Automation of the ACR MRI Low-Contrast Resolution Test Using Machine Learning

Jhonata E. Ramos[1], Hae Yong Kim[2]
Escola Politécnica
Universidade de São Paulo
São Paulo, Brazil
[1] jhonata.emerick@usp.br, [2] hae@lps.usp.br

F. B. Tancredi
Imaging Research Center – Dept. of Radiology
Hospital Israelita Albert Einstein
São Paulo, Brazil
felipe.tancredi@einstein.br

Abstract— **Magnetic Resonance Imaging (MRI) is a powerful, widespread and indispensable medical imaging modality. The American College of Radiology (ACR) recommends weekly acquisition of phantom images to assess the quality of scanner. Usually, these images must be analyzed by experienced technicians. Automatic analysis of these images would reduce costs and improve repeatability. Some automated methods have been proposed, but the automation of two of the ACR image quality tests remains open problem. Reports on the high- and low-contrast resolution tests are scarce and so far none of the proposed methods produce results robust enough to allow replacing human work. We use Machine Learning to emulate, with high accuracy, the detection of 120 low-contrast structures of ACR phantom by an experienced professional. We used a database with 620 sets of ACR phantom images that were acquired on scanners of different vendors, fields and coils, totaling 74,400 low-contrast structures. Technicians with more than 10 years of experience labeled each structure as 'detectable' or 'undetectable'. Machine learning algorithms were fed with image features extracted from the structures and their surroundings. Among the five methods we tested, Logistic Regression yielded the largest area under the ROC curve (0.878) and the highest Krippendorff's alpha (0.995). The results achieved in this study are substantially better than those previously reported in the literature. They are also better than the classifications made by junior technicians (with less than 5 years of experience). This indicate that the ACR MRI low-contrast resolution test may be automated using Machine Learning.**

Keywords - MRI, Quality Assurance, Visual Perception, Machine Learning.

I. INTRODUCTION

Magnetic Resonance Imaging (MRI) is the medical imaging modality that provides the largest range of image contrast [1]. It does not only produce images where bones can be distinguished from soft tissues, as in X-rays, but it also allows radiologists to detect small structures with low image contrasts, such as a meniscus tear, a myocardial infarction or an ovarian cyst. MRI scanners play a fundamental role in medical diagnosis.

Like any other medical measuring instrument, an MRI scanner must undergo periodical quality assurance (QA) tests and be recalibrated, promptly, whenever necessary. The American College of Radiology (ACR) MRI Accreditation Program is the most regarded and widespread MRI QA program [2]. In the US, the program is mandatory for institutions providing public healthcare, so most scanners operating in that country adhere to the program. But due to its prestige, the ACR accreditation is also sought by foreign international institutions that understand the importance of monitoring the performance of its scanners and consider ACR's recommendations fairly suitable.

The ACR recommends testing MRI scanners weekly. The QA image test consists of acquiring and evaluating images from the ACR multi-purpose phantom. Phantom is an object with known geometry and composition used to test imaging instrument. It is constructed to give origin to images or 3D volumes where one can measure the quality of the image signal. The ACR phantom is multi-purpose because there are at least 7 different QA metrics that can be extracted from its images. Only a few are monitored weekly. A full test using the phantom can take more than an hour; weekly tests take approximately 20 minutes. Automation of these tests, especially the low-contrast resolution test could improve the repeatability and reliability of QA measures, and save significant healthcare resources. There have been a few attempts to automate the ACR QA MRI tests [3, 4, 5, 6, 7] but, to date, none of the proposed methods has been recognized by the College.

Two of the tests are entirely dependent on the visual perception of an experienced operator and their automation has been challenging: the high and low-contrast resolution tests. In these tests, the operator must indicate whether a given set of structures in the phantom can be detected in the image (i.e., differentiated from the background). If these two tests could be automated, probably the entire ACR test could be done without

the presence of an experienced operator, reducing costs and improving repeatability.

In this paper, we address the automation of the low-contrast resolution test. We found two studies in the literature addressing this test [3, 4], but the correlations between the human and computer outputs were low.

II. ACR LOW-CONTRAST RESOLUTION TEST

The low-contrast resolution test estimates the detection of structures with small signal differences, using the typical protocol in a given equipment. It uses the 4 last slices of the typical 11-slice acquisition in the ACR phantom. In slices 8-11, the phantom has plastic disks of varying thicknesses, each with 30 holes. Holes are arranged in 10 radial triplets and the size of the holes decreases from 7mm to 1.5mm, clockwise (Fig. 1A). While holes are filled with phantom's 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). The thinner the plastic disk, the more contribution of the ionic solution to the signal; and the contrast between the holes and their background decreases. The phantom was conceived to produce disk hole contrasts from 5% (slice 11) to 1.5% (slice 8), as shown in figure 2. The test consists in counting how many of the 40 triplets can be seen in the slices 11 through 8. A triplet is considered visible only if all of its 3 holes can be clearly detected. A high counting indicates that the image quality is good and allows resolving small structures even when the contrast is low.

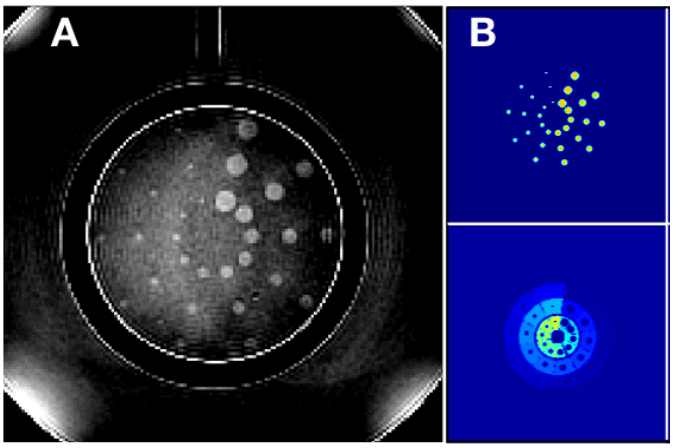

Figure 1. (A) A typical image of ACR MRI Phantom, slice 10; (B) two masks utilized to obtain the image characteristics.

The threshold of detectability is a manifestation of human perception, and empirical models describe it fairly well when the images are sharp and free from artifacts. Human perception in complex cases requires more sophisticated models. The method proposed by Fitzpatrick [4] to automate the ACR low-contrast resolution test is derived from the Rose Model of visual perception and exemplifies the difficulty in predicting the human detectability in a real scenario using a simplistic model. A group from the Mayo Clinic developed that model further [3]. They evaluated the performance of the algorithm using a database much larger than that used by the former study and found a correlation between human and computer outputs – measured using Krippendorff's alpha metric, a non-parametric index that measures the agreement between observations [8] – of 0.652, which is only modest and does not allow using their method in place of a trained professional.

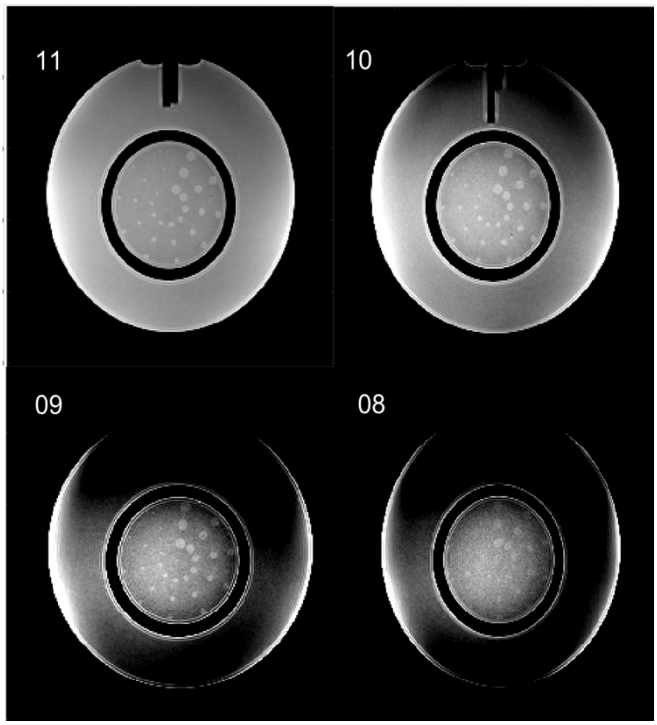

Figure 2. T1 images of slices 11 to 8 of the ACR Phantom. MRI images are acquired as 16 bits unsigned integer images.

III. EXPERIMENTS

The result of an ACR low-contrast test is the total number of visible triplets. A triplet is considered visible if all the 3 holes can be detected. That is, the operator counts triplets, but evaluates the visibility of each hole individually. Signal imperfections affect differently the visibility of each hole in the triplet so that the operator often detects 1 or 2 of its holes and not the other(s). To increase the power of our learning algorithms, we modeled the visibility of each hole individually, not the triplet. The operator usually stops counting the triplets at the last visible one. However, this does not necessarily mean that the holes in the remaining triplets cannot be detected. There may be many more visible holes and even entirely visible triplets. To take advantage of the all the available information, our technologists classified all the 120 holes of each acquisition, regardless of the total triplet counting.

We have in our database 620 ACR phantom acquisitions in the last 12 months, obtained in 13 scanners of different vendors (Siemens, GE and Philips), magnetic fields (1.5T and 3.0T) and head coils (8, 12 and 32 channels). That means, we have 74,400 low-contrast structures imaged in a great range of conditions to train our classification algorithms. All image

processing was carried out using in-house algorithms programmed in Matlab and R language.

A. *Feature Extraction*

Each low-contrast structure ("hole") received a label consisting of 3 numbers: slice (values from 8 to 11), angle (1 to 10) and radial position (1 to 3). We used two types of spatial masks (Fig. 1B): (a) a circle approximately the size of the hole, to measure signal characteristics within the hole; (b) an outer area to measure signal characteristics in its vicinity, consisting of sector minus the circle.

The holes' coordinates vary from acquisition to acquisition. Masks were adjusted by: (a) co-registering a template of holes with the image of slice 11 (that has the highest contrast) to obtain the parameters of an affine transformation; (b) using these parameters to register each mask on slice 11. On slices 10-8, we registered the masks by rotating the mask of slice 11 counterclockwise in steps of 9.5º.

We used the circle and outer masks to compute the mean and the standard deviation in each mask, obtaining four features: (a) S_IN: the signal (mean) inside the hole; (b) N_IN: the noise (standard deviation) inside the hole; (c) S_OUT: the signal (mean) in the surrounding area; (d) N_OUT: the noise (standard deviation) in the surrounding area.

The three hole label numbers (slice, angle and radial position) were also used as features. We use them as features because: (e) slice - the contrast of the image depends on this number and it helps to classify correctly the visibility of the holes; (f) angle - the radius of the hole depends on its angle in the slice, and the larger the hole, the easier it can be detected; (g) position - usually, the outer holes are more distorted and difficult to visualize than the inner holes.

Actually, we tested much more features before concluding that the most important features are the chosen seven, as described in Table I. The area under ROC curve increases slightly when we add some more features. However, as the improvements are marginal, we decided to discard them, in order to make the model simple and more easily interpretable.

TABLE I. DESCRIPTION OF THE FEATURES WE USED.

Variable	Type	Description
S_IN	Numerical	Signal inside the hole
N_IN	Numerical	Noise inside the hole
S_OUT	Numerical	Signal in the surrounding area
N_OUT	Numerical	Noise in the surrounding area
Angle	Categorical	The angle of the hole that indicates its size
Slice	Categorical	Slice where the hole is located
Position	Categorical	Position of the hole in triplet

B. *Predicted variable*

The visibility of the low-contrast structures is given as the total number of triplets visible to the operator. We already had our database with 620 acquisitions, where each of the slices 8-11 had been manually analyzed and assigned the number of visible triplets. However, the supervised machine learning algorithms should be trained with more detailed feedback, informing the visibility of each individual hole.

With the aid of an in-house application, our senior technicians, professionals with at least 10 years of experience, revisited the data and provided an answer (visible/invisible) for each hole. The application basically consisted of a pair of windows displayed side-by-side, as shown in figure 3: one where the technician could click in the holes he/she deemed "detectable"; and the other, a blank screen where red circles would pop up to provide a clue that the mouse click had been effective. A subsequent click in the same region switched the status back to "undetectable"; and so forth.

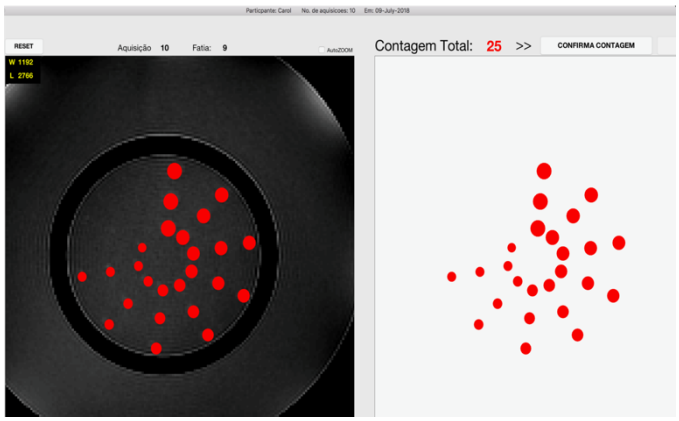

Figure 3. in-house application where the technician could click in the holes he/she deemed "detectable"

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 620x120=74,400 holes were labeled as either "detectable" or "undetectable" by experienced technicians under strictly controlled conditions.

C. *Machine Learning methods*

We tested 5 Machine Learning (ML) methods [9] to predict the responses of the technicians. The dataset was randomly divided into 70% of the entries for the training and 30% for the testing. To avoid overfitting, we set all the parameters using only the training base, and we used the test base only to verify

the performance of the algorithms. Algorithms were programmed and tested using the R language.

Logistic Regression (LR) is the most frequently used method in binary classification problems [10], such as the one tackled in this study. It has been implemented using the standard generalized linear model method in R.

Support Vector Machine (SVM) is a very popular ML classification method and is known to perform quite well in binary classification problems, splitting the feature space with hyperplanes. This algorithm has been implemented using the R's package e1071 [11].

Random Forest (RF) is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes of the individual trees. It has been implemented using R's randomForest package [12].

Extreme Gradient Boosting (XGB) consists of an ensemble of weak prediction models, typically decision trees, and optimizes an arbitrary differentiable loss function. We used R's xgboost package [13].

Neural Network (NNet) with a single hidden-layer of 10 units was also tested. This feed-forward net was implemented using R's nnet package [14].

Table II shows the main parameters we used in each method. For LR method, the only non-default parameter was "family=binomial". We selected the parameters for the XGB method using cross validation with grid search. All other methods were automatically tuned using Caret's Package default search grid.

TABLE II. THE CHOSEN PARAMETERS OF THE ML ALGORITHMS.

Technique	Object class	Parameters
LR	glm	family = binomial
SVM	train	svmRadial tuneLength =10
RF	randomForest	ntree = 500
XGB	xgb.cv	eta = c(0.1,0.7) max_depth = c(0,15) nrounds = c(25,300) max_delta_step = c(0,7) subsample = c(0.5,0.7) objective = "reg:logistic" nthread = 4 verbose = 0 nfold = 10 metrics = "auc"
NNet	nnet	size = 10 decay = 0.001

D. Performance Metrics

We evaluated the performance of the ML algorithms using AUC - the area under the ROC (Receiver Operating Characteristic) curve. AUC has been used in medical diagnostics since the 1970s, and is probably the best index of prediction accuracy available. AUC=1 indicates that predictions are perfectly accurate while AUC=0.5 indicates they are pure guessing.

We also calculated the Krippendorff's alpha (Kripp.alpha) metric, a non-parametric index that measures the agreement between observations [8]. We used R's package *irr* that implements this metric. It yields a value from -1 to 1, where 1 indicates perfect agreement, 0 indicates no agreement beyond chance and negative values indicate inverse agreement.

IV. RESULTS

Table III summarizes the obtained results. The obtained results are substantially better than those previously reported: Ehman et al. [3] obtained Krippendorff's alpha of 0.652 while our best alpha is 0.995.

The method with the largest AUC was LR (logistic regression) with area of 0.878±0.056, where 0.878 is the mean of the areas obtained by 10-fold cross-validation and 0.056 is the standard deviation. Figure 4 shows the ROC curves of LR model for the train and test bases. LR also yielded the the highest Krippendorff's alpha (0.995). It is noteworthy that there is no guarantee that AUC and Krippendorff's alpha will agree that a specific algorithm is the best.

We tried to solve this problem without using machine learning and did not get good results. Thresholding the signal-to-noise ratio did not work well. We also tried to include the area of the hole into the formula without success.

TABLE III. AREA UNDER THE ROC CURVE (AUC) AND KRIPPENDORFF'S ALPHA OF ML TECHNIQUES. THE NOTATION $X\pm Y$ INDICATES MEAN X AND STANDARD-DEVIATION Y OF THE 10-FOLD CROSS-VALIDATION.

	LR	SVM	RF	XGB	NNet
AUC	0.878 ±0.056	0.781 ±0.08	0.873 ±0.086	0.855 ±0.042	0.758 ±0.054
Kripp.alpha	0.995	0.993	0.917	0.750	0.994

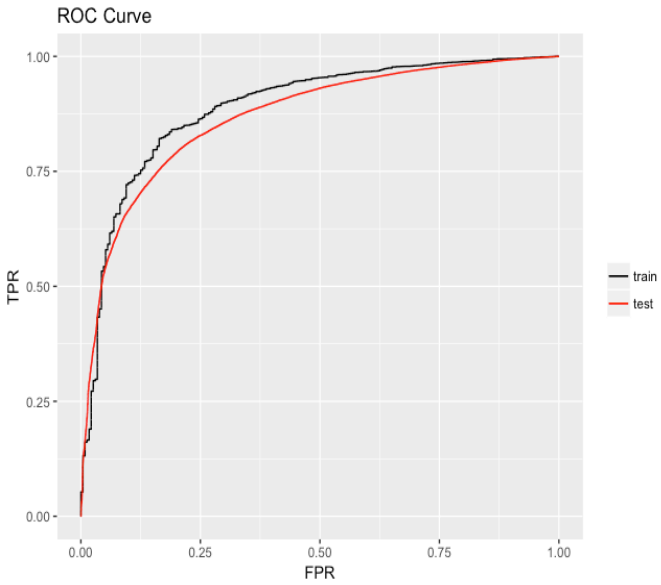

Figure 4. ROC curves of the logistic regression model (LR) for the train and test base

To assess the quality of our method, we compared the answers of junior technicians (with less than 5 years of experience) with our algorithm, considering the answers of senior technicians (with more than 10 years of experience) as "gold standard". The first row of Table 4 indicates that junior technicians classified correctly 82% of all holes; and classified correctly only 34% of undetectable holes and 84% of detectable holes. To measure the performance of our algorithm, we thresholded the output of LR model (that yielded the best results) using criterium "ROC01", that minimizes the distance between ROC plot and point (0,1). The second row of Table IV indicates that LR model classified correctly 84% of all holes, 68% of undetectable holes and 87% of detectable holes. In conclusion, our algorithm is better than junior technicians in classifying the holes as detectable/undetectable.

Table V shows the result of junior technicians and our algorithm applied only to slice 8, the one with the lowest contrast and therefore the most difficult to visualize. Again, our algorithm is better than junior technicians (considering the answers of senior technicians as correct ones).

We note that even senior technicians may disagree on the classification of a hole. As we have, at this moment, only one classification made by a senior technician per hole, we cannot calculate the dispersion of their responses. In the near future, we intend to collect more classifications of our images made by senior technicians, in order to calculate possible dispersion and to obtain a better "gold standard".

TABLE IV. ACCURACY, SENSITIVITY AND SPECIFICITY CONSIDERING THE ANSWERS OF SENIOR TECHNICIAN AS "GOLD STANDARD", APPLIED TO ALL SLICES (8-11). JUNIOR TECHNICIANS ARE THOSE WHO HAVE LESS THAN 5 YEARS OF EXPERIENCE. LR MODEL WAS THRESHOLDED TO MINIMIZE THE DISTANCE BETWEEN ROC PLOT AND POINT (0,1).

Professional experience	**Accuracy**	**Sensitivity**	**Specificity**
Junior technicians	0.824	0.343	0.844
LR model	0.842	0.677	0.868

TABLE V. ACCURACY, SENSITIVITY AND SPECIFICITY CONSIDERING THE ANSWERS OF SENIOR TECHNICIAN AS CORRECT, APPLIED ONLY TO SLICE 8 (THE ONE WITH THE LOWEST CONTRAST).

Professional experience	Accuracy	Sensitivity	Specificity
Junior technicians	0.583	0.560	0.584
LR model	0.690	0.617	0.784

V. CONCLUSIONS

To our knowledge, this is the first attempt to automate the ACR MRI low-contrast resolution test using Machine Learning. We fed five learning algorithms with features extracted from the ACR phantom images, and with labels (detectable/undetectable) assigned by senior technicians with more than 10 years of experience. Among the five methods we tested, Logistic Regression yielded the largest area under the ROC curve (0.878) and the highest Krippendorff's alpha (0.995). The results achieved in this study are substantially better than those previously reported in the literature. Also, the results are better than those obtained when junior technicians (with less than five years of experience) labels the image structures manually. This indicates that it may be possible to replace human operator in ACR low-resolution test.

ACKNOWLEDGMENT

We thank the Hospital Israelita Albert Einstein for assisting us in the analysis of the phantom images, and the Foundation for Research Support of the State of São Paulo (FAPESP) for the financial support (grant n. 2015/27022-0 received by FBT).

REFERENCES

[1] Brown, R. W., Haacke, E. M., Cheng, Y. C. N., Thompson, M. R., & Venkatesan, R. Magnetic resonance imaging: physical principles and sequence design. John Wiley & Sons, 2014.

[2] ACR website: http://www.acraccreditation.org, accessed February 26,2018.

[3] Ehman, Morgan O. et al. Automated low-contrast pattern recognition algorithm for magnetic resonance image quality assessment. Medical physics, v. 44, n. 8, p. 4009-4024, 2017.

[4] Fitzpatrick, Atiba Omari. Automated Quality Assurance for Magnetic Resonance Imaging with Extensions to Diffusion Tensor Imaging. 2005. Doctoral dissertation. Virginia Tech.

[5] Sun, J., Barnes, M., Dowling, J., Menk, F., Stanwell, P., & Greer, P. B. An open source automatic quality assurance (OSAQA) tool for the ACR MRI phantom. Australasian physical & engineering sciences in medicine, v. 38, n. 1, p. 39-46, 2015.

[6] Davids, Mathias et al. Fully-automated quality assurance in multi-center studies using MRI phantom measurements. Magnetic resonance imaging, v. 32, n. 6, p. 771-780, 2014.

[7] Panych, Lawrence P. et al. On replacing the manual measurement of ACR phantom images performed by MRI technologists with an automated measurement approach. Journal of Magnetic Resonance Imaging, v. 43, n. 4, p. 843-852, 2016.

[8] Landis, J. Richard; Koch, Gary G. The measurement of observer agreement for categorical data. biometrics, p. 159-174, 1977.

[9] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. No. 10. New York, NY, USA:: Springer series in statistics, 2001.

[10] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[11] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. e1071: misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien. R package version 1.6-7. 2015.

[12] Liaw, A.; Wiener, M. Classification and regression based on a forest of trees using random inputs. R Package

[13] Chen, Tianqi, Tong He, and Michael Benesty. "xgboost: Extreme gradient boosting. R package version 0.4–4." (2016

[14] Ripley, Brian; Venables, William. nnet: Feed-forward neural networks and multinomial log-linear models. R package version, v. 7, n. 5, 2011.