Machine learning for MRI quality assurance
Aprendizagem de máquina no controle de qualidade de RM

RAMOS, JHONATA E.[*]; TANCREDI, F. B.[†]; KIM, Hae Yong[*]
[*]Escola Politécnica - USP
[†]Centro de Pesquisa em Imagem - Hospital Israelita Albert Einstein
jhonata.emerick@usp.br, felipe.tancredi@einstein.br, hae@lps.usp.br

***Abstract*—Magnetic Resonance Imaging (MRI) is a powerful, versatile and all-important medical imaging method. The image quality of a MRI scanner is assessed from measurements in phantom images (phantom is a test object of known geometry and composition). The American College of Radiology (ACR) accreditation program recommends 7 periodic measures in its multi-purpose phantom. The image acquisition and manual analysis take approximately 30min. Such tests are essential for quality assurance; but also costly. Several automated methods have been proposed, but reports on the low-contrast resolution test are nebulous and results unconvincing. This test is entirely dependent on (the visual perception of) the operator. We propose the use of Machine Learning (ML) to emulate the operator's ability resolving the the 120 low-contrast structures of the ACR phantom, and here report results of a short proof-of-concept study on the new method. We have used 38 sets of images acquired in a 1.5T scanner, which totaled 4,560 structures to be classified between 'detectable' or 'undetectable.' As predictors of detectability, we have utilized 22 image features – extracted from the structure itself, its surroundings and other areas –, such as mean, min and max signals, noise, size and position of the structure, shading, sharpness and ringing levels. An operator labeled each structure as 'detectable' and 'undetectable.' Among the various machine supervised learning methods that we have tested, the xgboost was the one that resulted in the best accuracy: Area Under the Curve of 97.3%. These are preliminary but extremely encouraging results. We believe that after small adjustments in our feature extraction algorithms and training of our classifier with a larger data set we will be able to demonstrate that ACR MRI tests can be fully automated with the aid of ML.**

***Keywords*— MRI; Quality Assurance; Machine Learning.**
***Classification*— doctorate degree.**
***Category*— (Doctorate degree): Beginner**

Resumo—A ressonância magnética (RM) constitui poderoso, versátil e talvez o mais importante método de imagem médica. Assim como a qualidade de geração de imagens de um scanner de RM pode ser avaliada a partir de medidas extraídas de imagens de *phantom* (um objeto teste com geometria e composição conhecidas). O programa de acreditação do Colégio Americano de Radiologia (ACR) recomenda 7 medidas periódicas em seu *phantom* multi-propósito. A aquisição e análise manual das imagens leva aproximadamente 30 minutos. Tais testes são essenciais para o controle de qualidade, porém onerosos. São várias as propostas de automação desse processo. Contudo até hoje nenhum estudo apontou uma solução satisfatória para a medida de resolução em baixo contraste – uma medida que depende exclusivamente (da percepção visual) do operador. Nossa proposta consiste em utilizar aprendizagem de máquina para emular a capacidade do operador de resolver as 120 estruturas de baixo contraste do *phantom* ACR. Para a prova de conceito do novo método utilizamos 38 séries de imagens adquiridas em um equipamento de 1.5T, que nos forneceram um total de 4.560 estruturas a serem classificadas como 'visíveis' ou 'não-visíveis'. Como preditores, utilizamos 22 features de imagem – que extraímos das estruturas, seus entornos e alhures –, tais como sinais mínimo, médio e máximo, ruído, posição e tamanho da estrutura, sombreamento, definição de bordas e nível de *ringing*. Cada estrutura foi classificada por um operador como 'visível' ou 'não-visível'. Entre os diversos métodos de treinamento supervisionado que testamos, o método xgboost ofereceu o melhor resultado: 97.3% de AUC (*Area Under Curve*). São resultados preliminares, mas muito encorajadores. Acreditamos que após alguns ajustes no processo de extração de *features* e treinamento do algoritmo classificador com uma base de dados maior seremos capazes de demonstrar a tese de que o controle de qualidade RM do ACR pode ser totalmente automatizado com auxílio de ML.

***Palavras-chave*— Ressonância Magnética, Controle de Qualidade, Aprendizagem de Máquina.**
***Classificação*—Doutorado.**
***Categoria*—(Doutorado): Iniciante**

I. INTRODUÇÃO

A Ressonância Magnética (RM) é modalidade de imagem médica que oferece a maior gama de contrastes de imagem [1]. Produz imagens onde se podem resolver estruturas ósseas como Raio-X, mas também outras onde se podem resolver estruturas com tênues diferenças de composição (ie. contraste inerentemente baixo), como um pequeno infarto no miocárdio.

Assim como qualquer instrumento de medida, um scanner de RM deve passar por testes de controle de qualidade. E como todo instrumento de imagem médica esse teste consiste em avaliar imagens de um objeto de composição e geometria conhecidas, apelidado de *phantom*. Um scanner com bom desempenho produz imagens fidedignas do *phantom* e – por extrapolação – também do corpo humano.

O programa de controle de qualidade mais conhecido e difundido é aquele preconizado pelo Colégio Americano de Radiologia (ACR) [2]. Aqui no Brasil apenas algumas instituições de ponta como o Hospital Albert Einstein aderem

ao programa; nos EUA, a acreditação pelo ACR é compulsória para quem atende o sistema público de saúde.

O programa de controle de qualidade de RM do ACR prevê 7 testes em seu *phantom* multi-propósito. A aquisição e análise manual das imagens para extração das de qualidade levam aproximadamente 30 minutos. São 25 horas de tempo de máquina e mão de obra dedicada ao programa ACR por ano, por scanner. Um custo que justifica a procura por ferramentas tecnológicas para automação o processo.

Existem algumas propostas de automação do programa de controle de qualidade em RM do ACR [3,4]. Entretanto, a automação de um dos testes permanece um problema em aberto. Encontramos um único estudo na literatura que trata do assunto [5] e os resultados apresentados não são convincentes. Tratamos aqui do teste de resolução de estruturas em baixo contraste, um teste que depende exclusivamente da percepção visual do operador. Nossa proposta consiste da utilização de *Machine Learning* (ML) para emular a tarefa humana de resolver as estruturas de baixo contraste do *phantom* ACR.

II. RESOLUÇÃO EM BAIXO CONTRASTE

O teste de resolução em baixo contraste tem como objetivo averiguar a capacidade do scanner de oferecer imagens com qualidade suficiente para permitir diferenciar estruturas com pouco contraste com relação ao fundo. São adquiridos quatro cortes axiais no *phantom* ACR, na região onde se encontram finas películas de acrílico perfurada, cada uma com 30 furos de diâmetros variados e organizados em 10 *triplets* radiais (Fig.1).

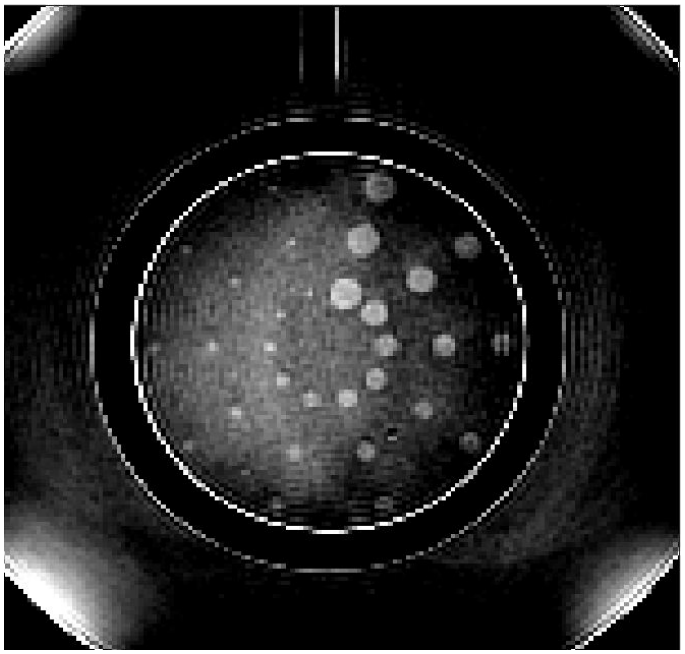

Fig 1 – Phantom do ACR para RM – imagem típica do corte axial 10

Os furos de um mesmo raio têm o mesmo diâmetro, que diminui gradativamente, no sentido horário, indo de 7mm até 1.5mm. As películas de cada tomo possuem diferentes espessuras, que é o que determina o contraste entre o fundo e os furos. Os furos possuem sinal da solução de NaCl/NiCl que preenche o *phantom*, enquanto que o sinal proveniente da região da película (fundo) é uma composição entre o sinal da solução e o sinal da estrutura de acrílico; que varia dependendo da espessura da película. A diferença entre eles – ie. o contraste – diminui progressivamente do corte 11 até o corte 8. A posição dos *triplets* também muda ligeiramente.

O teste de resolução consiste essencialmente em contar quantos dos 10 *triplets* de furos podem ser resolvidos em cada um dos tomos 11 a 8. Inicia-se a contagem a partir da fatia 11, onde o contraste é maior; e dos raios com furos de maior diâmetro para os de menor diâmetro. O *triplet* é considerado visível quando todos os 3 furos que o compõem podem ser claramente visíveis. Quanto melhor a qualidade de imagem, maior o número de raios visíveis.

III. VARIÁVEIS E MODELAGEM

O resultado do teste de baixo contraste do ACR é expresso em termos da contagem de *triplets* de furos. No entanto, o operador avalia a visibilidade de cada furo individualmente. Logo, a modelagem da leitura do operador também é realizada furo a furo. Foram testados alguns classificadores, todos baseados em treinamento supervisionado; e uma base de dados com 4.560 furos. Em todos os testes utilizamos 70% do conjunto de dados para o treinamento e 30% para teste. Imagens de *phantom* foram obtidas em equipamento de 1.5T ao longo de 40 semanas.

A seguir tratamos da definição de variáveis preditivas e da modelagem da variável resposta, ie. leituras do operador. As leituras de visibilidade foram obtidas com aplicativo desenvolvido em Matlab que permite ajustes de imagem semelhantes aos disponíveis na *workstation* do scanner (tais como *zoom*, pan e janelamento) e responde a cliques de mouse nas regiões dos furos para colecionar as leituras: o primeiro clique assinalando variável de resposta 1–'visível' e subsequentes alternando valor, por ex. para 0 – 'não-visível'.

A. Preditores

Uma estrutura é percebida numa imagem (ie. detectada, visualizada) quando estamos seguros de que algo (ie. a estrutura) se destaca do fundo. A percepção visual de estruturas depende tanto do tamanho e sinal da estrutura tanto quanto da relação desses com seus entornos. Isso sem contar os diversos problemas de qualidade de imagem, como sombreamento, distorções, efeitos de borda e de amostragem. Para modelagem da visibilidade dos furos do *phantom* ACR foram extraídos 22 *features* da imagem, localmente, na região do furo, e alhures (Tabela I). A extração de *features* locais foi baseada em mascaramento da imagem original com 3 tipos de máscaras: um consistindo da região do furo propriamente dita; um

segundo representando a região periférica ao furo; e um terceiro representando as demais adjacências. Esses ROIs foram criados através do corregistro de um template dos 30 furos com a fatia 11, onde todos os furos são sempre visíveis. O posicionamento dos ROIs nas demais fatias foi realizada por extrapolação.

Diversas *features* foram extraídas das imagens com a intenção de aumentar o poder preditivo dos modelos de ML, porém nem todas as variáveis são utilizadas no modelo final.

Tabela I. RESUMO DAS VARIÁVEIS PREDITIVAS

Feature	Descrição
SLICE	Fatia onde se encontra a estrutura (8-11)
RADIUS	Raio onde se encontra a estrutura (1-10)
POSITION	Posição da estrutura no raio (1-3)
GHOST	Nível de Ghosting da imagem
RING	Nível de artefato de borda relativo ao raio
EDGE	Definição de bordas da estrutura
S_*	Sinal na região de interesse
N_*	Desvio padrão do Sinal na região de interesse

* As features S_ e N_ são calculadas em diversas regiões de interesse da imagem. S_ ainda recebe índices de mean, min e max.

B. Métricas de Desempenho

Para a validação da capacidade preditiva dos modelos de ML foram usados o teste de Kolmogorov-Smirnov (KS), curva ROC (*Receiver Operating Characteristic*) e o coeficiente de Gini [6]. A estatística de **Kolmogorov-Smirnov (KS)** é uma estatística não paramétrica para testar se as funções de distribuição de probabilidades de dois grupos são iguais. O valor do KS do modelo é a maior diferença entre as distribuições acumuladas das probabilidades dos grupos de furos "visíveis" e "não-visíveis". O valor da estatística pode variar entre 0 e 1, sendo que quanto mais próximo de 1, maior o poder discriminatório do modelo.

A **Curva ROC** é uma métrica para avaliação de modelos, que permite estudar a variação para as medidas de sensibilidade e especificidade, para diferentes pontos de corte.

A área sob a curva ROC (**AUC**) é um método bastante utilizado porque é uma medida global de desempenho independente de limites de corte, geralmente empregados na construção da matriz de confusão. Quanto mais próximo de 1 for a área, melhor o desempenho.

O **coeficiente de Gini** é determinado a partir da construção da curva ROC, definido como quociente entre áreas. Quanto maior o coeficiente maior será a separação entre "visíveis" e "não-visíveis".

C. Modelos e Resultados

A variável resposta é do tipo binária ('visível' / 'não-visível'), assim a técnica de classificação que surge como candidata natural é a regressão logística. Para efeito de comparação, outras duas técnicas de classificação bem populares foram avaliadas: *Support vector machine* (SVM) e XGBOOST. Todos os dados foram analisados no software R e foram utilizadas 4.560 *entries* entre treinamento e teste dos classificadores. Um resumo dos resultados pode ser encontrado na Tabela II.

a) Regressão logística: Candidato natural a um problema de classificação. O modelo de regressão logística binomial está incluso na família de modelo lineares generalizados (i.e., *generalized linear model*). Para a implementação utilizou-se a função glm do R. A regressão logística apresentou um desempenho levemente inferior ao XGBOOST quando se olha o AUC (0.972 vs. 0.971) e o GINI (0.944 vs. 0.943) em ambas as bases de treino e teste, porém o KS da regressão logística ficou consideravelmente abaixo quando comparado ao do XGBOOST (0.840 vs. 0.853).

b) XGBOOST: A implementação desta técnica foi realizada fazendo uso do pacote xgboot do R. O modelo *gradiente boost* (GB) é de regressão aditiva, na qual os termos são árvores decisórias obtidas após simples partição recursiva. O XGBOOST do R se trata de uma implementação mais eficiente e escalável do GB, além de ser uma técnica muito eficaz, que na prática gera ótimos resultados de classificação e tem sido escolhida para solução de para vários problemas de *Machine Learning*. No presente estudo a técnica apresentou desempenho superior em relação às outras duas testadas, se mesmo que a diferença seja modesta quando comparada à de regressão logística: AUC de 0.973 vs. 0.971. A Figura 2 mostra a curva ROC (na realidade as duas obtidas das amostras de treino e teste). O fato das curvas estarem próximas do canto superior esquerdo do diagrama, o que confere maior área sob a curva, demonstra seu grande poder discriminante.

c) Support Vector Machine (SVM): O processo decisório em problemas de reconhecimento de padrões pode ser realizado através de funções que dividem o espaço de características (*features*) em regiões. A forma mais simples de fazer isso é através de hiperplanos. SVM baseia-se nessa estratégia ao construir um tipo especial de hiperplano, o Hiperplano de Margem Máxima. Aqui a implementação foi feita utilizando o pacote e1071 do R. Dentre todas as técnicas testadas foi a que apresentou o pior desempenho de acordo com todas as métricas. Este modelo precisa ser revisto em relação ao *overfitting*.

Tabela II. DESEMPENHO DOS MODELOS DE CLASSIFICACAO

Base/Métrica	KS	GINI	AUC
Regressão Logística			
Treino	0.840	0.944	0.972
Teste	0.853	0.943	0.971
SVM			
Treino	0.978	0.975	0.987
Teste	0.696	0.651	0.825
XGBOOST			
Treino	0.960	0.991	0.995
Teste	0.907	0.946	0.973

IV. CONCLUSÃO

Este trabalho analisou o desempenho de 3 métodos de *Machine Learning*, com o objetivo de modelar a capacidade de percepção humana em uma tarefa específica: resolver estruturas de baixo contraste em imagens de *phantom* de RM, isto é de classificar estruturas como visíveis ou não-visíveis em condições de baixo contraste. Entre os 3 modelos testados, a estratégia XGBOOST apresentou resultados superiores. Vários atributos do modelo podem ainda ser alterados na tentativa de melhorar seu desempenho, como por exemplo, taxa de aprendizado, profundidade da árvore decisória.

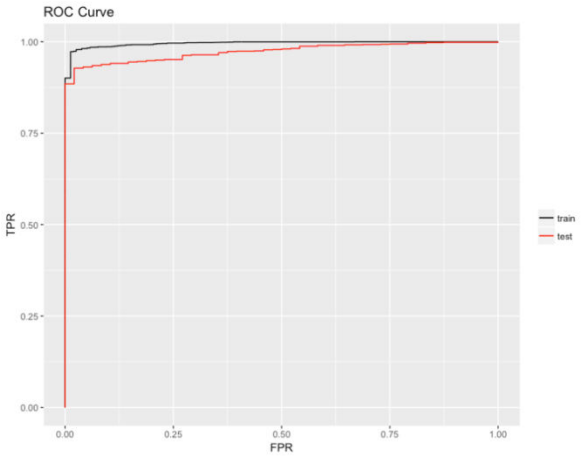

Fig 2 - Curva ROC DO MODELO XGBOOST

Como futuro desenvolvimento pretendemos realizar ajustes no processo de extração de *features*, e usar uma base de dados 100 vezes maior, que leve em conta mais imagens e leitores. Com esses ajustes esperamos que o método se torne pelo menos 200 vezes mais confiável (ie. não cometa mais do que 1 erro de leitura a cada 10 anos de teste) para com isso demonstrar que com ajuda de técnicas de aprendizagem de máquina os testes de controle de qualidade de RM do ACR podem ser completamente automatizados.

AGRADECIMENTOS

Agradecemos ao Hospital Israelita Albert Einstein, pela permissão de análise das imagens de *phantom*, e à Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP, pelo apoio financeiro ao projeto (processo nº 2015/27022-0, auxílio recebido por FBT).

REFERÊNCIAS

[1] BROWN, Robert W. et al. Magnetic resonance imaging: physical principles and sequence design. John Wiley & Sons, 2014.

[2] Site do ACR: http://www.acraccreditation.org

[3] M-n 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 Imaging, pp. n/a–n/a, Sep. 2015.

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

[5] 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," *Australasian Physical & Engineering Sciences in Medicine*, vol. 38, no. 1, pp. 39–46, Oct. 2015.

[6] ZHANG, Lingling et al. The Measurement of Distinguishing Ability of Classification in Data Mining Model and Its Statistical Significance. In: International Conference on Computational Science. Springer, Berlin, Heidelberg, 2009. p. 578-587.