10.1148/ryai.2020190035

Appendix E1

1. Supplementary Materials and Methods

The study protocol was approved by the ethical committee of the Geneva State, with a waiver concerning the requirement to obtain informed consent. The requirement for informed consent was waived because the chest radiographs (CXR) were part of a routine protocol, involving neither changes in patient clinical management nor additional diagnostic procedures.

1.1. Datasets

1.1.1. Training dataset.—

We retrospectively reviewed our institutional database for all consecutive patients who underwent upright frontal CXR between January and December 2017 using the radiography models of two different manufacturers (the most commonly employed in our institution): *Philips DigitalDiagnost* (later called 'DD') and *Siemens Fluorospot Compact FD* (later called 'FCFD'). Overall, 6,528 CXR were retrieved (3,264 CXR from each manufacturer), anonymized, resized to a resolution of 512 × 512 pixels (8 bits), and exported into PNG format. CXR were then preprocessed to remove symbols added at the image periphery after acquisition (eg, 'L' or 'R' signs) so as to avoid biases when training the generative adversarial network (GAN) model, as these acquisition symbols may be unbalanced between the DD and FCFD manufacturers, thus biasing the GAN model development. A Sobel filter (24) was applied to generate a region-of-interest (ROI) created around pixels with normalized gray levels above 0.95, designed to detect symbols and build an inpainting mask. The symbols were removed from the mask, and the missing surface was interpolated using the biharmonic method (25). To ensure that this preprocessing did not modify CXR beyond acquisition symbols, 10% of the training dataset was independently reviewed by two radiologists. Additionally, a new dataset of 50 CXR containing a pacemaker (not used elsewhere) were similarly preprocessed and reviewed by the same radiologists for preprocessing errors. None of the reviewed CXR exhibited errors in central regions of the image containing the patient's body.

1.1.2. Independent testing dataset #1: identification of manufacturer.—

We applied the GAN network after its development to upright frontal CXR acquired at our institution between July and August 2018, using either the DD or FCFD manufacturer. CXR were retrospectively retrieved from our institutional database, anonymized, resized, exported to PNG, and preprocessed as for the CXR of the training dataset. In total, 914 CXR were included in the testing set (457 CXR acquired using a DD manufacturer and 457 CXR acquired using a FCFD manufacturer, with about a quarter of CXR containing abnormal finding for each manufacturer). All CXR were reviewed by a radiologist for quality approval before being filtered by the GAN, to remove obviously "wrong" CXR (eg, XR incorrectly labeled as CXR in DICOM header, and not containing chest).

1.1.3 Independent testing dataset #2: diagnostic performance of RF.—

We applied our GAN network to a dataset of 200 upright frontal CXR from selected patient with and without congestive heart failure (CHF), acquired at our institution between January and March 2019 using DD or FCFD manufacturer and retrospectively retrieved. This independent testing dataset included 100 CXR from DD manufacturer (50% from patients with CHF), and 100 CXR from FCFD manufacturer (50% from patients with CHF). Patients with CHF were identified in our institutional database if a diagnosis code of CHF was included in their hospital discharge summary. All patients included were diagnosed with CHF based on clinical assessment, CXR findings and elevated serum B-type natriuretic peptide (at a cutoff of 100 ng/L as a marker of CHF, similar to Seah et al (2018) (26)), as part of regular patient diagnosis and management made during hospitalization. Only one CXR was taken by patients to avoid duplicates and bias in machine-learning classification. All CXR were also reviewed by a radiologist for quality approval before being filtered by the GAN. This second dataset was used to assess changes in CHF identification by ML classifier depending on the nature of RF (see "Material and Methods,") and not a comparison between radiologist versus RF classification performance.

1.2. Generative Adversarial Network (GAN) Model

1.2.1. GAN model development.—

Generative adversarial network (GAN) is a new kind of deep learning algorithm that has recently emerged (11,12). GANs are systems of neural networks contesting against each other in a given task. Briefly, a generator network tries to produce counterfeit images as close as possible to the authentic ones, while a discriminator network tries to differentiate the real images from fake ones produced by the generator network. With training, both networks are prone to become more efficient, with the generator network producing fake images very close to the real ones. Several GAN architectures have been developed to cover a wide range of applications, given that these networks can learn to mimic any data distribution kind. They have been used to transform images from one source domain to a target domain (13,14), such as generating maps from satellite pictures or counterfeit images from a renowned painter based on photography. A specific GAN type, called cycle-GAN, has been successfully applied in texture-translation from one source image to a target image (15).

We used a cycle-GAN model adapted from (15) to translate texture information from CXR acquired using one manufacturer to the other (eg, from DD to FCFD) and vice-versa (from FCFD to DD) (Fig 2). This model is based on two neural networks of the generator type: a generator receiving an original CXR from the DD set (native DD or nDD) and modifying it to match the FCFD type (fake FCFD or fFCFD) and another generator receiving an original FCFD image (native FCFD or nFCFD) and processing it to match the DD type (fake DD or fDD). These two generators depend on two other neural networks of the discriminator type: a discriminator trained to distinguish between native FCFD images from fake FCFD images, and another discriminator trained to distinguish native DD images from fake DD images (Fig 2). The generators are optimized based on the corresponding discriminator feedback and trained to mislead it (ie, produce fake images similar to the target ones).

We reconstructed the original images following texture-translation by passing it sequentially through the two generators (eg, a DD image is first translated to FCFD texture by

the first generator, then back to DD by the second generator). To preserve the cycle-consistency of the whole GAN, the generators were trained to minimize the absolute difference between the original image and reconstructed image (after passing through the two generators), using a reconstruction cost set as the L1 distance between the input and reconstructed images. Here, the cost was reinforced by adding a loss component to compare images at the level of small detail or texture (ie, high spatial frequency). This loss component was derived from a structural similarity index measure (SSIM) (27): Loss $(x, \hat{x}) = \varphi(1 - \text{SSIM}(x, \hat{x}))$, where $\hat{x}$ is the reconstructed version of x and φ is a parameter between 0 and 1, with a 11 × 11 box for the SSIM computation to focus textural details of the images. φ parameter is used to moderate the weight of the reconstruction loss compared with adversarial loss during backpropagation, and avoid small or large values favoring adversarial or reconstruction loss, respectively. Overall, SSIM measures the perceptual difference between two CXRs and quantifies image changes caused by texture-translation based on local gray level statistics.

The cycle-GAN is thus trained as follows: (i) the discriminator networks accurately recognize native CXR from each manufacturer (nDD and nFCFD); (ii) the generator networks translate CXR texture to mislead their respective discriminator network so it recognizes fake CXR as belonging to the target manufacturer and not to the original one (*e.g.;* fDD as belonging to DD manufacturer instead of its original FCFD one); (iii) the cycle-GAN must, additionally, maintain good cycle-consistency while accurately reconstructing fake CXR into their original image.

The whole model was trained on 472 × 472 patches of the input images that were cropped at random positions on the original 512 × 512 CXRs at each epoch. This means that two patches from the same image vary only at the image borders and catch the entire lung parenchyma, enabling regularization of the model at boundaries, with instantaneous random cropping at the level of the single training step. The GAN model was trained using 95% of the original training data, with the remaining 5% being discarded for validation purposes. The model was implemented in Python (Version 3.6.2) using TensorFlow (Version 1.3.0).

1.3. Evaluation of Texture-Transfer Using GAN

To assess the quality of the CXR texture-translation between the two manufacturers performed by the GAN model, we used two generator networks of the GAN to produce fDD and fFCFD images from CXR of the independent testing set. Thus, we obtained 2 × 457 fake CXR of each type (fDD and fFCFD), paired with 2 × 457 original CXR of each type (nFCFD and nDD, respectively).

As a global measure of the GAN model accuracy, the structural similarity index (SSIM) was computed for all images of the testing set as the general indicator of cycle-consistency of the GAN, along with its corresponding 95% CI. A SSIM of 1 implies a perfect reconstruction of the image back to its origin after passing through the GAN, with this value decreasing to-1 when the original and reconstructed images are opposite (ie, negative covariance), whereas random correlation provides values closer to 0. As the GAN translates texture, we expect changes between the manufacturers to occur mostly at high spatial frequencies, with the objects' global structures (eg, thorax shape) hardly altered.

1.3.1. Radiomics features extraction.—

We used an original and fake CXR of the independent testing set to compare the RF reproducibility before and after GAN-translation. To that end, we first automatically segmented lung parenchyma of each CXR, then extracted 92 RF from the lung parenchyma of each CXR, and finally compared the ability of the GAN texture-translation to reduce the RF difference between the two manufacturers (ie, DD and FCFD). The lung parenchyma segmentation was first performed by using an open-source pretrained deep-learning model (28). All segmentations were visually inspected by a senior resident in radiology (4 years of experience). RF were then extracted from the lung parenchyma using a sliding-window method with pyradiomics (Version 1.2.0 (16),). The 92 RF extracted are default features available in pyradiomics (ie, without image filtering), with the exception of shape features (16).

1.3.2. Radiomics features reproducibility.—

To test for a reduction in the intermanufacturer RF variability, before and after GAN texture-translation, we computed the concordance correlation coefficient (CCC), as defined by Lin (17), for each RF and each image type (nDD, nFCFD, fDD, and fFCFD). As the GAN should translate texture from one manufacturer to the other, we hypothesized that CCC would be increased when comparing one native CXR type with its opposite fake one (ie, nDD versus fDD; nFCFD versus fFCFD), as compared with the CCC between native images alone (ie, nDD versus nFCFD) or fake images alone (ie, fDD versus fFCFD). We have, thus, computed the CCC for each RF, such as between nDD and nFCFD, between nDD and fDD, and between nFCFD and fFCFD. Similar to (18), RF with CCC of 0.85 or greater were considered as reproducible. We also reported the percentage of RF with CCC ≥ ,0.80, ≥ 0.85, and ≥ 0.90 for each RF class, for nDD versus nFCFD, nDD versus fDD, and nFCFD versus fFCFD, similar to (18).

1.3.3. Machine-learning classification of a radiographic model.—

As the GAN is likely to reduce the intermanufacturer RF difference, we hypothesized that machine learning (ML) classifiers trained to recognize the manufacturer, as based on RF extracted from native CXR, would be misled when trying to identify the manufacturer of fake CXR. We, thus, trained five common ML classifiers to identify the manufacturer of native CXR, using the 92 RF previously extracted. We trained Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression (LogReg), and Random Forest (RFo) classifiers in *scikit-learn* (Version 0.19.0 (29);), with default parameters for each classifier. We then assessed the performance of these five ML classifiers in distinguishing the manufacturer for four comparisons using 10-fold cross-validation: 1) nDD versus nFCFD; 2) fDD versus fFCFD; 3) nDD versus fFCFD; 4) and fDD versus nFCFD. Correct manufacturer recognition was defined as the original manufacturer for native CXR, and as the target manufacturer for fake CXR (eg, FCFD class for original DD image translated to FCFD by the GAN). Thus, if ML classifiers identified fake CXR as belonging to the target manufacturer class, they would be misled by the GAN-texture-translation. We compared the accuracy of each ML model as to distinguishing nDD versus nFCFD, nDD versus fFCFD, fDD versus nFCFD, and fDD versus fFCFD.

1.3.4. Radiological classification of a radiographic model.—

Given that CXR characteristics are likely to depend on image acquisition and processing specific to each manufacturer, we hypothesized that experienced radiologists would accurately

distinguish the manufacturer when reviewing native CXR *(i.e.*, not translated by the GAN). As the GAN architecture has been suggested to translate image characteristics from one type to the other (15), we additionally hypothesized that radiologists would recognize the fake CXR as belonging to the target manufacturer rather than to the original one (eg, DD image translated to FCFD manufacturer (ie, fFCFD) would be recognized as FCFD ones). Taken together, the radiologist would accurately identify the manufacturer of native CXR, yet be misled by the GAN translation in recognizing the manufacturer of fake CXR.

To test these hypotheses, we asked two radiologists to review native and fake CXR and identify the manufacturer of each one (DD versus FCFD). Most of the CXR showed no active disease, and radiologists were instructed to focus on identifying manufacturer only. The radiologists were first trained to recognize native DD and FCFD images (training bloc) by reviewing 240 CXR (four blocks of 60 CXR, with 50% of each manufacturer type, randomly taken from the testing set). They were then asked to recognize the manufacturer (DD versus FCFD) of 240 native and fake CXR (similarly taken from the testing set and nonoverlapping with those of the training bloc), in four counterbalanced blocs of 60 images (test blocs), with an equal proportion of each image type (nDD, nFCFD, fDD, and fFCFD). The two radiologists were board-certified radiologist (12 and 19 years of experience) working as attending thoracic radiologist at our University Hospital and were not involved in GAN model development. In these test blocs, the radiologists read CXR together to reach a consensus about manufacturer type and were blinded to the type and proportion of native and fake CXR. They were merely asked to identify the manufacturer type, regardless of whether the image was processed or not by the GAN. As for ML classifiers, correct manufacturer recognition was defined as the original manufacturer for native CXR and the target manufacturer for fake CXR. Indeed, if radiologists identified the fake CXR as belonging to the target manufacturer type, they would be misled by the GAN.

For practical reasons, it was not possible to cross-validate the accuracies of the radiologists. Therefore, confusion matrices and permutation testing were applied to ensure our results were not observed by chance. Permutation testing was performed for each above-mentioned comparison by randomly permuting 10,000 CXR labels and computing a classification score for each iteration, thereby obtaining a permutation distribution. The *P* value returned from this permutation testing approximates the probability that a true classification score would be obtained by chance. As for ML classifiers, we compared the accuracy of radiologists as to distinguishing nDD versus nFCFD, nDD versus fFCFD, fDD versus nFCFD, and fDD versus fFCFD.

References

24. Nixon MS, Aguado AS. Feature Extraction and Image Processing for Computer Vision. 3rd ed. Oxford: Academic Press, 2012.

25. Chui CK, Mhaskar H. Mra contextual-recovery extension of smooth functions on manifolds. Appl Comput Harmon Anal 2010;28(1):104–113.

26. Seah JCY, Tang JSN, Kitchen A, Gaillard F, Dixon AF. Chest Radiographs in Congestive Heart Failure: Visualizing Neural Network Learning. Radiology 2019;290(2):514–522 https://doi.org/10.1148/radiol.2018180887.

27. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13(4):600–612.

28. Vitali Petsiuk AK. Lung-segmentation-2d. https://github.com/imlab-uiip/lung-segmentation-2d. Published 2017.

29. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825–2830. http://www.jmlr.org/papers/v12/pedregosa11a.html.

Table E1. Concordance Correlation Coefficient between manufacturer before and after texture-translation for all radiomic features

		before	after texture-translation	
RF class	RF name	nDD vs nFCFD	nDD vs fDD	nFCFD vs fFCFD
First-order	10Percentile	0.41	0.65	0.57
	90Percentile	0.35	0.68	0.52
	Energy	0.42	0.69	0.53
	Entropy	0.40	0.97	0.96
	InterquartileRange	0.23	0.98	0.97
	Kurtosis	0.44	0.98	0.98
	Maximum	0.33	0.72	0.27
	MeanAbsoluteDeviation	0.23	0.98	0.97
	Mean	0.37	0.65	0.59
	Median	0.35	0.65	0.58
	Minimum	0.42	0.66	0.50
	Range	0.35	0.98	0.97
	RobustMeanAbsoluteDeviation	0.22	0.98	0.97
	RootMeanSquared	0.36	0.65	0.58
	Skewness	0.38	0.97	0.97
	TotalEnergy	0.42	0.69	0.53
	Uniformity	0.41	0.95	0.95
	Variance	0.32	0.98	0.97
GLCM	Autocorrelation	0.37	0.98	0.97
	ClusterProminence	0.39	0.97	0.96
	ClusterShade	0.39	0.97	0.96
	ClusterTendency	0.31	0.98	0.97
	Contrast	0.26	0.95	0.98
	Correlation	0.31	0.81	0.88
	DifferenceAverage	0.15	0.92	0.95
	DifferenceEntropy	0.25	0.91	0.94
	DifferenceVariance	0.26	0.97	0.98
	Id	0.31	0.88	0.91
	Idm	0.32	0.88	0.91
	Idmn	0.36	0.81	0.87
	Idn	0.43	0.80	0.85
	Imc1	0.42	0.80	0.85
	Imc2	0.36	0.83	0.87
	InverseVariance	0.36	0.90	0.94
	JointAverage	0.43	0.98	0.97
	JointEnergy	0.34	0.89	0.90
	JointEntropy	0.38	0.95	0.96

	MaximumProbability	0.36	0.88	0.88
	SumAverage	0.43	0.98	0.97
	SumEntropy	0.40	0.97	0.96
	SumSquares	0.30	0.98	0.97
GLDM	DependenceEntropy	0.30	0.90	0.91
	DependenceNonUniformity	0.41	0.89	0.92
	DependenceNonUniformityNormalized	0.37	0.89	0.92
	DependenceVariance	0.35	0.84	0.85
	GrayLevelNonUniformity	0.37	0.96	0.96
	GrayLevelVariance	0.32	0.98	0.97
	HighGrayLevelEmphasis	0.37	0.98	0.97
	LargeDependenceEmphasis	0.32	0.85	0.86
	LargeDependenceHighGrayLevelEmphasis	0.30	0.92	0.91
	LargeDependenceLowGrayLevelEmphasis	0.17	0.84	0.83
	LowGrayLevelEmphasis	0.26	0.91	0.91
	SmallDependenceEmphasis	0.32	0.89	0.92
	SmallDependenceHighGrayLevelEmphasis	0.34	0.97	0.97
	SmallDependenceLowGrayLevelEmphasis	0.42	0.90	0.91
GLRLM	GrayLevelNonUniformity	0.22	0.96	0.95
	GrayLevelNonUniformityNormalized	0.41	0.96	0.95
	GrayLevelVariance	0.31	0.98	0.97
	HighGrayLevelRunEmphasis	0.38	0.98	0.97
	LongRunEmphasis	0.30	0.83	0.84
	LongRunHighGrayLevelEmphasis	0.15	0.96	0.95
	LongRunLowGrayLevelEmphasis	0.28	0.85	0.86
	LowGrayLevelRunEmphasis	0.31	0.91	0.92
	RunEntropy	0.24	0.92	0.92
	RunLengthNonUniformity	0.38	0.88	0.91
	RunLengthNonUniformityNormalized	0.33	0.87	0.90
	RunPercentage	0.32	0.86	0.88
	RunVariance	0.29	0.82	0.83
	ShortRunEmphasis	0.32	0.87	0.89
	ShortRunHighGrayLevelEmphasis	0.38	0.98	0.97
	ShortRunLowGrayLevelEmphasis	0.29	0.93	0.93
GLSZM	GrayLevelNonUniformity	0.47	0.83	0.87
	GrayLevelNonUniformityNormalized	0.38	0.96	0.95
	GrayLevelVariance	0.29	0.98	0.97
	HighGrayLevelZoneEmphasis	0.39	0.98	0.97
	LargeAreaEmphasis	0.38	0.73	0.74
	LargeAreaHighGrayLevelEmphasis	0.23	0.78	0.76
	LargeAreaLowGrayLevelEmphasis	0.36	0.68	0.70
	LowGrayLevelZoneEmphasis	0.38	0.92	0.93
	SizeZoneNonUniformity	0.33	0.89	0.93
	SizeZoneNonUniformityNormalized	0.24	0.88	0.92
	SmallAreaEmphasis	0.15	0.88	0.91
	SmallAreaHighGrayLevelEmphasis	0.39	0.98	0.97
	SmallAreaLowGrayLevelEmphasis	0.33	0.94	0.93
	ZoneEntropy	0.36	0.86	0.88
	ZonePercentage	0.31	0.89	0.92

	ZoneVariance	0.38	0.71	0.71
NGTDM	Busyness	0.37	0.93	0.92
	Coarseness	0.35	0.84	0.86
	Complexity	0.20	0.96	0.97
	Contrast	0.36	0.94	0.96
	Strength	0.33	0.97	0.94

The table displays the Concordance Correlation Coefficient of all radiomic features, for the comparison between manufacturers before and after texture-translation. nDD: native DD images, nFCFD: native FCFD images, fDD: fake DD images, fFCFD: fake FCFD images. First-order = first order features, GLCM = gray level co-occurrence matrix, GLRLM = gray level run length matrix, GLSZM = gray level size zone matrix, NGTDM = neighboring gray tone difference matrix, GLDM = gray level dependence matrix.

Table E2: Manufacturer classification accuracy by machine-learning algorithms and radiologists

A. Machine-Learning Classifier				
Classifier	**nDD vs nFCFD**	**fDD vs nFCFD**	**nDD vs fFCFD**	**fDD vs fFCFD**
SVM	97.8 ± 2.1	97.5 ± 1.6	97.0 ± 2.3	96.6 ± 2.0
LogReg	98.8 ± 1.2	98.0 ± 1.5	95.6 ± 2.3	95.9 ± 1.9
LDA	99.3 ± 0.6	99.1 ± 0.9	96.2 ± 2.3	95.9 ± 2.7
QDA	93.5 ± 3.6	93.8 ± 3.0	92.5 ± 4.4	92.9 ± 3.7
RFo	95.6 ± 1.9	96.3 ± 1.4	95.8 ± 2.2	96.5 ± 1.4
B. Radiologists				
Classifier	**nDD vs nFCFD**	**fDD vs nFCFD**	**nDD vs fFCFD**	**fDD vs fFCFD**
Radiologists	85.0	88.3	74.6	77.9

A. Manufacturer classification using several machine-learning (ML) classifiers. Each of the five ML classifiers is trained to identify the manufacturer (DD versus FCFD) using native images alone, based on radiomics features. These five ML classifiers are then tested using new and independent testing sets, with both native and fake CXR. Correct classification of native images (nDD and nFCFD) suggests the identification of the original manufacturer type (DD and FCFD, respectively). As ML classifiers accurately learn to identify the manufacturer using native images, if GAN correctly translates the texture from one manufacturer to the other, the correct identification of the manufacturer for fake images means the identification of the translated manufacturer type (eg, original FCFD image translated to DD [ie, fDD] would be recognized as DD type). Significance level testing is assessed by positive ranked Wilcoxon using 10-fold cross-validation. **B.** Manufacturer (DD or FCFD) classification by experienced radiologists. As for ML classifiers, correct identification of the manufacturer for native images (nDD and nFCFD) means the identification of the original constructor, while the correct identification of the manufacturer for fake images (fDD and fFCFD) means the identification of the translated manufacturer type. Significance level is derived from permutation testing. All accuracies are expressed in % and are statistically above chance ($P < .01$, for ML and radiologist classification). nDD = native DD, nFCFD = native FCFD, fDD = fake DD, f,CFD fake FCFD. SVM = support vector machine, LogReg = logistic regression, LDA = linear discriminant analysis, QDA = quadratic discriminant analysis, RFo = random forest.