IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 September 2022, accepted 19 October 2022, date of publication 25 October 2022, date of current version 31 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3216838

RESEARCH ARTICLE

Using Convolutional Neural Network to Automate ACR MRI Low-Contrast Detectability Test

JHONATA EMERICK RAMOS[1], HAE YONG KIM[1], AND FELIPE BRUNETTO TANCREDI[2]
[1]Department of Engineering Electronic Systems, Escola Politécnica, Universidade de São Paulo, São Paulo 05508-010, Brazil
[2]RadSquare Tecnologia Ltda, São Paulo 04020-040, Brazil

Corresponding author: Jhonata Emerick Ramos (jhonata.emerick@usp.br)

This work was supported in part by the Foundation for Research Support of the State of São Paulo (FAPESP) under Grant 2015/27022-0, and in part by the National Council for Scientific and Technological Development (CNPq) under Grant 305377/2018-3.

ABSTRACT According to the American College of Radiology (ACR), the performance of magnetic resonance imaging (MRI) scanners should be monitored using phantom images acquired weekly. These quality assurance images are usually analyzed by a technician, but automated analysis has been proposed to reduce costs and improve repeatability. Reports on the automation of low-contrast detectability tests are scarce, and none can completely replace human work. In previous works, we have demonstrated that machine learning methods can be used to learn the subtleties of image quality and the visual assessment of technicians. We showed that machines are able to mimic human perception quite accurately. In these works, we used hand-designed image quality features. In the present work, we use a deep learning method to automatically design appropriate image features. By training this network on a large base with visual assessments from multiple technicians, we show that the machine can be taught to assess MRI image quality better than any technician alone, justifying its widespread adoption. Our dataset contained 12,000 binary responses to the detectability of low-contrast structures (''holes''). We used the median of the technicians' responses as the gold standard. To increase statistical power, we repeated training and testing 5 times, using 5-fold cross-validation. We obtained a mean AUC (area under the ROC curve) of 0.983 ± 0.003. At the point of equal error rate, the mean accuracy, sensitivity and specificity were $93.2 \pm 0.7\%$, numbers higher than those achieved by any technician alone. Applying the obtained model to a completely independent test dataset with 10,800 structures, we obtained an AUC of 0.979. The predictions of our model in classifying spokes (sets of 3 holes) agree in 93.83% of the cases with the median of the responses of the technicians. These results again are better than the responses of any individual technician. We conclude that the ACR test can be performed by a machine with greater reliability than individual technicians.

INDEX TERMS Artificial intelligence, convolutional neural network, magnetic resonance imaging, machine learning, quality assurance, American College of Radiology.

I. INTRODUCTION

Magnetic resonance imaging is a non-invasive method that generates 2-D or 3-D images of the anatomy or physiological processes of the body [1]. MRI offers a huge range of image contrasts without using contrast agents or ionizing radiation. The MRI scanner can be programmed to produce images that reveal fractures in bones, as X-ray, but can also be tuned to reveal the differences between muscle and fat; or to detect structures with tenuous compositional differences in relation to their surroundings. This ability of MRI tomography to distinguish small structures with low contrast makes it particularly useful in the radiological evaluation of meniscus tears, myocardial infarctions, prostate cancer, endometriosis, to name a few examples.

Like other medical instruments, the MRI scanner must be routinely subjected to quality assessment to ensure that the device is imaging within its specifications and that it meets

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Mei.

quality standards, such as those recommended by the ACR. The ACR has an extensive quality control program and issues certificates of adequacy in all medical imaging modalities, including MRI. In the United States, the ACR tests are part of national regulatory rules. In the rest of the world, they are adopted by institutions that recognize the importance of monitoring the quality of the radiological images they produce and adopt the ACR tests as part of good practices.

The ACR recommends that the performance of MRI scanners be monitored by repeating image quality tests every 7 days or less. Deviations in quality scores indicate that the clinical images generated by the scanner may be compromised and that it needs calibration or maintenance. Quality tests are performed on images of an object of known geometry and composition called phantom and include measurements of distortion, contrast and resolution. A good quality image should depict the anatomy under inspection with the correct dimensions and features, and allow the detection of small structures under low contrast conditions.

ACR tests that are based on direct measurements are objective, tend to be consensual, and their automation can be performed through simple image processing strategies. On the other hand, low and high contrast tests rely entirely on the operator's visual perception, and their automation has been a challenge. In these tests, the operator must indicate whether a given set of structures in the phantom can be resolved (that is, distinguished from the background) in the image. These tests involve a very subjective assessment, as they are a direct reflection of human visual perception and image manipulation techniques that vary between operators, making their automation quite challenging. If these two tests could be automated, probably the entire ACR test could be performed without the presence of an experienced operator, reducing costs and improving repeatability.

The high contrast test is less sensitive and human responses are almost consensual. However, human assessments on the low-contrast resolution test may disagree somewhat. This test is what allows you to monitor the scanner's performance in generating contrasted images of soft tissues, a hallmark of the imaging modality.

Our group has investigated new methods for automating the low-contrast resolution test of the ACR program. In previous works [2], we extracted manually-designed features from the test images and used them to feed conventional machine learning algorithms. The results were encouraging; however, the accuracy of the predicted values did not allow the complete replacement of the operator by the method. In the present work, we investigate an alternative to automate the test. It is well-known that convolutional neural network (CNN) can automatically design appropriate low-level filters to extract the fittest features that can best detect and classify objects in an image. We evaluate the performance of CNNs in detecting the small low-contrast structures of the ACR phantom in the test image. We also revised the visibility labels of our dataset assigned by the technicians, removing those with gross errors.

The remainder of this paper is organized as follows. Section 2 describes the ACR low contrast test and the related works in the literature. Section 3 describes our main experiments: (A) our dataset; (B) how we calculated the coordinates of the holes; (C) how we labeled the visibility of holes; (D) the structure of our convolutional neural network; (E) the results we obtained; and (F) the results we obtained using a completely independent data set. Section 4 presents additional discussions: (A) the differences between our old and new datasets; (B) the results obtained by feeding CNN with ROI indices; and (C) the results obtained using classical machine learning algorithms. Finally, we present our conclusions in Section 5.

II. ACR LOW-CONTRAST DETECTABILITY TEST

A. THE LOW-CONTRAST TEST

In the ACR imaging protocol, the low-contrast detectability test consists of the acquisition of four axial slices in the posterior region of the ACR phantom, where there are thin circular plastic films, each perforated with 30 holes of varying diameters, arranged in 10 radial spokes (the term ''spoke'' denotes the set of three holes of the same diameter, aligned radially – Fig. 1). The four axial slices correspond to slices 8–11 in the imaging protocol [3]. The plastic films that are in that position have different thicknesses which determine the contrast between the holes and the background of the image.

While holes are filled with ionic solution and give maximum signal intensity, signal from the background depends on the thickness of the disk made of a material that emits no signal (Fig. 2). The holes of the same spoke have the same diameter, which gradually decreases clockwise, going from 7.0 mm to 1.5 mm. All holes in a given slice have the same contrast level, being 1.4%, 2.5%, 3.6% or 5.1% depending on the slice (8 to 11). Holes rotate counterclockwise, by nearly 9 degrees from slice to slice. A spoke is considered visible when all 3 of its holes can be clearly detected. The ACR low-contrast test consists of counting how many of the 10 spokes can be detected in a given slice. For example, a possible recommendation is that the spoke count of a 1.5T system should be equal or higher than 28.

B. RELATED WORKS

The detectability threshold is a manifestation of human perception, and empirical models describe it quite well when images are sharp and free of artifacts. Human perception in complex cases requires more sophisticated models. The method proposed by Fitzpatrick [4] to automate the ACR low-contrast detectability test is derived from Rose's visual perception model [5] and exemplifies the difficulty in predicting human detection capability in a real scenario using a simplistic model.

Davids et al. [6] implemented methods for fully-automatic evaluation of MRI quality measurements. However, they did not implement the low-contrast detection test.

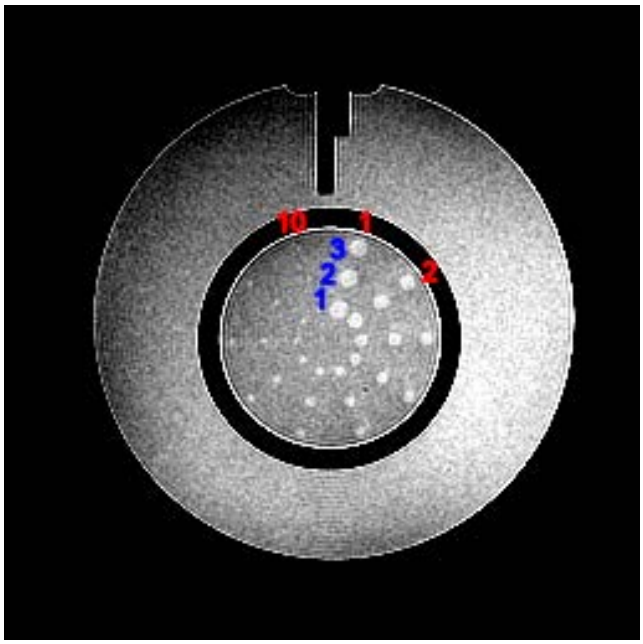
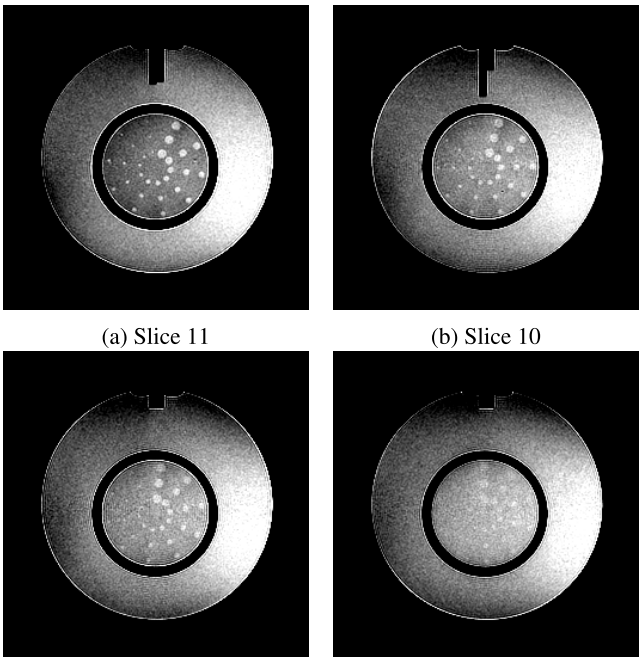


FIGURE 1. A typical image of a slice of the ACR MRI phantom. Red numbers 1-10 are the radial spoke indices (angle). Blue numbers 1-3 are the hole indices inside each spoke (radial position).



(a) Slice 11 (b) Slice 10

(c) Slice 9 (d) Slice 8

FIGURE 2. T1 images of slices 11 to 8 of the ACR Phantom. MRI images are acquired as 16-bit unsigned integer images, with 12 significant bits.

Sun et al. [7] described an open source automatic quality assurance tool for the ACR MRI test. For the low-contrast test, they implemented a module to assess the visual detection threshold, specific (according to the authors) for each user and computer monitor. They do not report the agreement between the responses of their system and human observers.

Panych et al. [8] described a solution to automate the high-contrast test. However, they point that the low-contrast test must still be done by a human.

Ehman et al. [9] devised an algorithm based on fuzzy logic to automate the low-contrast resolution test but found modest to low correlation between human and computer outputs.

Alaya et al. [10] estimated the hole visibility by drawing circular intensity profiles with varying radii, looking for peaks, computing peak contrasts based on intensity differences and applying an unspecified threshold. They compared the program's outputs with human readouts by counting the total number of holes, rather than evaluating the agreement of individual holes. This approach is even simpler than our previous work [2] because they use only intensity in and out of the hole and do not use the noise levels. Furthermore, they simply thresholded the ''contrast'', without using machine learning algorithms to mimic human operator.

Doi et al. [11] developed a CNN-based method to assess the low-contrast resolution of computed tomography (CT) images. CT is a completely different imaging modality from MRI. They focuse on assessing the quality of CT reconstruction in order to guide the development of new reconstruction algorithms. It is not directly related to the automation of the low contrast ACR MRI test.

Epistatou et al. [12] automate some of ACR MRI quality control tests evaluating four parameters: percent signal ghosting, percent image uniformity, signal-to-noise ratio (SNR), and SNR uniformity. They do not automate low-contrast test.

Teuho et al. [13] present the results of a software that automates five ACR MRI tests: geometric accuracy, slice thickness accuracy, slice position accuracy, image intensity uniformity and percent signal ghosting. They do not automate low-contrast test.

To the best of our knowledge, the automation of the ACR low-contrast test remains an open problem.

III. EXPERIMENTS

A. DATASET

Our dataset consists of 100 ACR phantom acquisitions made by 13 scanners of different vendors (Siemens, GE and Philips), magnetic fields (1.5T and 3.0T) and head coils (8, 12 and 32 channels), totaling 400 256 $\times$ 256 images. Each image has 30 low-contrast structures making up 12,000 ROI (Region Of Interest) images with 17 $\times$ 17 pixels.

B. COMPUTATION OF HOLE COORDINATES

Each hole received a tag consisting of 3 indices: slice (a number from 8 to 11), angle (or spoke, from 1 to 10) and radial position (1 to 3) – see Figs. 1 and 2. The coordinates of hole centers vary from acquisition to acquisition. We calculated these coordinates as follows:

1) We co-registered the image of slice 11 (which has the highest contrast) with the template of holes to obtain

the parameters of an affine transformation and thus calculated the hole coordinates in this slice.

2) We rotated the coordinates of the holes in slice 11 counterclockwise in steps of roughly 9 degrees to obtain the coordinates in slices 8-10.

C. LABELING THE VISIBILITY OF HOLES

A human being can only distinguish 700-900 levels of gray, even under ideal conditions [14]. As an MRI image is 12 bits or 1024 levels of gray, it is not possible to distinguish all shades in a still image, even with proper brightness/contrast adjustment (also known as windowing). Thus, the technician does not evaluate the visibility of a hole in a static image. She keeps dynamically changing the brightness/contrast of the image to check if the hole becomes visible under certain dynamically changing settings.

Using an in-house application, our technicians gave an answer (visible/invisible) for each hole. The application basically consisted of a pair of windows arranged side by side, as shown in Fig. 3: one where the technician could click on the holes she considered ''detectable''; and the other, a blank screen where red circles would appear to provide a clue that the mouse click was effective. A subsequent click on the same region changes the status back to ''undetectable'', and so on.

Slices were presented from 11 to 8. Technicians screened a batch of 10 acquisitions at a time and only eventually read 2 batches in the same day (with a minimum rest of 2 hours between the screening sessions). Images could be zoomed, panned and windowed. All sessions took place in the same dark room and using a single monitor with fixed presets. A total of $100 \times 120 = 12{,}000$ holes were labeled ''detectable'' or ''undetectable'' by experienced technicians under strictly controlled conditions.

The result of an ACR low-contrast test is the total number of visible spokes and a spoke is considered visible if all the 3 holes can be detected. That is, the operator counts spokes, but evaluates the visibility of each hole individually. To increase the power of our learning algorithms, we modeled the visibility of each hole individually, not of the spoke. Our technicians classified each of the 120 holes of each acquisition. The 30 holes of the 400 images were labeled by 4 to 7 technicians, assigning 0 to non-visible holes and 1 to visible ones.

It is humanly understandable that a technician could err in labeling thousands of samples. Thus, the authors of this paper reviewed the labeling of each of the 400 images and discarded those that contained gross errors. Gross errors include: declaring some holes to be invisible in an image I when they are clearly visible; claiming some holes in an image I to be visible when most other technicians and ourselves cannot see them; etc. In these cases, all labels given by that technician on image I have been discarded. Discarding gross errors, each image was labeled by 2 to 7 technicians, according to Tab. 1. For example, after discarding gross errors, 9 images were labeled by only 2 technicians, 33 images were labeled by 3 technicians, and so on. We defined as the gold standard the median of the technicians' responses (after discarding gross errors), rounding up to 1 in cases of a tie. This procedure resulted in 1,935 holes labeled as invisible and 10,065 ones labeled as visible.

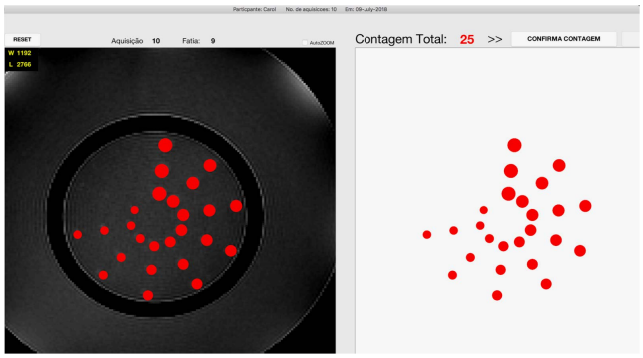


FIGURE 3. In-house application where technicians can click on the holes she deem "visible."

TABLE 1. Number of images by number of technicians who labeled them.

							total
number of images	9	33	244	44	66	4	400
number of technicians	2	3	4	5	6	7	

D. CONVOLUTIONAL NEURAL NETWORK

Recently, there has been a real revolution in image classification with the introduction of the deep convolutional neural network [15], [16], [17]. In CNNs, the pattern of connectivity between neurons is inspired by the organization of the animals' visual cortex [18]. CNN is a class of feed-forward artificial neural network designed to require as little pre-processing as possible. It automatically designs, using the sample images with labels, the low-level filters to extract useful features and the high-level filters to concatenate properly these features. In contrast, the extraction of low-level features have to be designed manually in the classic machine learning algorithms. This independence of a priori knowledge and human effort in the development of machine learning system is the biggest advantage of CNNs over the classic techniques.

In this work, we use a CNN, implemented in Keras/TensorFlow, to predict the visibility of the holes. Our system reads a CSV (Comma-Separated Values) file with the center coordinates of the holes and their gold standard visibility labels. It also reads 400 16-bit 256×256 images (with 12 significant bits) corresponding to the 100 MRI acquisitions. From this data, we extract 12,000 ROIs with 17×17 pixels around the center of each hole, with the corresponding visibility labels (Fig. 4). These 12,000 ROIs with labels are used as the training and test samples of our classification problem.

We use 5-fold cross validation to get reliable performance metrics along with standard deviations. For this, we randomly split 12,000 ROIs into 5 subsets of 2,400 ROIs each, without any stratification. In each fold, we take 4 subsets

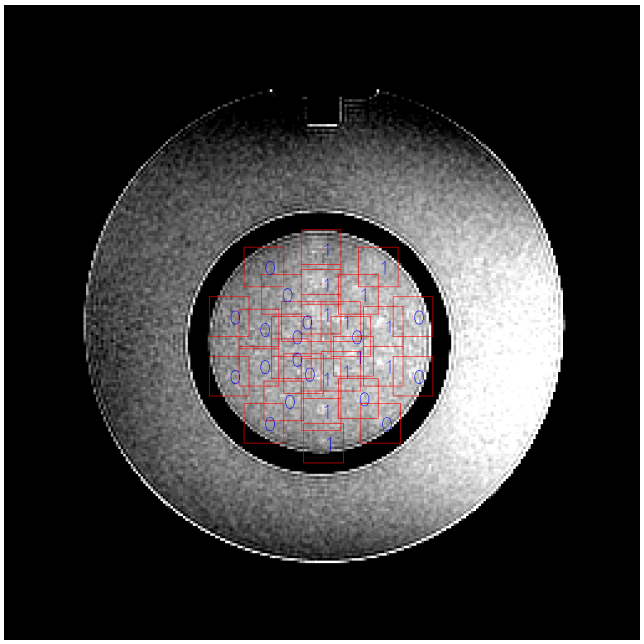


FIGURE 4. We extract ROIs with 17 × 17 pixels around the center of each hole. The labels 0/1 indicate that the hole is invisible/visible.

as the training samples and the remaining subset as the test sample.

We tested several CNN architectures and describe the one that generated the highest AUC. First, we compute the mean μ and the standard deviation σ of the training pixels to normalize both the training and test pixels $P_n = (P_o - \mu)/\sigma$, where P_o is the original 16-bit unsigned integer pixel value and P_n is the normalized 32-bit float pixel value.

Then, we make a simple data augmentation. We take each of the 9,600 training ROIs and shift one pixel in the north, south, east, and west directions, while keeping the original ROI. So the original 9,600 training ROIs become $5 \times 9{,}600 = 48{,}000$ data-augmented ROIs. We did not introduce sophisticated geometric deformations because the images were very small.

We use a simple CNN inspired by VGG (Visual Geometry Group [19]) model with the structure depicted in Fig. 5. It consists of the sequence of three VGG-inspired blocks (blue rectangles in Fig. 5) with the internal structure:

```
Conv2D(n, kernel=(3,3))
BatchNormalization()
Dropout(0.3)
Conv2D(n, kernel=(m,m))
BatchNormalization()
MaxPooling2D(pool=(2,2))
```

where the numbers of convolutions are $n = 64$, 96 and 128 in the first, second and third VGG-inspired blocks, respectively. All convolutional layers are followed by *relu* activation functions and use L_2 kernel regularizer with parameter 5×10^{-4}. All convolutional layers use kernel 3×3 with "same" padding (to keep input and output resolutions the same), except the second convolutional layer of the first VGG block that uses kernel size 2×2 with "valid" padding, to decrease the resolution of the image 17×17 to 16×16, in order to be divisible by 2 many times (suitable for sequence of 2×2 max-poolings).

The convolutions automatically design the sequence of filters to extract useful features and properly combine them. The batch normalization layers help in the convergence of the learning process and the dropout layer helps to avoid overfitting. The max-pooling layers decreases the resolution of the extracted features. The outputs of VGG blocks are 64 feature maps with 8×8 features (after the first block), 92 feature maps with 4×4 features (after the second block) and 128 feature maps with 2×2 features (after the third block). These features are "flattened", that is, converted into a 1-D vector with $2 \times 2 \times 128 = 512$ features, and go through the dense block (red rectangle in Fig. 5) consisting of two dense (fully-connected) layers:

```
Flatten()
Dense(60)
BatchNormalization()
Dropout(0.3)
Dense(1)
```

The first dense layer uses L_2 kernel regularizer with parameter 5×10^{-4} and *relu* activation. The second dense layer does not use kernel regularizer and uses *linear* activation. The output of the dense block is a prediction number between 0 and 1, so that the closer to 1 the more probably the hole is visible. We adopted "mean squared error" as the loss function and ADAM (Adaptive Moment Estimation) as the optimizer. The ADAM optimizer begins with its default learning rate of 0.001 that is reduced by factor of 0.9 whenever reaches a plateau of training accuracy. We used batch size of 32 and trained the CNN for 150 epochs.

E. RESULTS

Our system, like most artificial intelligence (AI) classification systems, does not return a binary answer. Instead, it returns a "grade" from 0 to 1, where the closer to 1, the more likely the hole is visible. Consequently, it is not possible to compute sensitivity and specificity directly from the system responses. It is necessary first to threshold the "grade" to obtain a Boolean answer and then calculate sensitivity and specificity (as well as type I and II errors). The ROC (Receiver Operating Characteristic) curve plots the sensitivities and specificities obtained by varying the threshold to all possible values between 0 and 1 and the area under the ROC curve measures the system performance regardless of the chosen threshold. Thus, AUC does not depend on the chosen threshold and this is the reason why it is so popular. Using the ROC plot, it is possible to calculate the sensitivity for a given specificity or the specificity for a given sensitivity. There is a special point on the ROC, called the equal error rate (EER) point, where accuracy, sensitivity and specificity all become equal. At this special point, it is possible to calculate accuracy,

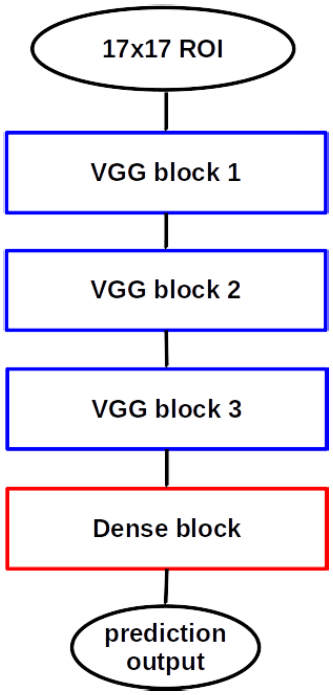


FIGURE 5. Structure of our CNN to classify the visibility of a ROI image.

TABLE 2. AUCs and EERs of our 5-fold experiments, without TTA.

	fold1	fold2	fold3	fold4	fold5	mean±std
EER	0.063	0.066	0.068	0.074	0.087	0.072±0.009
Ac., sen., spe. at EER	0.937	0.934	0.932	0.926	0.913	0.928±0.009
AUC	0.984	0.983	0.981	0.980	0.976	0.981±0.003

sensitivity and specificity without choosing a threshold value and, at the same time, obtain a metric that has an intuitive interpretation.

Without using test-time augmentation (TTA), we obtained the results described in Tab. 2. Our system yielded a mean AUC of 0.981 ± 0.003 with mean EER of $7.2 \pm 0.9\%$ (that is, the sensitivity, specificity and accuracy at the EER point are all equal to 92.8%). Fig. 6 depicts the ROC curves obtained without TTA.

Using TTA, performance improved even further. As we did with the training images, we shifted each test image in the north, south, east, and west directions. With that, each test image generated 5 images (4 distorted plus the original). We fed all these images to the AI system and averaged the 5 predictions. The results are shown in Tab. 3. All performance metrics improved slightly: AUC increased from 0.981 to 0.9833; accuracy, sensitivity and specificity at the EER point increased from 92.8% to 93.2%. Fig. 7 depicts the ROC curves obtained with TTA. The obtained AUC is quite high and the standard deviation is quite low, which means that similar results are obtained when repeating the experiments.

The performance measures of the 4 technicians that labeled all images of the dataset are shown in Tab. 4 (other technicians have only labeled parts of the dataset). To calculate the performance of a technician T, we cannot use the same gold

TABLE 3. AUCs and EERs of our 5-fold experiments, with TTA.

	fold1	fold2	fold3	fold4	fold5	mean±std
EER	0.063	0.066	0.068	0.074	0.087	0.068±0.007
Ac., sen., spe. at EER	0.940	0.938	0.934	0.928	0.920	0.932±0.007
AUC	0.985	0.985	0.985	0.982	0.979	0.983±0.003

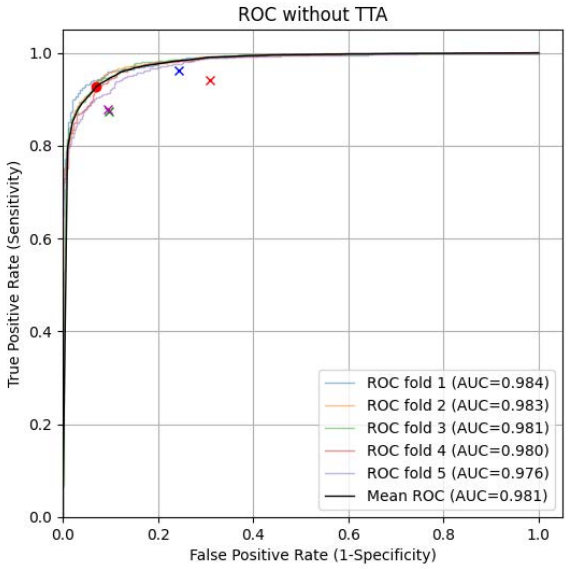


FIGURE 6. ROC curves obtained in 5-fold cross validation without TTA (faded colors) and the mean ROC curve (black). The red dot indicates the EER point. The four "X" marks indicate the sensitivity-specificity points of the four technicians.

TABLE 4. Accuracy, sensitivity and specificity of the 4 technicians.

	Technician 1	Technician 2	Technician 3	Technician 4
Accuracy	0.898	0.878	0.924	0.883
Sensitivity	0.940	0.873	0.961	0.878
Specificity	0.690	0.904	0.757	0.905

standard that we use to measure the performance of the AI system, as the technician T's own responses would go into the gold standard calculation. Thus, to compute the performance of a technician T, we used as the gold standard the median of the responses of the technicians excluding the response of the technician T herself. As before, gross errors were discarded from the gold standard calculation.

We can calculate the AUC of our CNN system as it returns a number in the range from 0 to 1. Meanwhile, technicians give binary responses (visible or invisible). From the binary responses, we can compute accuracy, sensitivity and specificity, but it is impossible to calculate AUC. Accuracy is not a good performance measure for our problem because our dataset is highly unbalanced. As we have many more visible holes (10,065) than invisible ones (1,935), a system/technician with a tendency to classify holes as visible will have a higher accuracy than one with a tendency to classify holes as invisible. Sensitivity and specificity are also not good performance measures, as there is a

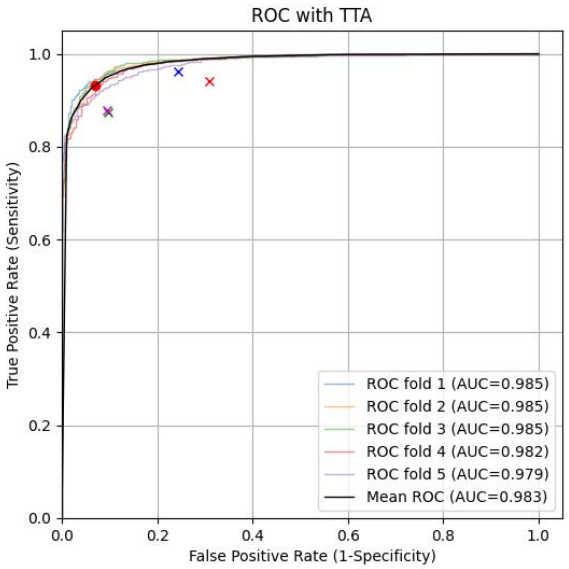


FIGURE 7. ROC curves obtained in 5-fold cross validation with TTA (faded colors) and the mean ROC curve (black).

trade-off between the two, such that increasing one causes the other to decrease. Therefore, in order to fairly compare the responses of the technicians with our system, we do not use accuracy, sensitivity or specificity alone. Instead, we plot the specificity-sensitivity points of the 4 technicians on our system's ROC curve. The four "X" marks in red, green, blue and magenta in Figs. 6 and 7 represent the performances of technicians 1 to 4, respectively. As all ROC curves of AI system are above these four points, we can conclude that AI system performs better than any individual technician.

F. TESTS IN AN INDEPENDENT DATASET

To further test our system, we used the ensemble of the five CNN models obtained above to classify a completely independent test dataset. This dataset consisted of 90 ACR phantom acquisitions, totaling $90 \times 4 = 360$ images with $360 \times 30 = 10{,}800$ holes. Three technicians T_1, T_2 and T_3 classified each hole as visible or invisible. As before, the authors of this paper reviewed the labeling of each image and discarded those containing gross errors. Discarding gross errors, each image was labeled by 1 to 3 technicians, according to Tab. 5. An ensemble model uses multiple models to obtain better predictive performance than could be obtained from any of the constituent model. In our case, the ensemble model averages the responses of the five models computed before in 5-fold cross validation.

We used as the gold standard the median of the technicians' responses, after discarding gross errors, rounding up to 1 in cases of a tie. This procedure resulted in 1,821 holes labeled as invisible and 8,979 labeled as visible. Using the ensemble model and TTA (4 shifted images plus the original), we obtained the ROC curve depicted in Fig. 8. We plot the specificity-sensitivity points of the technicians T_1 and T_2 on our system's ROC curve as "X" marks. As our ROC curves is above these points, we conclude that our system performs better than the two technicians. To calculate the performance of a technician T, the median of the responses of the technicians excluding the response of T herself was used as the gold standard. It was not possible to compute the performance of the technician T_3 because there were some images that were labeled only by herself (after discarding gross errors). Fig. 9 depicts some ROI images with the respective CNN's predictions. Note that the original images are 12-bit, but in this document they are represented as 8-bit, so there may be information in the original images that became invisible when the number of bits was reduced.

According to the ACR MRI manual [3], a spoke is considered visible if and only if all three of its constituent holes are visible. We computed the average spoke classification errors by acquisition, obtaining the error rates depicted in Tab. 6.

The error rate of our system (6.17%) using a suitable threshold (0.68) is much lower than those of the technicians T_1 and T_2 (12.31% and 14.67%). In the last two columns, we chose thresholds to result in false positive cases (51 and 42) similar to those of the technicians (51 and 43). Even in this situation, the error rates of our system (8.06% and 8.92%) are substantially lower than those of the technicians T_1 and T_2 (12.31% and 14.67%).

TABLE 5. Number of independent test images by number of technicians who labeled them.

				total
number of images	29	11	320	360
number of technicians	1	2	3	

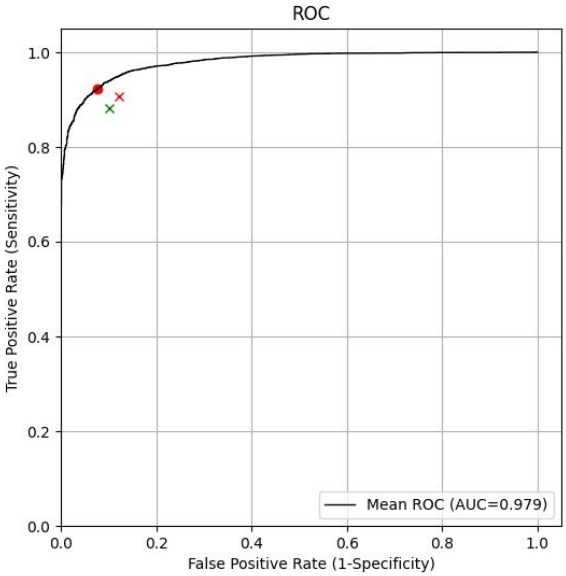
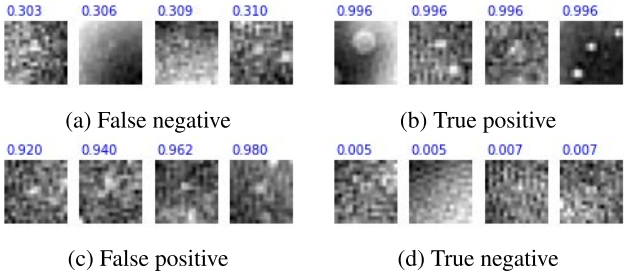


FIGURE 8. ROC curve obtained processing an independent test set with ensemble of 5 models and TTA. The two "X" marks indicate the sensitivity-specificity points of the technicians T_1 and T_2.



(a) False negative (b) True positive

(c) False positive (d) True negative

FIGURE 9. Examples FN, TP, FP and TN cases (using, for example, threshold 0.68). The blue numbers are CNN's predictions.

TABLE 6. Spoke classification errors using independent test images by the two technicians (T_1 and T_2) and by the AI system with different thresholds. According to gold standard, there are 2,779 visible spokes and 821 invisible ones.

	T_1	T_2	AI				
Threshold			0.66	**0.68**	0.70	0.84	0.87
TP	2516	2433	2654	2648	2638	2540	2500
TN	641	639	723	730	735	770	779
FP	51	43	98	91	86	51	42
FN	392	485	125	131	141	239	279
Errors	443	528	223	**222**	227	290	321
Error rate	12.31%	14.67%	6.19%	**6.17%**	6.31%	8.06%	8.92%

TABLE 7. Approval/rejection of MRI devices in disagreement with the gold standard by the two technicians and by the AI system with threshold 0.68.

	1.5T				3T			
	FP	FN	errors	er. rate	FP	FN	errors	er. rate
T_1	1	7	**8**	**12%**	0	8	8	33%
T_2	1	14	15	23%	0	9	9	38%
AI	5	3	**8**	**12%**	1	3	**4**	**17%**

In our test data, 66 acquisitions were made on 1.5T machines and 24 on 3T machines. Using criteria that the number of visible spokes n must be $n \geq 28$ and $n \geq 37$ to approve respectively 1.5T and 3T machines, the gold standard would have approved 49 (74%) of 1.5T and 10 (42%) of 3T machines. The two technicians and the AI system disagreed with the gold standard according to Tab. 7. In all cases, the AI system disagreed less or equally with the gold standard than the technicians T_1 or T_2. Most of the ''errors'' made by the technicians are of the false negative type, when they reject a machine that would have been approved by the gold standard. This means that T_1 and T_2 classified as invisible many holes that T_3 considered visible. There are technicians that tend to consider holes as visible or invisible.

Note that the gold standard is far from foolproof as it is just the median of the opinions of the technicians, eliminating the answers with gross errors. Furthermore, only 3 technicians labeled the test dataset, and some images were labeled by only 1 or 2 technicians.

IV. DISCUSSIONS

A. OLD AND NEW DATASETS

In our previous conference paper [2], we considered the answers of senior technicians, with more than 10 years of experience, as ''gold standard''. However, by carefully analyzing our dataset, we concluded that senior technicians make as many gross errors as newer technicians, and years of experience do not, by themselves, guarantee greater accuracy in classification. Thus, we changed the ''gold standard'' of visibility of holes in an image I to the median of responses from all technicians (regardless of years of experience) who did not make gross errors in classifying holes in I. If we found that a technician T made some gross errors in classifying holes in I, all labels for I provided by T were discarded.

B. CNN WITH ROI INDICES

We also tested feeding the CNNs with ROI indices, in addition to the ROI image itself. A ROI index is composed of three numbers: slice (from 8 to 11), angle (or spoke, from 1 to 10) and radial position (1 to 3) – see Figs. 1 and 2. We tested using them as features because, intuitively, they may help the classification:

- Slice – the contrast of the image depends on this number.
- Angle (spoke) – the diameter of the hole depends on this number.
- Position – usually, the outer holes are more distorted and difficult to visualize than the inner holes.

These numbers were normalized to range from -1 to $+1$ before entering the CNN, passing through a dense layer and being concatenated with 512 features extracted from the image. Contrary to the expectations, we did not get any improvement with this modification. This can mean that CNNs are able to extract this information from the image itself.

C. MANUALLY DESIGNED FEATURES AND CLASSIC MACHINE LEARNING

In our previous work [2], we extracted some manually designed features from ROI images and used conventional machine learning algorithms to achieve the maximum AUC of 0.878. In the present work, using CNNs, we achieved a much higher AUC (0.983). However, the two works are not directly comparable because they use different ''gold standard'' labels. To fairly compare the two approaches, we repeated the previous experiments using the new dataset.

As in [2], we used the four main features extracted from the ROIs:

- S_{in}: The average signal (mean value) inside the hole represented as a float variable normalized to the range between 0 and 1.
- N_{in}: The noise (standard deviation) inside the hole normalized to the range between 0 and 1.
- S_{out}: The average signal in the surrounding area, normalized to the range between 0 to 1.
- N_{out}: The noise in the surrounding area, normalized to the range between 0 and 1.

Fig. 10 depicts the masks used to compute these features.

As in [2], we also used the three ROI indices as features: slice (from 8 to 11), angle (1 to 10) and radial

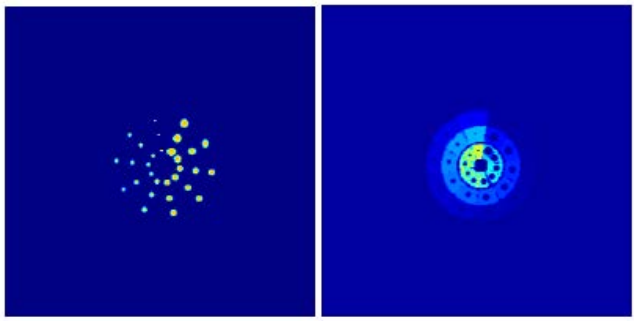


FIGURE 10. The masks used to compute the average and standard deviation inside the holes (left) and in the surrounding areas (right).

TABLE 8. Average of 5-fold cross validation results using classic machine learning algorithms with and without the three ROI indices (slice, angle and position).

	With ROI indices				Without ind.
	MSE	EER	Acc. at EER	AUC	AUC
Log. Regres.	0.098±0.006	0.166±0.008	0.834±0.008	0.906±0.006	0.716±0.020
Sup. Vec. Mac.	0.099±0.005	0.175±0.013	0.825±0.013	0.892±0.006	0.724±0.032
Rand. Forest.	**0.088±0.004**	**0.126±0.005**	**0.874±0.005**	**0.942±0.002**	**0.888±0.011**
Mult. Percept.	0.101±0.009	0.164±0.009	0.836±0.009	0.908±0.006	0.747±0.016
Ext. Grad. Boo.	0.091±0.004	0.126±0.004	**0.874±0.004**	**0.943±0.003**	0.884±0.009

position (1 to 3) – Figs. 1 and 2. We tested five classic machine learning algorithms: Logistic Regression, Support Vector Machine, Random Forest, Multilayer Perceptron and Extreme Gradient Boosting, provided by Scikit-Learn library. The averages of 5-fold cross validation results using classic machine learning algorithms are described in Tab. 8.

Random Forest and Extreme Gradient Boosting yielded quite good results (AUCs of 0.942±0.002 and 0.943±0.003), but substantially inferior to CNNs (AUC of 0.983 ± 0.008). It is possible to draw the same conclusion from the accuracy, sensitivity and specificity at the EER point: Random Forest and Extreme Gradient Boosting yielded 87.4 ± 0.5% and 87.4 ± 0.4% while CNNs yielded 93.2 ± 0.7%. This means that the average signal and noise inside and outside the hole are suitable features for classifying the visibility of the holes (although CNNs are even better).

Unlike CNNs, classic machine learning algorithms seem to heavily rely on the three ROI indices. Considerably worse results are obtained when they are withdrawn: compare the last two columns of Tab. 8.

V. CONCLUSION

In this paper, we have proposed to automate ACR MRI low-contrast detectability test using convolutional neural network. Apparently, this is the first work that actually manages to emulate the perception of a human observer in the ACR low-contrast test.

We created a dataset with 100 ACR phantom acquisitions, totaling 12,000 holes. Experienced technicians labeled each hole as "visible" or "invisible" and the median of technicians' responses were considered the gold standard label. We divided the dataset into 5 subsets and used 5-fold cross validation to train and test the AI system 5 times. We obtained a mean AUC of 0.983 ± 0.003 and a mean accuracy of 93.2 ± 0.7% at the EER point, that are better than any of the individual technicians' results.

We repeated the experiments using an independent test set, obtaining an AUC of 0.979. The classification of spokes by the AI system agrees with the gold standard more than any individual technician. The AI system's decisions to approve or reject MRI machines also agree more or equally with the gold standard than the decisions made by any individual technician. These results show that this test can be confidently automated using CNNs.

We also used manually designed features (signal and noise inside and outside of the holes) and classic machine learning algorithms to do the same task, obtaining a mean AUC of 0.943 ± 0.003 and a mean accuracy of 87.4 ± 0.4% at the EER point. These results show that CNNs are superior to the classic machine learning algorithms using manually-designed features.

ACKNOWLEDGMENT

The authors would like to thank technicians Wanderlir Alves de Faria and Ana Paula Carvalho Caldeira Pires from the Radiology Department, Hospital Israelita Albert Einstein, for their help in analyzing the phantom images.

REFERENCES

[1] R. W. Brown, Y. C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan, *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Hoboken, NJ, USA: Wiley, 2014.

[2] J. E. Ramos, H. Y. Kim, and F. B. Tancredi, "Automation of the ACR MRI low-contrast resolution test using machine learning," in *Proc. 11th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Beijing, China, Oct. 2018, pp. 1–6.

[3] (2015). *ACR Magnetic Resonance Imaging Quality Control Manual*. [Online]. Available: https://www.acr.org/-/media/ACR/Files/Clinical-Resources/QC-Manuals/MR_QCManual.pdf

[4] A. O. Fitzpatrick, "Automated quality assurance for magnetic resonance imaging with extensions to diffusion tensor imaging," Doctoral dissertation, Virginia Tech, Blacksburg, VA, USA, 2005.

[5] A. Rose, "A unified approach to the performance of photographic film, television pickup tubes, and the human eye," *J. Soc. Motion Picture Eng.*, vol. 47, no. 4, pp. 273–294, 1946.

[6] M. Davids, F. G. Zöllner, M. Ruttorf, F. Nees, H. Flor, G. Schumann, and L. R. Schad, "Fully-automated quality assurance in multi-center studies using MRI phantom measurements," *Magn. Reson. Imag.*, vol. 32, no. 6, pp. 771–780, Jul. 2014.

[7] J. Sun, M. Barnes, J. Dowling, F. Menk, P. Stanwell, and P. B. Greer, "An open source automatic quality assurance (OSAQA) tool for the ACR MRI phantom," *Australas. Phys. Eng. Sci. Med.*, vol. 38, no. 1, pp. 39–46, Mar. 2015.

[8] L. P. Panych, J.-Y.-G. Chiou, L. Qin, V. L. Kimbrell, L. Bussolari, and R. V. Mulkern, "On replacing the manual measurement of ACR phantom images performed by MRI technologists with an automated measurement approach," *J. Magn. Reson. Imag.*, vol. 43, no. 4, pp. 843–852, Apr. 2016.

[9] M. O. Ehman, Z. Bao, S. O. Stiving, M. Kasam, D. Lanners, T. Peterson, R. Jonsgaard, R. Carter, and K. P. McGee, "Automated low-contrast pattern recognition algorithm for magnetic resonance image quality assessment," *Med. Phys.*, vol. 44, no. 8, pp. 4009–4024, Aug. 2017.

[10] I. B. Alaya and M. Mars, "Automatic analysis of ACR phantom images in MRI," *Current Med. Imag. Formerly Current Med. Imag. Rev.*, vol. 16, no. 7, pp. 892–901, Sep. 2020.

[11] Y. Doi, A. Teramoto, A. Yamada, M. Kobayashi, K. Saito, and H. Fujita, "Estimating subjective evaluation of low-contrast resolution using convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 44, no. 4, pp. 1285–1296, Dec. 2021.

[12] A. C. Epistatou, I. A. Tsalafoutas, and K. K. Delibasis, "An automated method for quality control in MRI systems: Methods and considerations," *J. Imag.*, vol. 6, no. 10, p. 111, Oct. 2020.
[13] J. Teuho, V. Saunavaara, S. Heikkinen, I. Ranta, J. Saunavaara, and M. Teräs, "MRI quality control and system performance using ACR phantom tests and automated software," *Phys. Medica*, vol. 52, p. 121, Aug. 2018.
[14] T. Kimpe and T. Tuytschaever, "Increasing the number of gray shades in medical display systems—How much is enough?" *J. Digit. Imag.*, vol. 20, no. 4, pp. 422–432, 2006.
[15] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.
[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.
[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[18] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Netw.*, vol. 16, nos. 5–6, pp. 555–559, Jul. 2003.
[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

JHONATA EMERICK RAMOS was born in Brasília, Brazil, in 1981. He received the B.E. degree in aeronautical engineering from the University of São Paulo, Brazil, in 2009, and the master's degree in economics from Fundação Getúlio Vargas, in 2015. He is currently pursuing the Ph.D. degree in computer engineering with the University of São Paulo.

He founded a logistics startup called Rapiddo that was sold to iFood, in September 2018. Then, he founded two other startups that use machine learning, helping Latin America's largest companies to make better decisions. He was a Professor at FIA/USP, between 2019 and 2020, in the following subjects: statistics, data mining, big data, machine learning, and deep learning. He is the coauthor of one patent and coauthor of the book *Econometrics with Eviews—Essential Guide to Concepts and Applications*. He worked in Brazil in large companies, such as Embraer and Itaú.

HAE YONG KIM was born in South Korea. He received the B.S. and M.S. degrees (Hons.) in computer science and the Ph.D. degree in electrical engineering from the Universidade de São Paulo (USP), São Paulo, Brazil, in 1988, 1992, and 1997, respectively.

He is currently an Associate Professor with the Department of Electronic Systems Engineering, USP. He is the author of more than 100 articles and holds three patents. His research interests include image processing, machine learning, medical image processing, and computer security.

Dr. Kim and colleagues received the 6th edition of the Petrobras Technology Award in the "Refining and Petrochemical Technology" category, in 2013, the "Best Paper in Image Analysis" Award at the Pacific-Rim Symposium on Image and Video Technology, in 2007, and the Thomson ISI Essential Science Indicators "Hot Paper" Award, for writing one of the top 0.1% of the most cited computer science papers, in 2005.

FELIPE BRUNETTO TANCREDI received the Ph.D. degree in biomedical engineering from the University of Montreal, in 2014. He is the coauthor of ten scientific articles and two patents, served as a reviewer for different journals in the field of MRI and neuroscience, and was a member of the Editorial Board of the Prestigious *Journal of Cerebral Blood Flow and Metabolism*. He worked at GE HealthCare and private/public hospitals before founding RadSquare, in 2019, a healthtech specialized in image processing automation. He was responsible for structuring the MRI quality assurance program at Albert Einstein Jewish Hospital and for obtaining the ACR accreditation for all its 14 scanners, in 2015, a pioneering work in Brazil.