion locations. Thus visual inspection must continue to be used in conjunction with the proposed quantitative metric.

6. Conclusion

We show that machine learning techniques that generate medical images with a low Fréchet inception distance (FID) as the target population or a similar FID as the current standard method likely retains the necessary and appropriate diagnostic information. In doing so, we validated a two-level, vector quantized variational autoencoder (VQ-VAE-2) that is a lightweight and fast method to compress medical images while preserving diagnostic utility. Depending on the required diagnostic task, some loss of image quality may be an acceptable trade-off to further increase compression or decrease acquisition time. Additional works that correlate quantitative metrics to diagnostic performance can lead to their use as an objective proxy measurement of diagnostic quality, which will result in faster development and deployment of generative machine learning (ML) techniques to clinic.

Acknowledgments

The authors thank PhysioNet and the Stanford ML Group for the MIMIC-CXR and the CheXpert datasets, respectively. The authors would like to acknowledge funding support from the RSNA Medical Student Research Grant and the T32 NIH T32 GM007280 Medical Scientist Training Program Grant. The data in this paper were used in a dissertation as partial fulfillment

of the requirements for a PhD degree at the Graduate School of Biomedical Sciences at Mount Sinai.

References

- Alison M Bonnyman, Colin E Webber, Paul W Stratford, and Norma J MacIntyre. Intrarater reliability of dual-energy x-ray absorptiometry-based measures of vertebral height in postmenopausal women. *J. Clin. Densitom.*, 15(4):405–412, October 2012.
- Kendrick Boyd, Kevin H Eng, and C David Page. Area under the Precision-Recall curve: Point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases*, pages 451–466. Springer Berlin Heidelberg, 2013.
- Jacob Cohen. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20(1):37–46, April 1960.
- Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals.
- Mike Folk, Gerd Heber, Quincey Koziol, Elena Pourmal, and Dana Robinson. An overview of the HDF5 technology suite and its applications. In *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases*, AD ’11, pages 36–47, New York, NY, USA, March 2011. Association for Computing Machinery.
- Peter Hall, Rob J Hyndman, and Yanan Fan. Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika*, 91(3):743–750, September 2004.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two Time-Scale update rule converge to a local nash equilibrium. June 2017.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A Mong, Safwan S Halabi, Jesse K Sandberg, Ricky Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. January 2019.
- Alistair E W Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-Ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*, 6(1):317, December 2019.
- Yang Lei, Joseph Harms, Tonghe Wang, Yingzi Liu, Hui-Kuo Shu, Ashesh B Jani, Walter J Curran, Hui Mao, Tian Liu, and Xiaofeng Yang. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. *Med. Phys.*, 46(8):3565–3581, August 2019.
- Feng Liu, Miguel Hernandez-Cabronero, Victor Sanchez, Michael W Marcellin, and Ali Bilgin. The current role of image compression standards in medical imaging. *Information*, 8(4):131, October 2017.
- Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Z. Med. Phys.*, 29(2):102–127, May 2019.
- Allister Mason, James Rioux, Sharon E Clarke, Andreu Costa, Matthias Schmidt, Valerie Keough, Thien Huynh, and Steven Beyea. Comparison of objective image quality metrics to expert radiologists’ scoring of diagnostic quality of MR images.

IEEE Trans. Med. Imaging, 39(4):1064–1072, April 2020.

Mark I Neuman, Edward Y Lee, Sarah Bixby, Stephanie Diperna, Jeffrey Hellinger, Richard Markowitz, Sabah Servaes, Michael C Monuteaux, and Samir S Shah. Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children. *J. Hosp. Med.*, 7(4):294–298, April 2012.

Dong Nie, Roger Trullo, Jun Lian, Li Wang, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with deep convolutional adversarial networks. *IEEE Trans. Biomed. Eng.*, 65(12):2720–2730, December 2018.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, Francis G Blankenberg, Jayne Seekins, Timothy J Amrhein, David A Mong, Safwan S Halabi, Evan J Zucker, Andrew Y Ng, and Matthew P Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.*, 15(11):e1002686, November 2018.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse High-Fidelity images with VQ-VAE-2. June 2019.

T O Smith, A Cogan, S Patel, M Shakokani, A P Toms, and S T Donell. The intra- and inter-rater reliability of x-ray radiological measurements for patellar instability. *Knee*, 20(2):133–138, March 2013.

Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. March 2017.

Andreas Thelle, Miriam Gjerdevik, Thomas Grydeland, Trude D Skorge, Tore Wentzel-Larsen, and Per S Bakke. Pneumothorax size measurements on digital chest radiographs: Intra- and inter- rater reliability. *Eur. J. Radiol.*, 84(10):2038–2043, October 2015.

Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, David Firmin, Jennifer Keegan, Greg Slabaugh, Simon Arridge, Xujiong Ye, Yike Guo, Simiao Yu, Fangde Liu, David Firmin, Pier Luigi Dragotti, Guang Yang, and Hao Dong. DAGAN: Deep De-Aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE Trans. Med. Imaging*, 37(6):1310–1321, June 2018.

Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Deep ADMM-Net for compressive sensing MRI. In D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 10–18. Curran Associates, Inc., 2016.

Appendix A. Sample Images from Three Tested Compression Algorithms



Figure 5: Current standard JPEG2000 compression and reconstruction.



Figure 6: Standard compression VQ-VAE and its reconstruction. Radiologists reported that JPEG2000 and standard compression VQ-VAE radiographs looked identical and would have made the same diagnoses based on either of the two reconstructions.



Figure 7: High Compression VQ-VAE and its reconstruction. Radiologists noted decreases in the spatial resolution, but the radiograph was not completely rendered nondiagnostic.

Appendix B. Internal Consistency of Radiologists

Reviewing Radiologist	% Agreement Diagnosis	% Agreement Likert Score
Radiologist 1	40/40 (100%)	38/40 (95.0%)
Radiologist 2	40/40 (100%)	37/40 (92.5%)
Radiologist 3	40/40 (100%)	38/40 (95.0%)

Table 3: In the set of 120 radiographs that each radiologist reviewed, 5 radiographs in the second half of the dataset were duplicates of the previously seen radiograph in the first half of the dataset. From the 5 radiographs, 8 diagnostic tasks and the associated Likert scores were evaluated for percent agreement. All radiologists had 100% agreement for all diagnostic criteria. Of the 6 Likert scores that changed, none changed by more than a difference of 1 Likert score: Two reviews changed from 5 to 4, two reviews from 3 to 2, and two reviews from 4 to 5.