JHONATA EMERICK RAMOS

Automatização do teste de baixo contraste do colégio americano de radiologia

São Paulo
2022

JHONATA EMERICK RAMOS

Automatização do teste de baixo contraste do colégio americano de radiologia

Versão Corrigida

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Doutor em Ciências.

Área de Concentração:
Engenharia de Computação

Orientador:
Prof. Dr. Hae Yong Kim

São Paulo
2022

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, **23** de **Novembro** de **2022**

Assinatura do autor: Jhonata F. Bunos

Assinatura do orientador: Ana Maria Kim

Catalogação-na-publicação

Ramos, Jhonata
Automatização do teste de baixo contraste do colégio americano de radiologia / J. Ramos -- versão corr. -- São Paulo, 2022.
101 p.

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Aprendizado de máquina 2.Teste de baixo contraste 3.Ressonância magnética 4.Colégio americano de radiologia I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

AGRADECIMENTOS

Aos meus pais Juarez Vieira e Maria Celina Emerick, meus irmãos: Jefferson Emerick Ramos, Jordão Emerick Ramos e Jennifer Emerick Ramos.

Ao professor e amigo Dr. Hae Yong Kim, por todos os ensinamentos e orientações que foram fundamentais no desenvolvimento deste trabalho e na minha formação acadêmica, e também pela paciência em conciliar o doutorado com as tarefas empresariais.

A todos os professores da Universidade de São Paulo (USP), em especial os da Escola Politécnica (POLI/USP) e Escola de Engenharia de São Carlos (EESC/USP), que sempre estiveram presentes em todos esses anos de USP. A USP foi fundamental para eu encontrar meu caminho: sou muito grato. Espero e vou me esforçar para poder retribuir à sociedade tudo o que me foi dado.

Ao meu sócio e amigo Felipe Brunetto Tancredi pela inspiração acadêmica que ajudou a dar origem a este trabalho, além da motivação de sempre seguir em frente, mesmo nos momentos mais tortuosos.

Ao meu sócio e amigo Carlos Relvas pela compreensão e pelas muitas vezes que parou o que estava fazendo para me apoiar; fico sem palavras para agradecer.

Ao meu sócio e amigo Felipe Barjud Pereira do Nascimento pelas ideias e provocações para tentar nos fazer melhores como pessoas e empresa. Meu muito obrigado.

Ao meu sócio e amigo Gustavo Bernardo por toda a compreensão durante a construção deste trabalho, e pela ajuda de sempre. A todos os colaboradores da *Datarisk*, que me ensinam diariamente e me ajudam a sonhar grande.

A todos os meus amigos de graduação e pós-graduação que sempre me ensinaram muito.

Enfim, a todos que me apoiaram e apoiam, só tenho a agradecer, pois sempre estive sobre os ombros de gigantes.

RESUMO

RAMOS, J. **Automatização do teste de baixo contraste do colégio americano de radiologia**. 2022. 96p. Dissertação (Doutorado) - Escola Politécnica da Universidade de São Paulo, São Paulo.

A Ressonância Magnética (MRI - do inglês *Magnetic Ressonance Imaging*) é uma modalidade de imagem médica poderosa, difundida e indispensável. O ACR (American College of Radiology) recomenda que o desempenho das máquinas de ressonância magnética seja monitorado, repetindo os testes de qualidade de imagem a cada 7 dias ou menos. Testes de qualidade são realizados em imagens de um objeto de geometria e composição conhecidas, denominado *phantom*. Uma máquina em bom estado deve gerar imagem que retrate a anatomia sob inspeção com as dimensões e características corretas, e permitir a detecção de pequenos furos em condições de baixo contraste. Alguns métodos automatizados foram propostos na literatura, mas a automação de dois dos testes do ACR, de alto e baixo contraste, continua sendo um problema em aberto. Esta tese apresenta uma proposta para automatizar o teste de baixo contraste do ACR. Este teste, geralmente, é feito por técnicos analisando a imagem de *phantom*. No entanto, a análise automatizada seria capaz de reduzir custos, melhorar a repetibilidade e confiabilidade das medidas de controle de qualidade. Os trabalhos sobre automação dos testes de baixo contraste do ACR são escassos e, até onde sabemos, nenhum deles produziu resultados robustos o suficiente que permita substituir o trabalho humano.

Podemos separar esta tese em duas fases principais. Na primeira, consideramos as respostas dos técnicos seniores, com mais de 10 anos de experiência, como nosso padrão ouro. Utilizamos um banco de dados com 620 conjuntos de imagens *phantom* ACR, que foram adquiridos em máquinas de diferentes fornecedores, campos e bobinas, totalizando 74.400 furos de baixo contraste. Técnicos com mais de 10 anos de experiência rotularam cada furo como visível ou invisível. Os algoritmos de aprendizado de máquina foram alimentados com características obtidas manualmente, com objetivo de extrair informações dos furos e seus arredores. Entre os cinco métodos testados, a regressão logística apresentou a maior área sob a curva ROC (0,878) e o maior alfa de *Krippendorff* (0,995). Os resultados alcançados nesta fase do trabalho já são substancialmente melhores do que os relatados anteriormente na literatura. Também são melhores do que as classificações feitas por técnicos juniores, com menos de 5 anos de experiência. Estes primeiros resultados já são um indicativo de que o teste de resolução de baixo contraste, ACR MRI, pode ser automatizado usando as técnicas de aprendizado de máquina.

Aprendizado profundo (*deep learning*) faz parte de uma família de métodos de

aprendizado de máquina baseados em redes neurais artificiais. Entre as técnicas de aprendizado profundo, redes neurais convolucionais têm sido usadas na classificação de imagens, produzindo resultados comparáveis e, em alguns casos, superiores aos de especialistas humanos. Na segunda fase, usamos redes neurais convolucionais para emular a detecção de furos, estruturas de baixo contraste, em um *phantom* ACR. Analisando cuidadosamente nosso conjunto de dados, concluímos que os técnicos seniores cometem tantos equívocos quanto os técnicos menos experientes e, portanto, anos de experiência não garantem, por si só, maior precisão na tarefa de classificação. Assim, na segunda fase da tese, alteramos o padrão ouro, de visibilidade dos furos para a mediana das respostas de todos os técnicos que não cometeram equívocos na classificação dos furos na imagem, independente dos anos de experiência. Utilizamos um subconjunto de 100 aquisições de *phantom* ACR da fase anterior, totalizando 12.000 furos. Para obter robustez estatística, repetimos o treinamento e o teste 5 vezes, usando validação cruzada de 5 vezes (*5-fold cross validation*). Obtivemos uma AUC média (*Area Under ROC Curve*) de 0,983±0,003 e uma acurácia média de 93,2±0,7% no ponto de EER (*Equal Error Rate*). Aplicando o modelo obtido a um conjunto de dados de teste completamente independente com 10.800 furos, obtivemos uma AUC de 0,979. As previsões do nosso modelo na classificação dos *spokes* - conjunto de três furos do mesmo diâmetro, alinhados radialmente - concordam em 93,83% dos casos com a mediana das respostas dos técnicos. Esses resultados são melhores do que as respostas de qualquer técnico individualmente. Concluímos que nosso sistema pode substituir o técnico humano no teste de baixo-contraste do ACR e ainda pode fornecer respostas em tempo real, para ajudar no treinamento de novos técnicos envolvidos no processo.

Palavras-Chave: Aprendizado de máquina, rede neural convolucional, teste de baixo contraste, ressonância magnética, inteligência artificial, colégio americano de radiologia, percepção visual, qualidade de imagem.

ABSTRACT

RAMOS, J. **Automatização do teste de baixo contraste do colégio americano de radiologia**. 2022. 96p. Dissertation (Doctorate) - Escola Politécnica da Universidade de São Paulo, São Paulo.

Magnetic Resonance Imaging (MRI) is a powerful, widespread and indispensable medical imaging modality. The American College of Radiology (ACR) recommends that the performance of MRI machines be monitored by repeating the image quality tests every 7 days or less. Quality tests are performed on images of an object of known geometry and composition called phantom. A machine in good condition must generate an image that depicts the anatomy under inspection with the correct dimensions and characteristics; and allow the detection of small structures under low contrast conditions. Some automated methods have been proposed in the literature, but the automation of two of the ACR tests, high and low contrast, remains an open problem. This thesis presents a proposal to automate the ACR low contrast test. This test is usually done by technicians analyzing the phantom image, but automated analysis would reduce costs, improve repeatability and reliability of quality control measures. Work on automating ACR's low-contrast tests is scarce and, as far as we know, none of them has produced results robust enough to replace human work.

We can separate this project into two main phases. In the first one, we consider the answers of senior technicians, with more than 10 years of experience, as our "gold standard". We used a database with 620 sets of ACR phantom images that were acquired on machines from 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 manually designed features to extract information from structures and their surroundings. Among the five methods tested, Logistic Regression presented the largest area under the ROC curve (0.878) and the largest Krippendorff alpha (0.995). The results achieved in this phase of the work are already substantially better than those previously reported in the literature. They are also better than classifications made by junior technicians (less than 5 years of experience). These early results are already an indication that the ACR MRI low-contrast resolution test can be automated using machine learning techniques.

Deep learning is part of a family of machine learning methods based on artificial neural networks. Among deep learning techniques, convolutional neural networks have been used in image classification, producing results comparable to, and in some cases superior to, those of human experts. In the second phase, we use convolutional neural networks to emulate the detection of low contrast structures ("holes") in an

ACR phantom. Carefully analyzing our dataset, we concluded that senior technicians make as many mistakes as less experienced technicians, and therefore years of experience do not, by themselves, guarantee greater accuracy in the classification task. Thus, in the second phase, we changed the "gold standard" from hole visibility to the median of the responses of all technicians (regardless of years of experience) who did not make gross mistakes in classifying the holes in the image. We used a subset of 100 phantom ACR acquisitions from the previous phase, totaling 12,000 holes. For statistical robustness, we repeated training and testing 5 times, using 5-fold cross validation. We obtained an average AUC (*Area Under the ROC Curve*) of 0.983±0.003 and an average accuracy of 93.2±0.7% at the EER point (*Equal Error Rate*). Applying the model obtained to a completely independent test dataset with 10,800 structures, we obtained an AUC of 0.979. The predictions of our model in the classification of spokes (sets of 3 holes) agree in 93.83% of cases with the median of the technicians' answers. These results are better than the answers of any individual technician. We conclude that our system can replace the human technician in the ACR low-contrast test and can still provide real-time answers to help in the training of new technicians involved in the process.

Keywords: Machine learning, convolutional neural network, low contrast test, magnetic resonance imaging, artificial intelligence, American College of Radiology, visual perception, image quality.

"Transire suum pectus mundoque potiri "

-Frase cunhada no verso da medalha fields, que em uma tradução livre significa: superar os limites da inteligência e conquistar o universo-

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACR - (American College of Radiology) Colégio Americano de Radiologia

MRI - (Magnetic Resonance Imaging) Ressonância Magnética

IA - (Artificial Intelligence) Inteligência Artificial

ML - (Machine Learning) Aprendizado de Máquina

CNN - (Convolutional Neural Network) Rede Neural Convolucional

ROI - (Region Of Interest) Região de Interesse

LGPD - Lei Geral de Proteção de Dados

CQ - (Quality Assurance) Controle de Qualidade

PIU - (Percent Image Uniformity) Percentagem Integral de Uniformidade

SUMÁRIO

1 INTRODUÇÃO

As pesquisas em saúde têm proporcionado tratamentos cada vez mais avançados para diversas doenças. Os equipamentos utilizados para aquisição de dados digitais como: imagens visíveis, imagens infravermelho, ressonância magnética, áudio e outras fontes, têm possibilitado a obtenção de dados relevantes para diagnósticos (WARING; LINDVALL; UMETON, 2020). Além disso, estes dados permitem o desenvolvimento de processos de automação com foco em otimização. Técnicas de Inteligência Artificial (IA) e Aprendizado de Máquina (ML - sigla do inglês *Machine Learning*) (BONACCORSO, 2017) são cada vez mais exploradas em laboratórios e centros de pesquisa que lidam com doenças complexas como cânceres, problemas neurológicos e automatização de diagnóstico.

A ressonância magnética é um método não invasivo que gera imagens 2-D ou 3-D da anatomia ou processos fisiológicos do corpo. (BROWN et al., 2014) Os equipamentos de ressonância magnética usam radiação eletromagnética não-ionizante - campos magnéticos fortes, ondas de rádio e gradientes de campo - para gerar imagens do interior do corpo. Essa tecnologia oferece uma enorme variedade de contrastes nas imagens sem usar agentes de contraste ou radiação ionizante. Este equipamento pode ser programado para produzir imagens que revelam fraturas em ossos, como raio-X, mas também pode ser preparado para destacar as diferenças entre músculo e gordura, ou registrar estruturas com diferenças tênues de composição em relação

ao seu entorno. Essa capacidade de distinguir pequenos furos em baixo contraste o torna particularmente útil na avaliação de lesões de menisco, infartos do miocárdio, câncer de próstata, endometriose, para citar alguns exemplos.

Assim como outros instrumentos médicos, a máquina de ressonância magnética deve ser rotineiramente submetida a um controle de qualidade, para garantir que o dispositivo esteja gerando imagens de acordo com suas especificações e que atenda os padrões de qualidade, como os recomendados pelo ACR. A instituição possui um extenso programa de controle de qualidade e emite certificados de adequação em todas as modalidades de imagens médicas, incluindo as de ressonância magnética. Nos Estados Unidos, os testes do ACR fazem parte das normas regulatórias nacionais, enquanto no resto do mundo, os testes são adotados por instituições que reconhecem a importância do monitoramento da qualidade das imagens radiológicas produzidas, como parte das boas práticas.

O ACR recomenda que o desempenho dos *scanners* de ressonância magnética seja monitorado repetindo os testes de qualidade de imagem a cada 7 dias ou menos. Desvios nos índices de qualidade indicam que as imagens clínicas geradas pelo dispositivo podem estar comprometidas e que ele precisa de calibração ou manutenção. Testes de qualidade são realizados em imagens de um objeto de geometria e composição conhecidas, denominado *phantom* (RADIOLOGY et al., 2015). Os testes incluem medidas de distorção, contraste e resolução, entre outros que serão descritos em detalhes no próximo capítulo. Uma imagem de boa qualidade deve retratar a anatomia sob inspeção com as dimensões e características corretas, e permitir a detecção de pequenos furos em condições de baixo contraste.

Os testes que se baseiam em medidas diretas tendem a ser consensuais e objetivos, de modo que a sua automação pode ser realizada por meio de técnicas

convencionais de processamento de imagens. Os testes de baixo e alto contraste, por outro lado, dependem inteiramente da percepção visual do operador. Nesses testes, o operador deve indicar se um determinado conjunto de furos do *phantom* pode ser detectado na imagem, ou seja, diferenciado em relação ao fundo. Esses testes envolvem uma avaliação bastante subjetiva, na medida em que são um reflexo direto da percepção visual humana e da técnica de manipulação do sinal de imagem, que varia entre os operadores, tornando a automação desses testes bastante desafiadora.

Se esses dois testes, baixo e alto contraste, pudessem ser automatizados, provavelmente todo o teste ACR poderia ser realizado sem a presença de um técnico experiente, reduzindo custos envolvidos no processo e melhorando a repetibilidade. As avaliações realizadas pelos técnicos no teste de resolução de baixo contraste são menos consensuais. Este teste é o que permite monitorar o desempenho do *scanner* na geração de imagens contrastadas de tecidos moles, uma marca registrada desta modalidade de imagem. A Figura 1 tem como objetivo facilitar o entendimento do leitor de todo o processo previamente descrito.

Nosso grupo investigou novos métodos para automatizar o teste de resolução de baixo contraste do programa do ACR. Na primeira fase, extraímos manualmente características das imagens de teste, em seguida utilizamos estas informações como entrada para algoritmos convencionais de aprendizado de máquina. Apesar dos resultados terem sido animadores, a precisão dos algoritmos não permitiu substituir completamente o técnico humano. Na segunda fase, investigamos uma nova alternativa para automatizar o teste. É bem conhecido que a rede neural convolucional (CNN - sigla do inglês *Convolutional Neural Network*) é capaz de criar automaticamente filtros de baixo nível mais apropriados para extrair as características mais relevantes, e assim detectar e classificar melhor os objetos em uma imagem (LECUN

Figura 1: Visão resumida dos processos e responsáveis envolvidos nos testes do Colégio Americano de Radiologia (ACR)

et al., 1989; KRIZHEVSKY; SUTSKEVER; HINTON, 2017; LECUN; BENGIO; HINTON, 2015). Avaliamos o desempenho de uma CNN simples na detecção de pequenos furos de baixo contraste nas imagens do *phantom* ACR.

1.1 Banco de Dados

Para a realização deste trabalho, uma base de dados foi estruturada a partir de 620 aquisições de imagens do *phantom* ACR realizadas ao longo de 12 meses, em *scanners* de ressonância magnética de diversos fabricantes (*Siemens*, *GE* e *Philips*), diferentes modelos (por exemplo, *narrow* e *wide bore*), campos magnéticos (1.5T e 3.0T) e com variadas antenas de recepção (eg. 8, 12, e 32 canais). Cada aquisição consiste de 4 imagens 256×256 *pixels*, e cada imagem possui 30 furos de baixo contraste, totalizando 74.400 furos capturados em uma ampla variedade de condições. As imagens obtidas do *phantom* mostram uma pluralidade de furos que podem ser discernidos.

Da literatura, sabemos que um ser humano só consegue distinguir 700-900 níveis de cinza, mesmo em condições ideais (KIMPE; TUYTSCHAEVER, 2007). Como uma imagem de ressonância magnética tem 12 bits ou 4096 níveis de cinza, não é possível distinguir todos os tons em uma mesma imagem, mesmo com o ajuste adequado de brilho/contraste (também conhecido como janelamento). Desta forma, o técnico não avalia a visibilidade de um furo em uma imagem estática, mas fica alterando dinamicamente o brilho/contraste da imagem para verificar se o furo se torna visível sob certas configurações.

Os algoritmos de aprendizado de máquina supervisionados devem ser treinados com *feedback*, informando a visibilidade de cada furo individualmente. Usando um programa desenvolvido em MATLAB (MATLAB; SIMULINK,), os nossos técnicos deram uma resposta (visível/invisível) para cada furo. A aplicação consistia basicamente em um par de janelas exibidas lado a lado (Figura 2), onde o técnico podia clicar nos furos que considerasse visíveis; e do outro lado, uma tela em branco onde apareceriam círculos vermelhos para indicar ao profissional se o clique com o mouse foi efetivo. Um clique subsequente na mesma região muda o status de volta para invisíveis; e assim por diante.

Durante o processo de rotulação dos furos, as fatias foram apresentadas aos técnicos em ordem decrescente de contraste, ou seja, da 11 para a 8. Os técnicos avaliaram um lote com 10 aquisições por dia, em alguns poucos casos avaliaram dois lotes no mesmo dia, com um intervalo mínimo de 2 horas entre as sessões. A ferramenta desenvolvida possui funções de janelamento que são comuns na rotina destes profissionais. Todas as sessões ocorreram na mesma sala escura e usando um único monitor com ajustes fixos. Como resultado, obtivemos um total de 74.400 furos rotulados como visíveis ou invisíveis, por técnicos experientes, sob condições estritamente controladas.

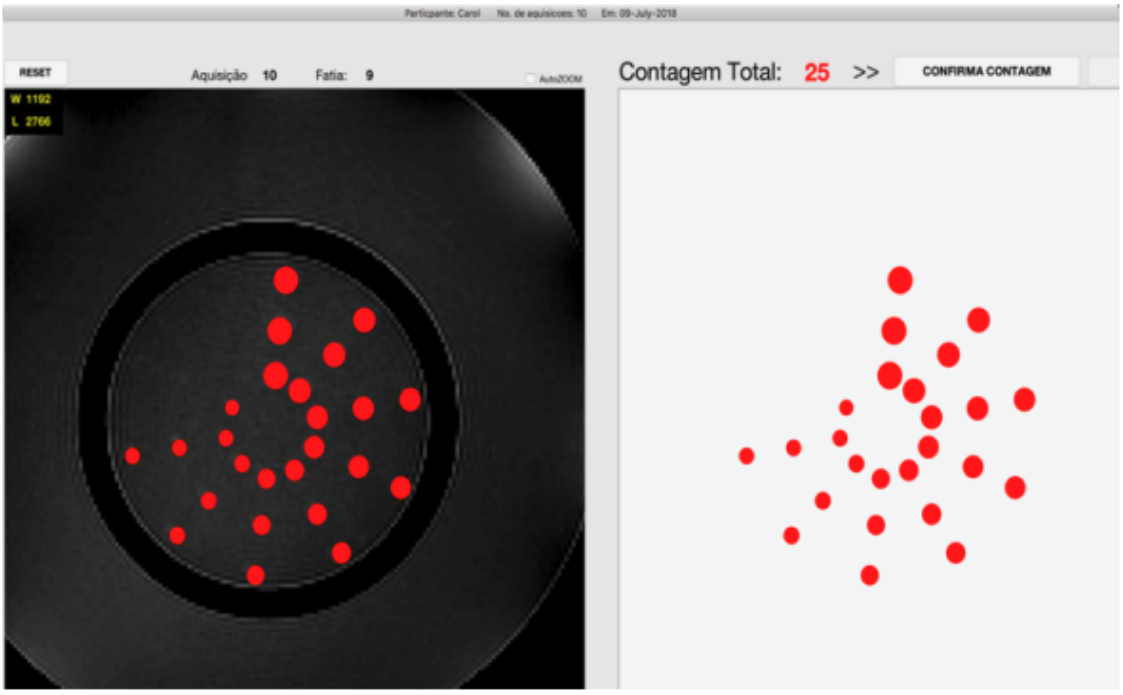

Figura 2: Aplicação desenvolvida internamente onde os técnicos podem clicar nos furos que consideram visíveis.

Para a segunda fase, utilizamos um subconjunto de 100 aquisições da base de dados da fase anterior. Como cada aquisição consiste de 4 imagens e cada imagem possui 30 furos, temos 12.000 imagens ROI (*Region Of Interest*) com 17×17 *pixels*. A escolha da região de interesse com esta dimensão, foi feita para poder conter o furo por inteiro, mesmo os grandes, e para conter um pouco do *background* nas bordas. É necessário que o *background* esteja contido dentro do ROI para obtermos estatísticas desta região. Escolhemos um número ímpar de *pixels*, para que seja possível identificar o pixel central.

1.2 Revisão Literatura

Tomógrafos são instrumentos que fornecem dados essenciais para o diagnóstico de doenças e, portanto, imprescindíveis na medicina moderna. A tomografia, seja ela em 2D ou 3D, por ressonância magnética é particularmente útil porque oferece

uma gama enorme de contrastes de imagem prescindindo do uso de agentes de realce para produzir imagens que revelam fraturas em ossos, como na modalidade de raios-x, mas também podem ser preparadas para destacar as diferenças entre músculo e gordura; ou para registrar furos com tênues diferenças de composição em relação aos seus entornos.

Do mesmo modo que outros instrumentos médicos, o tomógrafo de MRI deve ser rotineiramente submetido a testes de qualidade, o que visa garantir que o tomógrafo esteja gerando imagens conforme suas especificações que satisfaçam minimamente os padrões de qualidade, como aqueles propostos pelo ACR.

O limiar de detectabilidade visual é uma manifestação da percepção humana, e modelos empíricos o descrevem muito bem quando as imagens são nítidas e livres de artefatos. A percepção humana em casos complexos requer modelos mais sofisticados. O método proposto por Fitzpatrick (FITZPATRICK, 2005) para automatizar o teste de detectabilidade de baixo contraste do ACR é derivado do modelo de percepção visual de Rose (ROSE, 1946) e exemplifica a dificuldade em prever a capacidade de detecção humana em um cenário real usando um modelo simplificado.

Davids et al. (DAVIDS et al., 2014) implementaram métodos totalmente automáticos para avaliar medidas de qualidade das imagens de MRI. No entanto, eles não implementaram o teste de detecção de baixo contraste do ACR.

Sun et al. (SUN et al., 2015) descreveram uma ferramenta, de código aberto, para avaliação automática do teste de qualidade do programa do ACR. Para o teste de baixo contraste, implementaram um módulo para avaliar o limiar (*threshold*) de detecção visual, que é específico, segundo os autores, para cada usuário e monitor de computador. O artigo não relata a concordância entre as respostas de seu sistema e os observadores humanos.

Panych et al. (PANYCH et al., 2016) descreveram uma solução para automatizar o teste de alto contraste do ACR. No entanto, eles recomendam que o teste de baixo contraste ainda deve ser realizado por um ser humano.

Ehman et al. (EHMAN et al., 2017) desenvolveram um algoritmo baseado em lógica *fuzzy* para automatizar o teste de resolução de baixo contraste, mas obtiveram uma baixa correlação entre as respostas dos técnicos e da solução desenvolvida. Utilizando a métrica alfa de *Krippendorff* (KRIPPENDORFF, 2011), índice não paramétrico que mede a concordância, encontraram uma correlação de 0,652 entre os técnicos e a solução proposta, o qual é modesto e não permite utilizar o método em substituição a um profissional treinado.

Alaya et al. (ALAYA; MARS, 2020) estimaram visibilidade dos furos desenhando o perfil de intensidade de sinal em círculos de raio variável, identificando picos de sinal, calculando o contraste a partir da sua diferença e aplicando um limiar que não foi especificado. Os autores compararam as saídas do programa com as leituras humanas a partir da contagem total de furos, e não fazendo uma avaliação da concordância de leituras furo a furo. Esta abordagem é ainda mais simples do que o nosso trabalho da primeira fase (RAMOS; KIM; TANCREDI, 2018), dado que utilizaram os níveis de sinal dentro e fora do furo, porém, não fizeram uso dos níveis de ruído. Ademais, os autores aplicam um valor de corte no contraste (que talvez tenha sido otimizado para o conjunto de dados, mas isso não está descrito), sem aplicar algoritmos de aprendizagem para realmente simular a resposta do operador.

Doi et al. (TERAMOTO et al., 2021) desenvolveram um método baseado em CNN para avaliar a resolução de baixo contraste de imagens de tomografia computadorizada. Uma modalidade de imagem completamente diferente da imagem de MRI. Os autores se concentram na avaliação da qualidade da reconstrução da

imagem de tomografia computadorizada, visando o desenvolvimento de novos algoritmos neste campo de pesquisa. Desta forma, não está diretamente relacionado com a automação do teste de baixo contraste do ACR.

Até onde sabemos, a automação do teste de baixo contraste do ACR ainda é um problema em aberto.

1.3 Objetivos, motivação e justificativa

A área da saúde pode se beneficiar das modernas técnicas de inteligência artificial, em especial devido a quantidade de dados que estas organizações coletam diariamente. Tais técnicas, podem apoiar desde a organização da informação até a identificação de quais dados são mais relevantes para a resolução de problemas. Assim, a principal motivação deste trabalho é apresentar uma solução para o teste de baixo contraste do Colégio Americano de Radiologia, utilizando técnicas clássicas e de aprendizagem profunda (*deep learning*), mais especificamente:

- A partir de um banco de dados com imagens oriundas de exames de ressonância magnética (MRI) de um objeto com composição e geometria conhecidas, o *phantom*, desenvolver uma solução de inteligência artificial. Uma solução que seja capaz de emular a percepção visual de um técnico experiente nas avaliações realizadas dentro do contexto do teste de resolução de baixo contraste do ACR.

- Comparar a eficiência do método proposto com a análise realizada por técnicos experientes que atuam na área.

- Propor novos métodos para automatizar o teste de baixo contraste do programa do Colégio Americano de Radiologia (ACR).

1.4 Proposta de abordagem

A proposta deste trabalho de doutorado é de desenvolver uma solução que permita a automatização do teste de baixo contraste do Colégio Americano de Radiologia (ACR). Para tal desenvolvimento, algumas abordagens que fizemos foram essenciais, sendo elas:

- **Parceria com a área da saúde**: para fazer uma pesquisa na área da saúde é fundamental estar inserido em um ambiente que disponibilize equipamentos e equipe capacitada em relação ao objeto de pesquisa. Sendo assim, firmamos parceria com um dos maiores hospitais da América Latina, o Hospital Albert Einstein, localizado em São Paulo, no Brasil;

- **Base de dados**: nesta etapa é feita a coleta e/ou extração da base de dados. De acordo com a Lei Geral de Proteção de Dados (LGPD), o processamento de dados sensíveis, assim como os dados da saúde das pessoas, devem ser tratados para o uso acadêmico de forma anonimizada. O objetivo final desta etapa é obter um banco de dados adequado para o treinamento dos algoritmos;

- **Modelagem**: nesta etapa são realizados os testes de diversas técnicas e/ou arquiteturas, com objetivo de alcançar resultados robustos, que suportem as melhores tomadas de decisão;

- **Resultados**: apresentar os resultados com qualidade e de uma maneira que a comunidade médica e científica consigam reproduzir, permitindo a evolução dos tópicos aqui abordados.

1.5 Principais contribuições

A máquina de ressonância magnética, como outros instrumentos médicos, deve ser rotineiramente submetida a controles de qualidade, para garantir que o dispositivo está gerando imagens de acordo com suas especificações. O Colégio Americano de Radiologia (ACR), possui um extenso programa de controle de qualidade e emite certificados de adequação em todas as modalidades de imagens médicas, incluindo as de ressonância magnética. Houveram algumas tentativas de automatizar os testes de imagem de ressonância magnética no contexto do ACR (DAVIDS et al., 2014; EHMAN et al., 2017; FITZPATRICK, 2005; PANYCH et al., 2016; SUN et al., 2015; ALAYA; MARS, 2020), mas até onde os autores conhecem, nenhum dos métodos propostos foi reconhecido pelo Colégio Americano de Radiologia. As avaliações realizadas pelos técnicos no teste de resolução de baixo contraste, geralmente, são menos consensuais. Este teste é o que permite monitorar o desempenho do *scanner* na geração de imagens contrastadas de tecidos moles, por exemplo. Desta forma, se este teste puder ser automatizado, provavelmente todo o teste do ACR poderia ser realizado sem a presença de um técnico experiente, reduzindo custos e melhorando a repetibilidade. Além disso, aparentemente este é o primeiro trabalho que realmente consegue emular a percepção de um observador humano no contexto de avaliação de furos de baixo contraste em uma imagem digital.

1.6 Componentes da tese

Esta tese está organizada em 6 capítulos:

- **Capítulo 1**: "Introdução", trata sobre a construção da base de dados que utilizamos nesta pesquisa. Descrevemos os objetivos, motivação e justificativa.

Além disso, detalhamos as principais contribuições deste trabalho.

- **Capítulo 2**: "Testes e procedimentos recomendados pelo Colégio Americano de Radiologia (ACR)", trata sobre os aspectos técnicos do programa de acreditação do ACR. Também são abordados procedimentos de controle de qualidade recomendados pelo ACR, que fornece um *design* próprio para o *phantom* e instruções detalhadas sobre os testes a serem realizados.

- **Capítulo 3**: "Métodos Clássicos", nesta primeira fase extraímos manualmente características das imagens e utilizamos algoritmos clássicos de aprendizado de máquina para emular a percepção de um observador humano no contexto do teste de baixo contraste do ACR.

- **Capítulo 4**: "Métodos de aprendizagem profunda", nesta segunda fase investigamos uma nova alternativa para automação do teste, fazendo uso de uma rede neural convolucional. Avaliamos o desempenho de uma CNN na detecção de pequenos furos de baixo contraste nas imagens do *phantom* ACR.

- **Capítulo 5**: "Discussões", neste capítulo apresentamos os resultados obtidos, além de discutirmos as vantagens e os ganhos obtidos com a abordagem utilizada na segunda fase do projeto em relação à primeira.

- **Capítulo 6**: "Conclusões", apresentamos as conclusões a partir dos resultados alcançados, além de revisitarmos os objetivos que foram inicialmente propostos.

1.7 Publicações

As publicações associadas à minha pesquisa de doutorado são:

1. Jhonata E. Ramos, F. B. Tancredi, Hae Yong Kim, “Aprendizagem de máquina no controle de qualidade de RM”, Workshop de Pós Graduação Engenharia de Computação, Escola Politécnica, USP, 2017.

2. RAMOS, Jhonata E.; KIM, Hae Yong; TANCREDI, F. B. Automation of the ACR MRI Low-Contrast Resolution Test Using Machine Learning. In: 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP- BMEI). IEEE, 2018. p. 1-6.

3. RAMOS, Jhonata E.; KIM, Hae Yong; TANCREDI, F. B. Automation of the ACR MRI Low-Contrast Resolution Test Using Machine Learning. Quality Improvement Reports Committee, CME Discussion. Radiological Society of North America (RSNA), 2019.

4. Jhonata E. Ramos, Hae Yong Kim, F. B. Tancredi, Automatização do Teste de Baixo Contraste do ACR Fazendo Uso de Aprendizado de Máquina, 49ª Jornada Paulista de Radiologia, São Paulo, 2019. Este trabalho recebeu um certificado de mérito.

5. Felipe BRUNETTO TANCREDI, Jhonata EMERICK RAMOS. Método para automação de teste de resolução em imagens digitais. Requerimento: BR 1020180712934, solicitação de Patente.

6. J. E. Ramos, H. Y. Kim and F. B. Tancredi, "Using Convolutional Neural Network to Automate ACR MRI Low-Contrast Detectability Test,"in IEEE Access, vol. 10, pp. 112529-112538, 2022, doi: 10.1109/ACCESS.2022.3216838.

2 TESTES E PROCEDIMENTOS RECOMENDADOS PELO COLÉGIO AMERICANO DE RADIOLOGIA (ACR)

Este capítulo trata sobre os aspectos técnicos do programa de acreditação do Colégio Americano de Radiologia, em especial na modalidade de ressonância magnética. Também são abordados os procedimentos de controle de qualidade recomendados, que fornecem um *design* próprio para o *phantom* e instruções detalhadas sobre os testes a serem realizados. A instituição recomenda uma rotina de testes e sugere valores de performance mínima para aceitação, bem como fluxo de trabalho e atribuição de responsabilidades aos envolvidos no programa de controle de qualidade.

2.1 Introdução

A ressonância magnética atualmente é um método de aquisição de imagem maduro e amplamente utilizado. No entanto, há uma variabilidade significativa na qualidade dos exames de MRI realizados em diferentes locais (RADIOLOGY et al., 2015). Alcançar todo o potencial desta modalidade de imagem requer atenção aos controles de qualidade, tanto no que diz respeito ao desempenho do equipamento quanto à execução dos estudos de imagem relacionados.

Em resposta às preocupações de médicos, pacientes e instituições, o ACR propõe um amplo programa de controle de qualidade. Nos Estados Unidos, todas as

instituições que prestam serviço ao governo devem aderir às recomendações do ACR e apresentar certificado(s) de acreditação para terem reembolsados seus custos. No resto do mundo o programa é adotado por instituições que reconhecem a importância de monitorar a qualidade das imagens que produzem.

O controle de qualidade de imagens médicas é essencial para o desenvolvimento e aplicação de métodos de análise ditos quantitativos, os quais permitem estabelecer critérios diagnósticos mais objetivos e recomendações médicas personalizadas. Métodos de imagem quantitativos são aqueles que permitem gerar imagens cujos *pixels* trazem informações objetivas sobre o tecido sob análise, tais como grau de rigidez, níveis de determinadas substâncias e/ou moléculas, taxa de fluxo sanguíneo, entre outros. Nessas imagens, os sinais dos *pixels* variam dentro de uma escala de valores com significado médico ou fisiológico. Portanto, além de oferecer informações sobre a relação entre sinais e sua distribuição, as imagens quantitativas oferecem um mapa de medidas, como se fosse uma balança (que nos serve porque mede peso em uma unidade física - Kg, por exemplo) de *pixels*. *Pixels* de imagens de diferentes pacientes ou do mesmo paciente examinado em tempos diferentes, podem ser comparados objetivamente, tal como acompanhamos a mudança de peso durante uma dieta, subindo numa balança. Portanto, para uma comparação de imagens válida, o(s) equipamento(s) deve(m) estar calibrado(s) e ser(em) submetido(s) a testes de controle periodicamente (RADIOLOGY et al., 2015), como acontece com qualquer equipamento de medida na indústria.

Variações no desempenho de *hardware* podem limitar a comparação de imagens obtidas em diferentes equipamentos de MRI, bem como entre imagens obtidas num mesmo equipamento mas em diferentes circunstâncias. Para uma comparação fidedigna, entre resultados clínicos e de pesquisa, é imprescindível que a qualidade das imagens, isto é, a calibração do *hardware*, onde as imagens são geradas, seja

rigorosamente controlada e padronizada. Programas de controle de qualidade têm, em geral, entre suas finalidades:

1. Assegurar que as imagens de ressonância magnética obtidas nos diversos tomógrafos sejam geradas em condições de paridade;

2. Estabelecer um fluxo de trabalho que facilite eventuais ações corretivas.

Um programa de controle de qualidade eficaz não eliminará todos os problemas, mas, em geral, permite a identificação dos mesmos antes que afetem os resultados clínicos.

O desempenho de máquina de MRI é comumente avaliado através de testes de imagem com um objeto de composição e geometria conhecidas, denominado *phantom*. Realizando-se uma série de medidas nas imagens desse objeto e comparando-as com valores esperados é possível determinar se o sistema opera dentro de suas especificações ou, caso contrário, identificar quais componentes necessitam de calibração.

Em relação ao tempo de realização dos testes, espera-se que o profissional treinado seja capaz de executá-los em até 45 minutos, de maneira a minimizar a interferência na prática clínica. Os testes devem ser realizados após decorrido um mínimo de 60 minutos, a contar da partida matinal do equipamento. Portanto, recomenda-se que sejam executados ao final da rotina da clínica diária, sempre que possível em um horário definido (agendado), já que sistematizações tendem a reduzir erros.

2.2 O *phantom* ACR

O *phantom* (Figura 3) multipropósito é um objeto utilizado para fazer a calibração das máquinas de ressonância magnética, recomendada pelo ACR. Os testes de

qualidade de imagem envolvem a coleta e a avaliação de uma série de imagens deste objeto. A avaliação das imagens deve ser realizada imediatamente após a aquisição, de forma a permitir a repetição do teste e da aquisição caso se faça necessário.

O *phantom* do ACR é um cilindro de acrílico com 190 mm de diâmetro e 148 mm de altura/comprimento, contendo em seu interior diversas estruturas de acrílico e plástico, cada uma delas servindo um propósito, tipo de medida de qualidade, diferente. O restante do interior do cilindro é preenchido por uma solução salina de Níquel, Cloro e Sódio. No seu exterior pode-se ler "*NOSE*" – do inglês, nariz – e "*CHIN*" – do inglês, queixo – que indicam como o objeto deve ser posicionado dentro do *scanner*. Não raro, durante a avaliação são detectadas falhas, que são consequências do mal posicionamento do *phantom* no interior do equipamento. Como prevenção a este tipo de retrabalho, pode-se utilizar um suporte para auxiliar no posicionamento do objeto. Além de diminuir o tempo de preparo e a frequência de erros, o uso do suporte também aumenta a reprodutibilidade dos testes. Vale ressaltar que, mesmo que um suporte seja adotado, é recomendado que o usuário/técnico não deixe as imagens para serem analisadas posteriormente.

O *phantom* possui estruturas internas com medidas e composição conhecidas. Para a calibração é injetada uma solução de cloreto de níquel, 10 mmolar, e cloreto de sódio, 75 mmolar, a qual preenche o espaço vazio dentro do mesmo. Pequenas ampolas no seu interior carregam soluções com os mesmos componentes, mas em diferentes concentrações, 20 e 15 mmolar respectivamente, servindo como referência para diferenças entre tempos de relaxação T1 e T2 (RADIOLOGY et al., 2015). Uma fração mínima é preenchida pelo ar, como forma de impedir que dilatações térmicas do fluido gerem danos ao invólucro, por tal motivo é normal observarem-se bolhas dentro do *phantom*.

Figura 3: Exemplo de um *phantom* multipropósito com *design* recomendado pelo ACR.

Fonte: Adaptado de *Magnetic resonance imaging quality control manual* (RADIOLOGY et al., 2015).

Medidas do limite de resolução são baseadas nas imagens de três pequenas matrizes de estruturas, com dimensões variadas, inseridas em uma barra espessa de 11 mm de comprimento. As estruturas são quadradas, com lados que medem: 0,9, 1,0 e 1,1 mm, e o espaçamento dessas estruturas são iguais às respectivas dimensões dos seus lados. Entre as extremidades laterais do *phantom* encontram-se duas barras com rampas de inclinação conhecida (razão de 1:10) e correndo em direção contrária no seu interior. Imagens dessa estrutura permitem estimar a exatidão da espessura de tomo prescrita. Uma grade larga, 10cm $\times$ 10cm com quadrados de 1,4 cm, localizada no centro do *phantom*, é utilizada para medidas de distorção. Na extremidade distal do *phantom*, encontram-se 4 discos de baixa densidade de contraste, que consistem em finas folhas de policarbonato de espessuras variadas e com furos redondos, também de variadas dimensões. Contribuições oriundas de volume parcial das folhas e solução produzem pequenas variações de sinal, e podem ser usadas para avaliar a capacidade do equipamento de distinguir objetos com baixo contraste. Nas extremidades distal e proximal do *phantom*, na sua porção ventral, existe um par de cunhas com 45^{o} e -45^{o} de inclinação. Cada uma tem 2 cm e suas rampas se cruzam a 1 cm da origem. A distância entre os pontos de interseção de cada par de rampas é de 10 cm. Essas rampas são usadas para avaliar desvios e espaçamento de tomo. É com base na imagem sagital dessas rampas que a prescrição dos tomos axiais é feita.

O erro mais comum de falha no processo de aquisição das imagens é o mal posicionamento do *phantom*, apenas 2 mm de deslocamento podem comprometer todo o teste. Desta forma, um adesivo com desenho de seta indica a orientação que o *phantom* deve ser posicionado. O uso de espumas pode ajudar a imobilização do *phantom* e alinhamento. O *landmark* deve ser feito no centro da linha de referência que o circunda, sendo que o alinhamento sagital é realizado com base na linha que

corre na face ventral. Uma bolha de nível presente no interior do *phantom*, na sua face ventral, permite o fino alinhamento do mesmo ao longo do eixo axial.

Utilizando-se o protocolo recomendado pelo ACR, que prevê imagens ponderadas na constante de relaxação T1, o sinal da solução salina é alto e o sinal do acrílico/plástico é baixo. Desse modo, *pixels* que representem a solução salina tem sinal máximo - branco - na imagem, e *pixels* que representem acrílico/plástico têm sinal mínimo - preto. *Pixels* de regiões que abarquem ambos materiais tem sinal intermediário. Por exemplo, é normal que *pixels* recaiam sobre a borda das estruturas de acrílico, sendo o mesmo tanto mais escuro quanto mais deslocado ao interior do furo estiver. Isso não significa, no entanto, que *pixels* totalmente no interior do furo sejam perfeitamente pretos. O *phantom* foi projetado para gerar imagens com dois tipos de sinal muito bem distintos - isto é, da solução vs. do acrílico/plástico, com exceção das regiões destinadas aos testes de baixo contraste, projetadas para gerar sinais intermediários. No entanto, suas imagens não são preto e branco, e sim em escala de cinza, com altíssimo contraste.

O *phantom* ACR é definido como multipropósito, pois existem pelo menos sete testes (Figura 4) quantitativos que podem ser executados a partir de suas imagens:

- Distorção na geometria;
- Relação sinal ruído e *ghosting*;
- Uniformidade de emissão/detecção de sinal;
- Espessura, posicionamento e espaçamento de/entre tomos;
- Resolução espacial em alto contraste;
- Detecção a baixo contraste.

Os testes de qualidade recomendados pelo ACR, que devem ser repetidos semanalmente, são os de distorção geométrica, resolução espacial, presença de artefatos, frequência central e precisão do deslocamento ao isocentro do magneto. Os testes de frequência central se resumem a registrar o valor da frequência de ressonância no momento do teste. O teste do isocentro se resume a medir se a região do *phantom* em que foi feita a marcação com os *lasers* de referência aparece na imagem com a coordenada Z próxima de 0. Os testes de distorção geométrica requerem que o operador faça a medição de distâncias – mais precisamente a altura e diâmetro interno do cilindro do *phantom* – com o uso de réguas virtuais manuseadas na tela através de um mouse.

O operador também avalia a série de imagens quanto a presença de artefatos, de qualquer tipo, que ele considere significativo de menção. Já nos testes de resolução, o operador deve indicar se é ou não capaz de diferenciar furos de diferentes dimensões e contrastes, contra o fundo da imagem. Um ponto importante a ressaltar é que os testes ditos de resolução não exigem qualquer tipo de medida – refletem a impressão do observador sobre se é ou não capaz de perceber um determinado furo contra o fundo da imagem. Testes como o de distorção geométrica, por exemplo, dependem do operador determinar o ponto de transição entre o sinal do interior do *phantom* e seu envoltório acrílico, sinais com extremo de contraste, e esticar uma régua virtual que toque dois pontos como esse, diametralmente opostos, em dadas direções. Nos testes de resolução, o observador verifica qual a menor dimensão que um furo deve ter para poder ser detectado nas dadas condições de imagem. Por condições, entendem-se níveis de sinal, contraste, ruído e imperfeições. O tempo total dispensado com o teste gira em torno de 15 a 25 minutos. Em tomógrafos ocupados com pacientes de emergência, por exemplo, os encaixes na agenda são a regra. Mas observa-se que as taxas de erro são menores quando os testes são executados em horários pré-

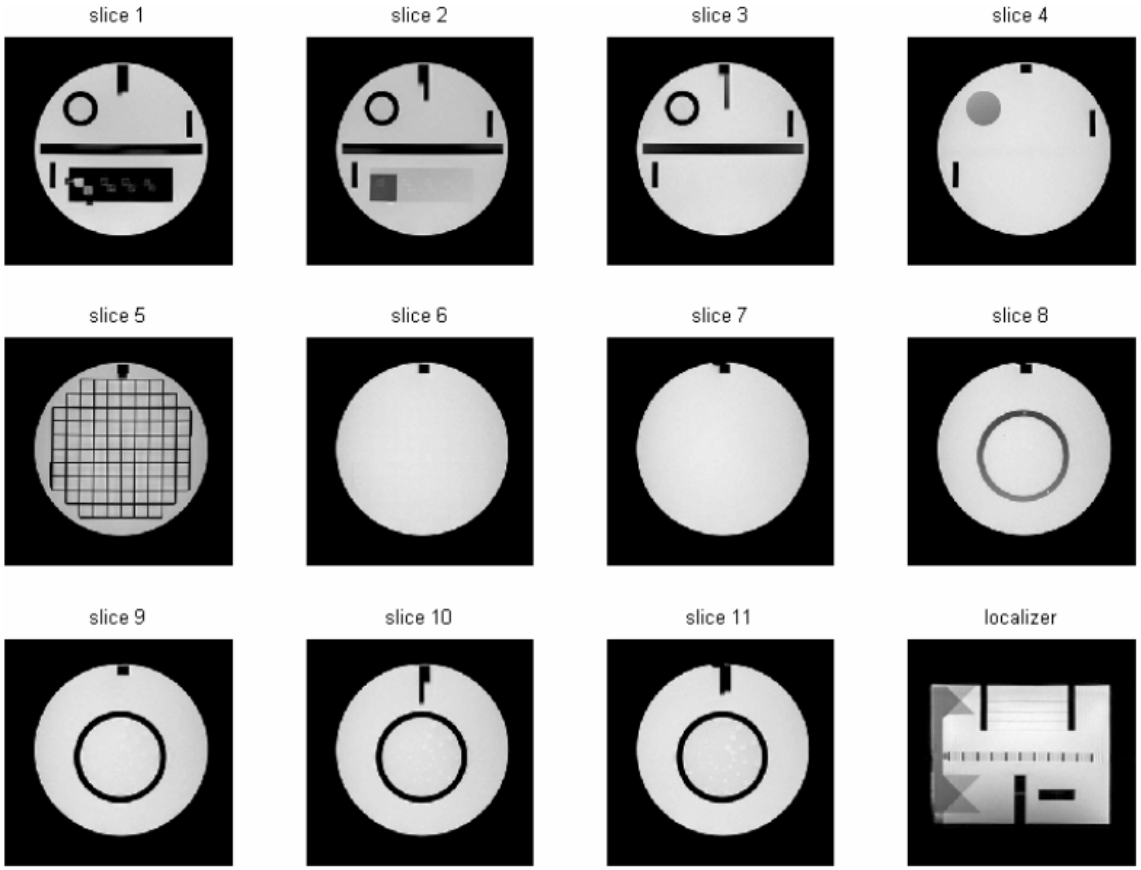

Figura 4: Exemplo de uma série de imagens utilizada para realizar os testes de acreditação do ACR.
Fonte: Adaptado de *Magnetic resonance imaging quality control manual* (RADIOLOGY et al., 2015).

agendados.

Os testes baseados em medidas diretas, como as medidas de comprimento, usando réguas virtuais dos testes de distorção geométrica, costumam ser consensuais, isto é, objetivos, de modo que sua automação pode ser realizada utilizando estratégias de tratamento de imagens convencionais. No entanto, os testes de resolução envolvem uma avaliação subjetiva, afinal, são um reflexo direto da percepção visual humana, o que torna sua automação um grande desafio. Os testes de controle de qualidade esbarram na dificuldade da reprodução dos resultados humanos dos mesmos testes mencionados, sobretudo nos testes de resolução.

2.3 Distorção da geometria

O teste de geometria consiste em medir distâncias entre extremidades do *phantom*, nas fatias axiais 1 e 5, obtidas com a sequência T1 ACR, bem como no localizador. Como a escala de cinza pode influenciar a localização aparente dos furos, recomenda-se que as medidas de distância sejam tomadas usando janela de brilho estreita. Isso é realizado baseado na imagem do localizador, da seguinte maneira:

- Ajustar a escala de cinza para 1 - normalização;
- Observar a região do *phantom* que contém solução, cujo sinal deve ser máximo;
- Ajustar o fundo de escala até que essa região fique toda branca;
- Aumentar o mínimo do brilho de monitor até que metade da região de líquidos torne-se escura (como ilustrado na Figura 6);
- Tomar nota desse nível, que representa a mediana do sinal da solução, e será usado como estimativa da média;
- Reduzir o nível de brilho até metade do valor anotado;
- Aumentar a intensidade de brilho até o valor anotado.

Após ajustada a janela, medir a distância crânio-caudal do *phantom* na imagem do localizador. Isso deve ser feito próximo ao eixo do *phantom*, à esquerda ou à direita da barra escura central. No corte axial 1, medir o diâmetro na direção anteroposterior e na direção laterolateral. Repetir o procedimento no corte axial 5, mas nesse tomo incluir medidas de diâmetro nas diagonais, o que é auxiliado pela estrutura interna em forma de grade. As linhas medindo o diâmetro, devem cruzar-se no centro do *phantom* e todas tocarem o limite da sua circunferência. Para o

comprimento interno do *phantom* na direção crânio caudal espera-se obter o valor de 148 mm. Em relação ao diâmetro interno o valor esperado é de 190 mm.

Para o teste ser aceito a diferença entre os valores medidos e esperados não devem diferir em mais de 2mm.

2.4 Relação sinal ruído e *ghosting*

O teste da relação sinal ruído permite estabelecer o nível de *ghosting* em percentagem do sinal central da imagem. Esse teste é realizado no tomo 7 do *phantom*, adquirido com a sequência T1 do ACR. Consiste basicamente em calcular a intensidade de sinal em 5 regiões diferentes do FoV (*Field of View*) da imagem, e calcular frações de sinal correspondentes.

No corte axial 7, desenhar um grande ROI no centro da imagem do *phantom*, como ilustrado na Figura 5, e como realizado no teste anterior. A área desse ROI deve ter entre $195cm^2$ e $205cm^2$ (ou $19.500mm^2$ e $20.500mm^2$) e deve ser posicionado bem ao centro da imagem, sem incluir o pequeno quadrado preto que aparece na porção superior, além de anotar o sinal médio no mesmo. Em seguida, deve-se desenhar outros 4 ROI's elípticos em torno do *phantom*, na área correspondente ao fundo: um na porção superior, um na inferior, um na lateral esquerda e outro na direita.

Para computar o nível de *ghosting* como uma fração do sinal primário, deve-se utilizar a equação (2.1):

$$ghosting = \frac{100\ ((sinal_{superior} + sinal_{inferior}) - (sinal_{esquerda} + sinal_{direita}))}{2\ (sinal_{ROIcentral})} \quad (2.1)$$

Em relação ao critério de aceitação do ACR, o teste é considerado positivo quando o nível de *ghosting* é menor que 2.5%.

2.5 Uniformidade de emissão/detecção de sinal

Nesse teste visa-se averiguar a homogeneidade de sinal no corpo do *phantom*, na região central da bobina de cabeça. Utiliza-se o tomo 7, que não contém furos no centro da imagem. As imagens são adquiridas com as sequências T1 e T2 do ACR.

Desenhar um ROI circular, de área $195 - 205cm^2$, no centro da imagem axial 7. O ROI não deve circunscrever o furo preto que aparece no topo da imagem. Diferentemente do teste anterior, não é necessário anotar o valor de intensidade média no ROI. Ajustar o nível de brilho para um mínimo, até que a área inteira do *phantom* fique branca. Aumentar o nível de brilho até que uma pequena porção dos *pixels* no interior do ROI torne-se preta (aproximadamente $1cm^2$ ou $100mm^2$). Esses *pixels* representam a região de menor intensidade de sinal do *phantom*. Desenhar um pequeno ROI circular de $1cm^2$ nesta região, como representado na Figura 6-centro. Em seguida registrar a intensidade de sinal média nesse ROI e continue aumentando o nível de brilho até que reste somente uma pequena região com sinal, no centro da imagem (Figura 6-direita). Essa é a região de máximo sinal. Desenhar um segundo ROI circular de $1cm^2$ nesta região. Anotar a intensidade de sinal média do ROI. A homogeneidade de intensidade de sinal ou Percentagem Integral de Uniformidade (PIU) é calculada através da equação (2.2) usando-se os valores anotados de alta e baixa intensidade de sinal:

$$PIU = \frac{100\ (1 - (intensidade_{alta} - intensidade_{baixa}))}{intensidade_{alta} + intensidade_{baixa}} \tag{2.2}$$

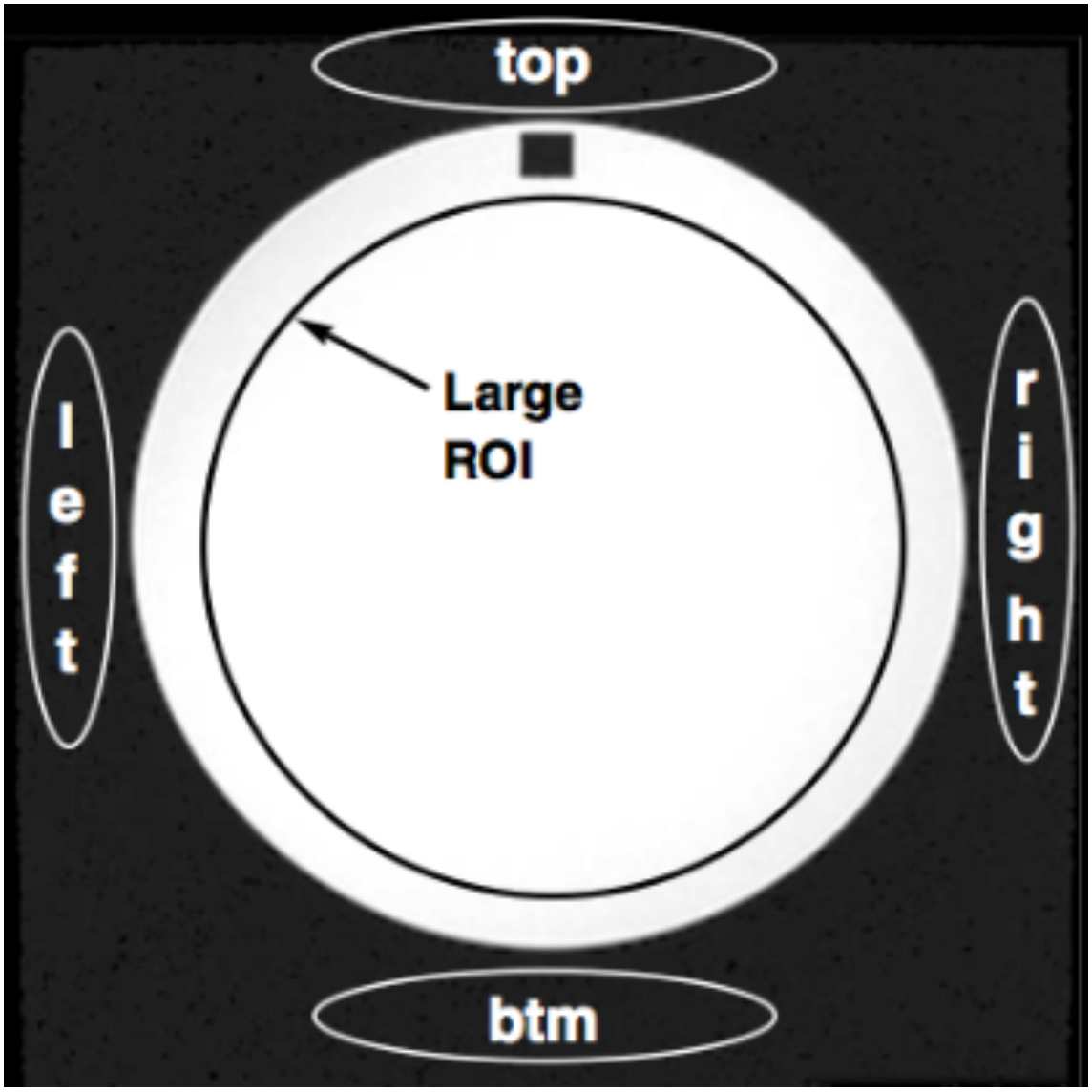

Figura 5: ROIs desenhados sobre o corte axial 7.
Fonte: Adaptado de *Magnetic resonance imaging quality control manual* (RADIOLOGY et al., 2015).

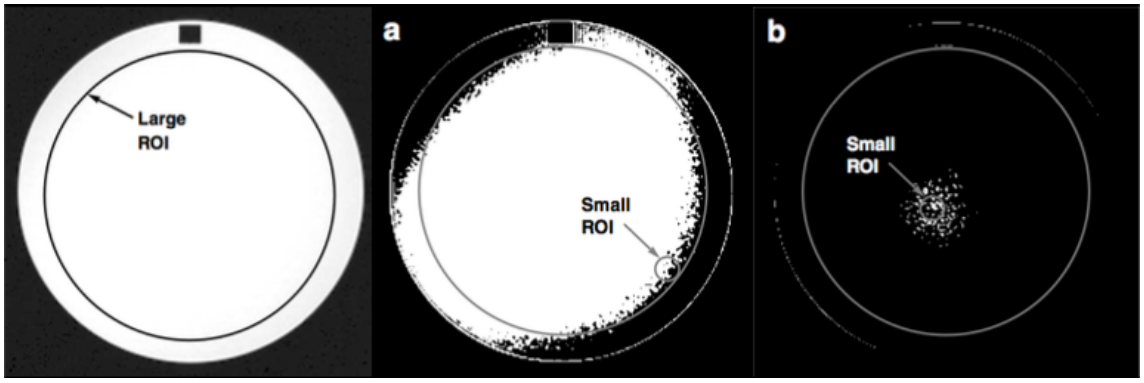

Figura 6: Exemplo de ROIs desenhados no corte axial 7 com intensidade de brilho ajustadas.
Fonte: Adaptado de *Magnetic resonance imaging quality control manual* (RADIOLOGY et al., 2015).

Em relação ao critério de aceitação do ACR: Para sistemas de 1.5T o PIU deve ser maior ou igual a 87.5%. Para sistemas de 3T o PIU deve ser maior ou igual a 82.0%.

Uma possível causa de falha nesse teste pode ser o mal posicionamento do *phantom* dentro da bobina de cabeça, cuja sensibilidade varia espacialmente. Outra possível razão de falha são artefatos de movimento.

2.6 Espessura, posicionamento e espaçamento entre tomos

Os presentes testes visam determinar se o equipamento realiza a aquisição de acordo com a prescrição geométrica, ou seja, se as imagens geradas representam os tomos que foram selecionados a partir do localizador, e se a espessura do tomo corresponde a 5 mm. O teste de espessura é realizado no tomo 1, enquanto que o de posicionamento é realizado nos tomos 1 e 11. Ambos os testes requerem avaliação com imagens adquiridas com as sequências T1 e T2 do ACR.

2.6.1 Teste de espessura

Medem-se os comprimentos de duas rampas no tomo 1. Amplificar a imagem em 2 a 4 vezes e ajustar o brilho da imagem para visualizar bem as rampas. Em seguida, desenhar 2 ROIs retangulares, um em cada rampa, na região central, e anotar o sinal em cada um deles, para se calcular a média entre eles. Reduzir o nível de cinza (*Level*) à metade do valor anterior, e deixar a janela (*Window*) no mínimo (zero), de maneira a obter contraste máximo entre a região central da rampa e suas extremidades. Medir o comprimento da região branca central como mostrado na Figura 7 (direita).

A medida de comprimento errônea pode levar ao erro na estimativa de espessura da fatia. Entretanto, vale salientar que, dado que a proporção entre comprimento de rampa e espessura de fatia é 10:1, o erro na estimativa de espessura é um décimo do erro da medida de comprimento.

Estima-se a espessura de fatia usando as medidas de comprimento das rampas inferiores ($comp_{inf}$) e superior ($comp_{sup}$) na equação (2.3):

$$espessura = 0.2\ \frac{comp_{sup}\ comp_{inf}}{comp_{sup} + comp_{inf}} \tag{2.3}$$

Em relação ao critério de aceitação do ACR, a espessura de fatia computada não deve diferir da espessura prescrita (5 mm) em mais de 0.7 mm.

2.6.2 Teste de posicionamento

Em ambos tomos 1 e 11, as cunhas cruzadas da parte central do *phantom* devem aparecer como barras pretas de igual comprimento na parte superior da imagem axial do *phantom*. Aumentar a imagem em 2 a 4 vezes, de maneira a bem visualizar os

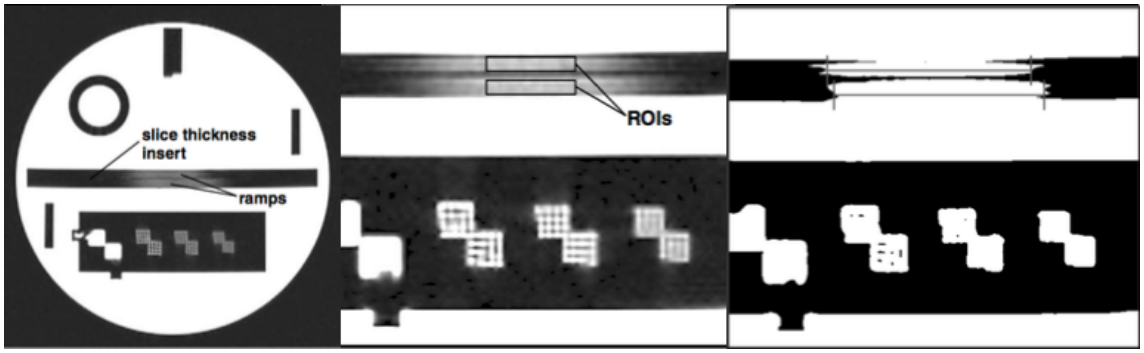

Figura 7: Imagens das barras laterais de espessura tomo no corte axial 1.
Fonte: Adaptado de *Magnetic resonance imaging quality control manual* (RADIOLOGY et al., 2015).

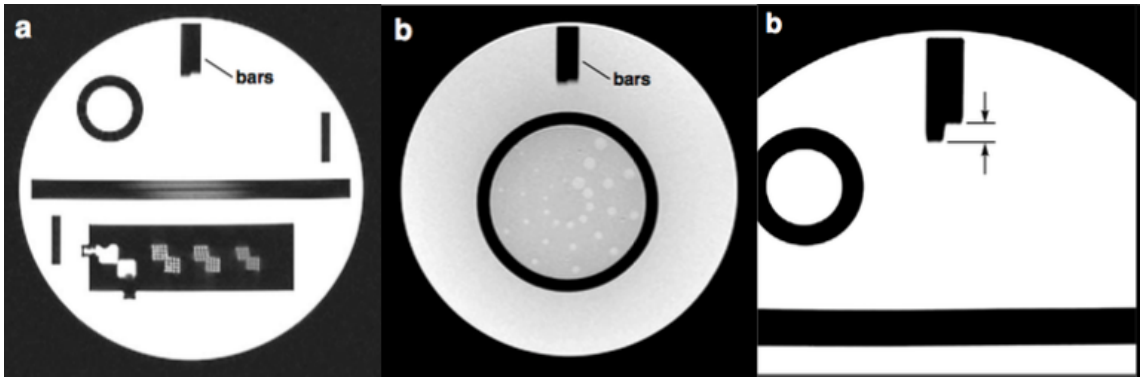

Figura 8: Imagens das barras longitudinais de posicionamento de tomo nos cortes axiais 1 e 11.
Fonte: Adaptado de *Magnetic resonance imaging quality control manual* (RADIOLOGY et al., 2015).

pequenos furos e ajustar os níveis de cinza e brilho para que os contornos das barras fiquem bem definidos. Em cada um dos tomos medir a diferença de comprimento entre as duas barras. Se a barra da esquerda for mais longa, usar um sinal de menos para denotar a diferença. Como as barras têm inclinação de 45°, a diferença de comprimento representa duas vezes a diferença de posicionamento prescrito, ou seja, uma medida de 2 mm de diferença representa um deslocamento de 1 mm.

Em relação ao critério de aceitação do ACR, a diferença de comprimento entre as barras deve ser menor ou igual a 4 mm.

2.7 Resolução espacial de alto contraste

Uma imagem de alta resolução permite a um observador detectar furos de pequena dimensão na imagem. O tomógrafo de ressonância deve gerar imagens com poder de resolução mínimos, que satisfaçam a determinados critérios de aceitação. Os critérios de aceitação podem diferir para sistemas de diferentes configurações. Espera-se, naturalmente, que tomógrafos de 3T gerem imagens com maior poder de resolução quando comparados a equipamentos de 1.5T, mesmo que com menor nível de homogeneidade de sinal. Portanto, tomógrafos de campos magnéticos diferentes são tratados como instrumentos diferentes e são testados observando-se diferentes critérios de aceitação, de acordo com suas potencialidades e limitações.

Este teste é realizado analisando-se visualmente as imagens do corte axial 1 (obtidas com as sequências T1 e T2 do ACR), que possui 3 pares de grades com furos de diâmetro e distanciamento variados. O par da esquerda possui furos de 1,1 mm, o do meio, 1,0 mm e o da direita, 0,9 mm. Colocar as grades de resolução no centro do monitor e aumentar o tamanho da imagem entre 2 e 4 vezes. Começar a análise partindo da esquerda, onde os furos têm maior dimensão. Ajustar a janela de brilho e escala de cinza de maneira a obter alto contraste dos furos das grades superiores. Quando os furos de uma linha qualquer da grade são distinguíveis, a grade é avaliada como observável da-esquerda-para-a-direita para aquela dimensão. Notar que, por distintos entende-se que as linhas podem ser identificadas como furos de maior intensidade que o fundo preto. Não é necessário que o furo seja identificado como quadrado perfeito, dado que isso é praticamente impossível em decorrência do espalhamento natural do sinal de MRI. Repetir esse procedimento analisando as grades inferiores. Quando os furos de uma coluna qualquer da grade são distinguíveis, a grade é avaliada como observável de-cima-para-baixo para respectiva dimensão.

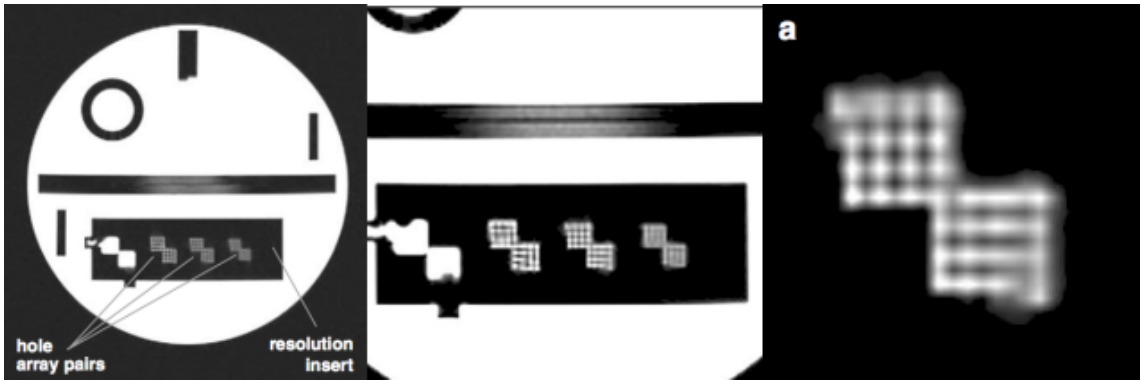

Figura 9: Imagens das grades de resolução no corte axial 1.
Fonte: Adaptado de *Magnetic resonance imaging quality control manual* (RADIOLOGY et al., 2015).

Repetir o procedimento para as grades com furos de menor dimensão.

As sequências T1 e T2 do ACR foram escolhidas para gerar uma resolução de $1mm^2$. Desta forma, as resoluções mínimas devem ser 1mm em ambas as direções para que o teste seja aceito.

2.8 O teste de baixo contraste do ACR

Em se tratando de detecção, é importante assegurar que, de fato, existe um furo, numa dada região da imagem que é diferente do fundo ou dos seus entornos. Um dos maiores determinantes da resolubilidade de uma imagem é o contraste, isto é, a diferença relativa entre o sinal capturado do furo no campo de visão e o sinal de fundo da imagem. Quanto menor o contraste, menor o poder de diferenciação de pequenos furos na imagem. Pois, partindo de um determinado nível de contraste, oferecido por um dado protocolo, os demais determinantes da resolubilidade de uma imagem são parâmetros de qualidade, tais como nível de sinal-ruído, que reflete sensibilidade do instrumento, e espalhamentos de sinal, que refletem outras limitações e imperfeições.

Um teste de resolução em baixo contraste é um teste que visa determinar o poder

de resolução das imagens geradas por um equipamento quando os furos apresentam baixo contraste, ou seja, quando os furos que se procuram diferenciar apresentam pouco destaque de sinal em relação aos seus entornos ou fundo em que se inserem. Assim, este teste tem como objetivo averiguar a capacidade do equipamento de oferecer imagens com qualidade suficiente para permitir diferenciar os furos de pouco contraste.

Com o objetivo de detectar furos com pequenas diferenças de sinal, seguindo o protocolo do próprio equipamento onde a aquisição é feita, o teste utiliza as últimas 4 fatias, região posterior do *phantom*, às quais correspondem as fatias 8-11 no protocolo de imagem do ACR. Nessa região existem finos filmes plásticos circulares, cada um perfurado com 30 furos de diâmetros variados, os quais diminuem de 7 mm para 1,5 mm, no sentido horário (Figura 10). Todos os furos de uma determinada fatia têm o mesmo nível de contraste, variando de 1,4%, 2,5%, 3,6% e 5,1%, de acordo com a fatia (8 a 11). Os furos também giram no sentido anti-horário, cerca de 9 graus de uma fatia para outra.

Um *spoke* é considerado visível quando todos os seus 3 furos podem ser claramente detectados. O termo *spoke* denota o conjunto de três furos de mesmo diâmetro, alinhados radialmente (Figura 12). O teste de baixo contraste ACR consiste em contar quantos dos 10 *spokes* podem ser detectados em uma determinada fatia, geralmente na fatia 8 ou 9, dependendo de qual o físico responsável considerar mais relevante para o sistema em análise. Por exemplo, uma possível recomendação é que a contagem de *spokes* de um sistema de 1,5T seja igual ou superior a 28.

Enquanto esses furos são preenchidos com a solução iônica do *phantom* e fornecem a intensidade máxima do sinal, o sinal do fundo depende da espessura do disco, feito de um material que não emite sinal. Quanto mais fino o disco plástico, maior a

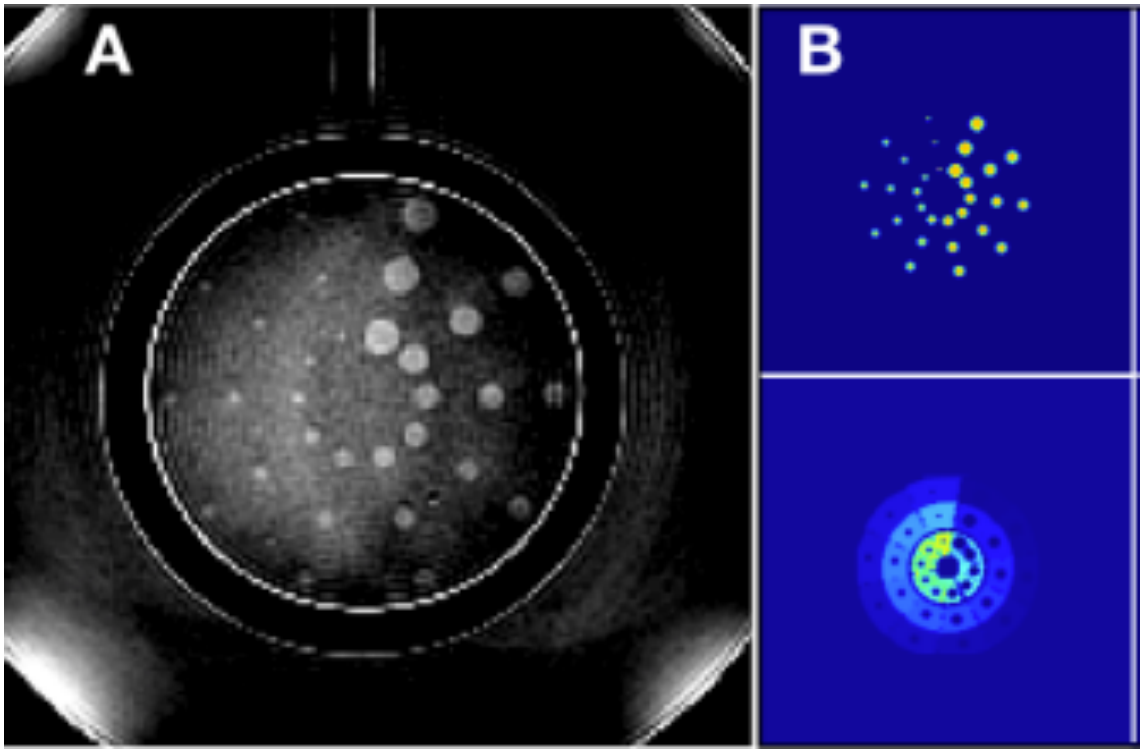

Figura 10: (A) Uma típica imagem do *phantom* utilizado para teste de ACR na fatia 10. (B) Duas máscaras utilizadas para se obter as características da imagem.

contribuição da solução iônica para o sinal, e o contraste entre os furos e seu fundo diminui (Figura 11).

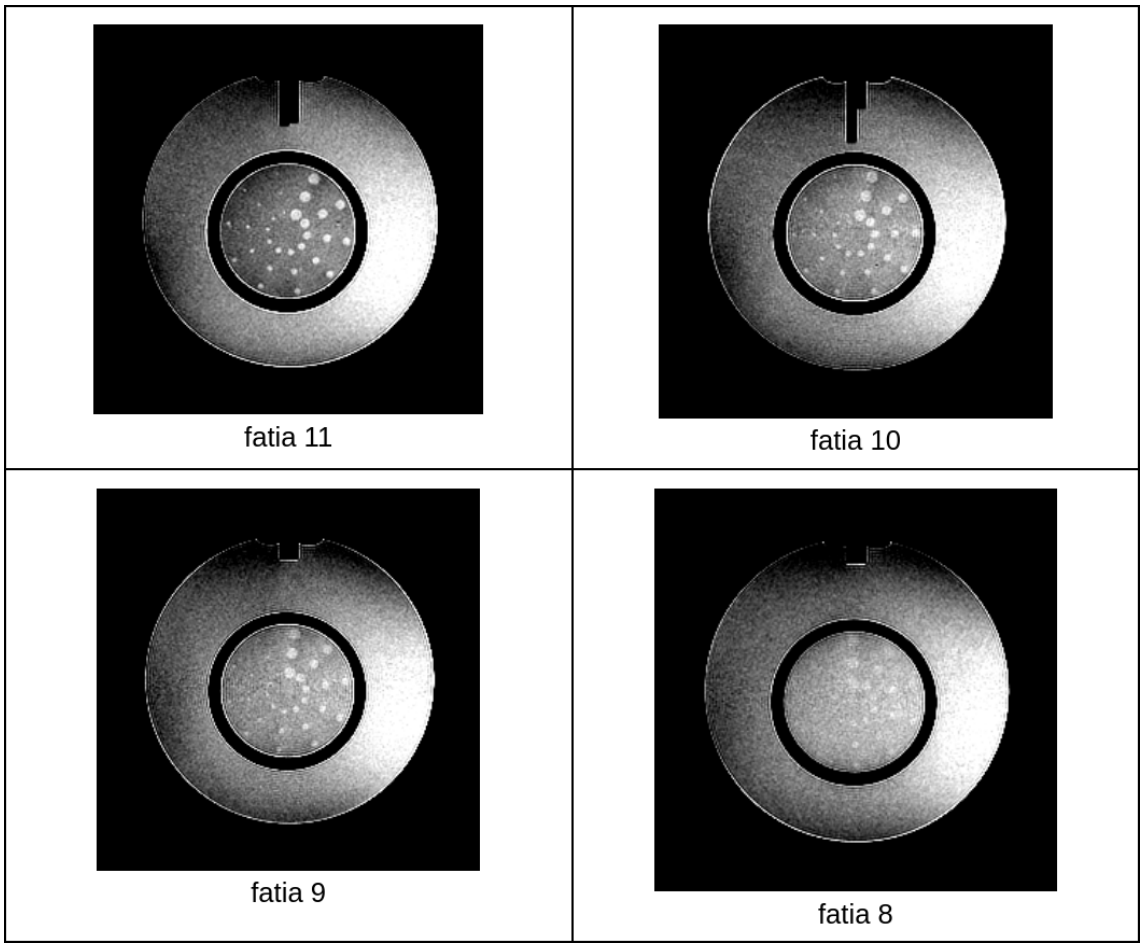

Figura 11: Imagens em T1 das fatias de 11 a 8 do *phantom* ACR. As imagens de ressonância magnética foram adquiridas como imagens de números inteiros de 16 bits, com 12 bits significativos.

Fonte: Adaptado de *Magnetic resonance imaging quality control manual* (RADIOLOGY et al., 2015).

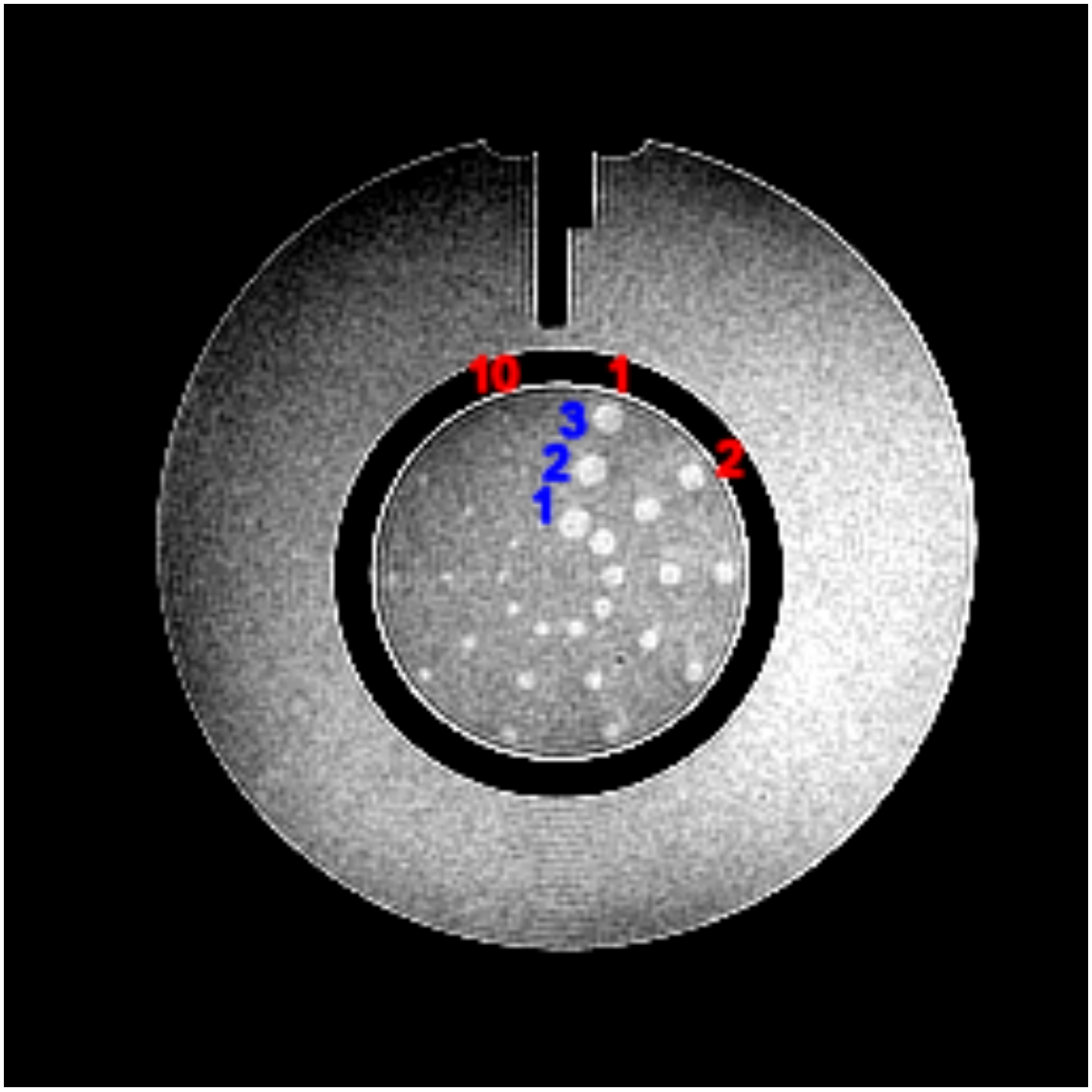

Figura 12: Uma típica imagem de ressonância magnética de uma fatia do *phantom* ACR. Os números vermelhos de 1 a 10 são os índices dos *spokes* (ângulo). Os números azuis de 1 a 3 são os índices de furos dentro de cada *spoke* (posição radial).

3 MÉTODOS CLÁSSICOS

Este capítulo descreve de forma detalhada a primeira fase desta tese. A extração das características foi realizada manualmente com base em três regiões de interesse. Essas características foram então utilizadas como variáveis explicativas, no processo de treinamento dos algoritmos clássicos de aprendizado de máquina, para emular a percepção de um observador humano, técnico, no teste de resolução de baixo contraste do ACR.

3.1 Experimentos

3.1.1 Extração de características

Para a modelagem da visibilidade dos furos do *phantom* ACR foram extraídas características específicas da imagem, a serem vistas mais adiante. Essa extração foi baseada na sobreposição da imagem original com 3 tipos de ROIs:

1. O primeiro sendo a região do furo propriamente dita;
2. O segundo representando suas adjacências (setor de círculo menos furo);
3. O terceiro formado por segmentos de reta no sentido radial (linhas), que cruzam a região de fundo entre os *spokes*.

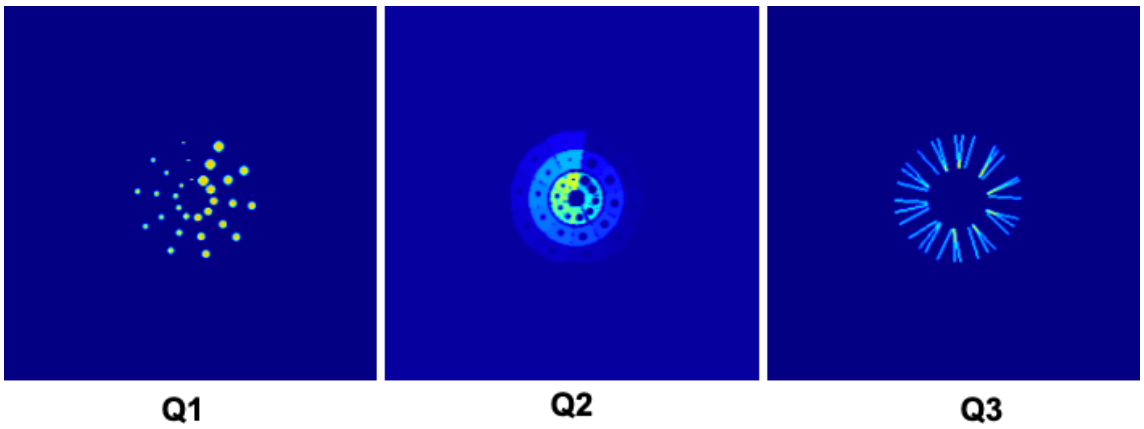

Figura 13: Máscaras utilizadas na extração das características da imagem.

Esses três tipos de máscara são mostrados na Figura 13, na qual o quadro Q1 da figura mostra a máscara furo, o quadro Q2 mostra a máscara setor e o quadro Q3 mostra a máscara linhas.

O mascaramento da imagem com os ROIs do tipo furo e setor serviu para se obter atributos dos furos. Dos ROIs se extraíram média e desvio padrão (ruído) do sinal - as duas métricas mais importantes ao descrever a qualidade de uma imagem.

Foi desenvolvida uma estratégia para posicionar esses ROIs na sua exata posição, baseada em corregistro de *templates* 2D, em cada uma das fatias 8 a 11, respectivamente.

Os furos de baixo contraste das fatias 8 a 11 se encontram dentro de um círculo cinza de bordas pretas. A primeira ação é localizar, em cada corte, o centro e o raio (R) desse círculo, que vão determinar as posições dos ROIs.

O círculo interno é dividido em 10 frações de 36 graus de abertura cada. Quando o procedimento é aplicado na fatia 8, as frações são dispostas de maneira que uma delas, digamos a fração número 1, esteja alinhada com o eixo vertical, na posição 90°. Quando o procedimento é aplicado na fatia 9, as frações apresentam uma rotação de 9° no sentido horário, na fatia 10, a rotação é de 18° e na fatia 11, de 27°.

Duas circunferências de raios correspondendo a 40% e 70% de R dividem o círculo em três frações da seguinte maneira:

1. Um círculo menor de 40% de R;
2. Uma coroa de raio interno 40% de R e raio externo 70% de R;
3. Uma coroa de raio interno 70% de R e raio externo 100% de R.

Os 30 ROIs denominados setores são formados pela interseção dos círculos e coroas com as frações.

No caso da interseção do círculo menor com as frações, temos como resultado 10 setores de círculo. No caso da interseção das coroas com as frações, temos como resultado 20 setores de coroa.

Em relação às linhas, são formados 30 segmentos de reta, 3 entre cada *spoke* de furos, na direção radial e de comprimento correspondendo a 40% de R, terminando na borda do círculo. Os segmentos são distantes 14°, 18° e 22° do centro dos setores.

A formação dos ROIs do tipo círculo se deu de maneira ligeiramente diferente. Os raios dos furos de baixo contraste são conhecidos. Os raios dos furos em cada um dos 10 *spokes* são: 7.0mm, 6.0mm, 5.0mm, 4.5mm, 4.0mm, 3.5mm, 3.0mm, 2.5mm, 2.0mm e 1.5mm. Para determinar as posições, optou-se por utilizar uma imagem adquirida em um equipamento de MRI de alto campo, antena de 32 canais de recepção, 10 repetições e um tempo de eco de 4ms. De maneira a minimizar distorções, medimos a posição relativa do centro de cada um dos furos de baixo contraste com relação ao centro do círculo de raio R. Tomou-se o cuidado de posicionar o *phantom* no centro do magneto sem qualquer rotação nos eixos x, y e z, e as imagens axiais também foram adquiridas sem qualquer inclinação. Foram colecionadas 120

coordenadas, expressas em termos do raio R e de um círculo de centro em $x = 0$ e $y = 0$. O posicionamento dos ROIs é realizado, então, ajustando-se o *template* de posições para cada imagem de acordo com o centro e raio (R) do furo detectado. A partir das coordenadas co-registradas são criados os ROIs circulares com diâmetro igual ao diâmetro conhecido do furo, acrescido de 3mm.

Além das características de imagem extraídas das fatias 8-11, outras métricas de qualidade extraídas de outros cortes do *phantom* ACR podem ajudar no desempenho do algoritmo. Assim, variáveis preditivas adicionais poderiam ser utilizadas no desenvolvimento da solução. Dentre elas o nível de *ghosting*, de homogeneidade de sinal e de distorção geométrica. Da mesma forma, condições de aquisição, tais como tipo de *hardware*, campo magnético do *scanner* de MRI e antena, e banda de recepção, que não são determinadas pelo ACR.

Portanto, depois de todo este procedimento, para cada ROI são extraídas média e desvio padrão do sinal. Além disso, cada furo recebeu um índice composto por 3 números: fatia (valores de 8 a 11), ângulo (1 a 10) e posição radial (1 a 3).

As coordenadas dos furos variam de uma aquisição para outra. Calculamos essas coordenadas da seguinte maneira:

1. Co-registramos a imagem da fatia 11, que possui o maior contraste, com o *template* de furos para obter os parâmetros de uma transformação afim, e calculamos as coordenadas dos furos nesta fatia;

2. Rotacionamos as coordenadas dos furos da fatia 11 no sentido anti-horário, em passos de aproximadamente 9,0° para obter as coordenadas dos furos nas fatias 8-10, respectivamente.

A partir dos ROIs descritos, extraímos quatro características principais:

- S_{in}: Sinal médio (valor médio) dentro do furo;
- N_{in}: Ruído (desvio padrão) dentro do furo;
- S_{out}: Sinal médio nas adjacências do furo;
- N_{out}: Ruído nas adjacências do furo.

Os três números que compõem o índice do furo, fatia, ângulo e posição radial, também foram utilizados como variáveis explicativas. Intuitivamente, os mesmos se relacionam diretamente com o problema, dado que:

1. Fatia - o contraste da imagem depende deste número e ajuda a classificar corretamente a visibilidade dos furos;
2. Ângulo - o raio do furo depende do seu ângulo na fatia, e quanto maior o furo, mais fácil de ser detectado;
3. Posição radial - geralmente, os furos externos são mais distorcidos e difíceis de se visualizar quando comparados aos furos internos.

Além das características de imagem extraídas das fatias 8-11, outras possíveis variáveis explicativas foram testadas com objetivo de aumentar a capacidade de predição da solução. Por exemplo, variáveis relacionadas às condições de aquisição, tais como: tipo do *hardware*, campo magnético do *scanner* de MRI e antena, e banda de recepção, que não é determinado pelo ACR. Ao adicionar estas variáveis notamos que a melhora no modelo era marginal, assim, optamos por descartá-las. A decisão tem como objetivo obter um modelo simples, eficiente e interpretável. Dessa forma, as variáveis finalistas estão descritas na Tabela 1.

Tabela 1: Descrição das variáveis finalistas do modelo.

Variável	**Tipo**	**Descrição**
S_IN	Numérica	Sinal dentro do furo
N_IN	Numérica	Ruído dentro do furo
S_OUT	Numérica	Sinal na área ao redor do furo
N_OUT	Numérica	Ruído na área ao redor do furo
Ângulo	Categórica	O ângulo do furo que indica seu tamanho
Fatia	Categórica	Fatia no qual está localizado o furo
Posição	Categórica	Posição do furo no triplete

3.1.2 Métodos de aprendizado de máquina

Testamos 5 métodos de aprendizado de máquina para prever as respostas dos técnicos. A base de dados foi dividida, treinada e analisada utilizando o método de validação cruzada (REFAEILZADEH; TANG; LIU, 2009) de *10-fold*, o que nos mostra uma média mais realista da performance de cada método, quando aplicado no banco de dados como um todo. Os algoritmos foram desenvolvidos e testados usando a linguagem de programação R (R-PROJECT,).

Regressão Logística (LR) é o método mais utilizado em problemas de classificação binária (JR; LEMESHOW; STURDIVANT, 2013), como o abordado neste projeto de pesquisa. É uma técnica estatística que, a partir de um conjunto de observações, gera um modelo que prevê uma variável de saída binária a partir de uma série de variáveis explicativas contínuas e/ou discretas. Este algoritmo foi implementado usando o método de modelo linear generalizado padrão em R.

Support Vector Machine (SVM) é um método de classificação muito popular por ter um bom desempenho em problemas de classificação binária, dividindo o espaço de atributos com hiperplanos. É um método não probabilístico que representa exemplos como pontos no espaço, mapeados de forma que os exemplos em cada categoria sejam divididos por um hiperplano. Os novos exemplos são, então, mapeados para

o mesmo espaço e previstos para pertencer a uma categoria com base em qual lado do hiperplano eles são dispostos. Este algoritmo foi implementado usando o pacote do R e1071 (MEYER et al., 2015).

Random Forest (RF) é um método de aprendizado que constrói uma grande quantidade de árvores de decisão em tempo de treinamento. A decisão final é tomada ponderando as respostas das diversas árvores. O método foi implementado usando o pacote *randomForest* do R (LIAW; WIENER, 2020).

Extreme Gradient Boosting (XGB) consiste em um conjunto de modelos de previsão fracos, geralmente árvores de decisão, e otimiza uma função de perda diferenciável. Utilizamos o pacote *xgboost* do R (CHEN et al., 2016) para implementação.

A Rede Neural (NNet) com uma única camada oculta de 10 unidades foi a última metodologia testada. Esta rede *feed-forward* foi implementada usando o pacote *nnet* do R (RIPLEY; VENABLES; RIPLEY, 2016).

A Tabela 2 descreve os principais parâmetros que utilizamos em cada um dos métodos. Para o método LR, o único parâmetro não padrão foi "família=*binomial*". Selecionamos os parâmetros para o método *XGB*, usando validação cruzada combinado com ferramentas utilizadas para automatizar o processo de ajuste dos parâmetros (*Caret Package*).

Para a avaliação do desempenho dos algoritmos de aprendizado de máquina utilizamos o AUC - a área sob a curva ROC (*Receiver Operating Characteristic*) (HANLEY; MCNEIL, 1982). O AUC têm sido adotado em problemas de classificação na área médica desde 1970. O AUC=1 indica que as previsões são perfeitamente precisas, enquanto AUC=0,5 significa que o modelo não tem capacidade de separação de classe.

Tabela 2: Principais parâmetros utilizados em cada um dos métodos de aprendizado de máquina testados.

Técnicas	**Classe**	**Parâmetros**
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 nforld = 10 metrics = "auc"
NNet	nnet	size = 10 decay = 0.001

Calculamos a métrica alfa de *Krippendorff* (*Kripp.alpha*), um índice não paramétrico que mede a concordância entre as observações (LANDIS; KOCH, 1977). Usamos o pacote *irr* do R para implementação dessa métrica. Seu valor varia de -1 a 1, onde 1 indica concordância perfeita, 0 indica nenhuma concordância e valores negativos indicam concordância inversa.

3.2 Resultados

As médias dos resultados da validação cruzada *10-fold*, usando os métodos de aprendizado de máquina clássico, estão descritos na tabela 3. Ehman et al. (EHMAN et al., 2017) obtiveram o alfa de *Krippendorff* de 0,652, enquanto que nosso melhor alfa é de 0,995. Portanto, os resultados obtidos até aqui são substancialmente melhores do que os relatados na literatura.

O método que obteve o maior AUC foi a LR (regressão logística) com área de

Tabela 3: Área abaixo da curva ROC (AUC) e do alfa de *Krippendorff*, para cada uma das técnicas de aprendizado de máquina. A Notação XY indica média X e desvio padrão Y nas 10-*folds* de validação cruzada.

	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
alfa de Kripp.	0.995	0.993	0.917	0.750	0.994

0,878±0,056, onde 0,878 é a média das áreas obtidas pela validação cruzada de 10 vezes e 0,056 é o desvio padrão. A Figura 14 mostra a curva ROC média para as *10-fold* no modelo LR, considerando bases de treino e teste. A regressão logística também produziu o alfa de *Krippendorff* mais alto, 0,995. Vale ressaltar, que não há garantia de que a AUC e o alfa de *Krippendorff* concordem que um algoritmo específico seja o melhor.

Para avaliar a qualidade do nosso método, comparamos as respostas dos técnicos juniores, com menos de 5 anos de experiência, com o nosso algoritmo, considerando as respostas dos técnicos seniores, com mais de 10 anos de experiência, como padrão ouro. A primeira linha da Tabela 4 indica que os técnicos juniores classificaram corretamente 82% de todos os furos. Classificaram corretamente apenas 34% dos furos indetectáveis e 84% dos furos detectáveis.

Para medir o desempenho do nosso algoritmo, limitamos a saída do modelo LR, que produziu os melhores resultados, usando o critério "*ROC01*", que minimiza a distância entre o gráfico ROC e o ponto (0, 1). A segunda linha da Tabela 4 indica que o modelo LR classificou corretamente 84% de todos os furos, 68% dos furos indetectáveis e 87% dos furos detectáveis.

Em conclusão, nosso algoritmo é melhor que os técnicos juniores na classificação dos furos como visíveis/invisíveis. A Tabela 5 mostra o resultado dos técnicos juniores e nosso algoritmo aplicado apenas à fatia 8, aquele com o menor contraste e,

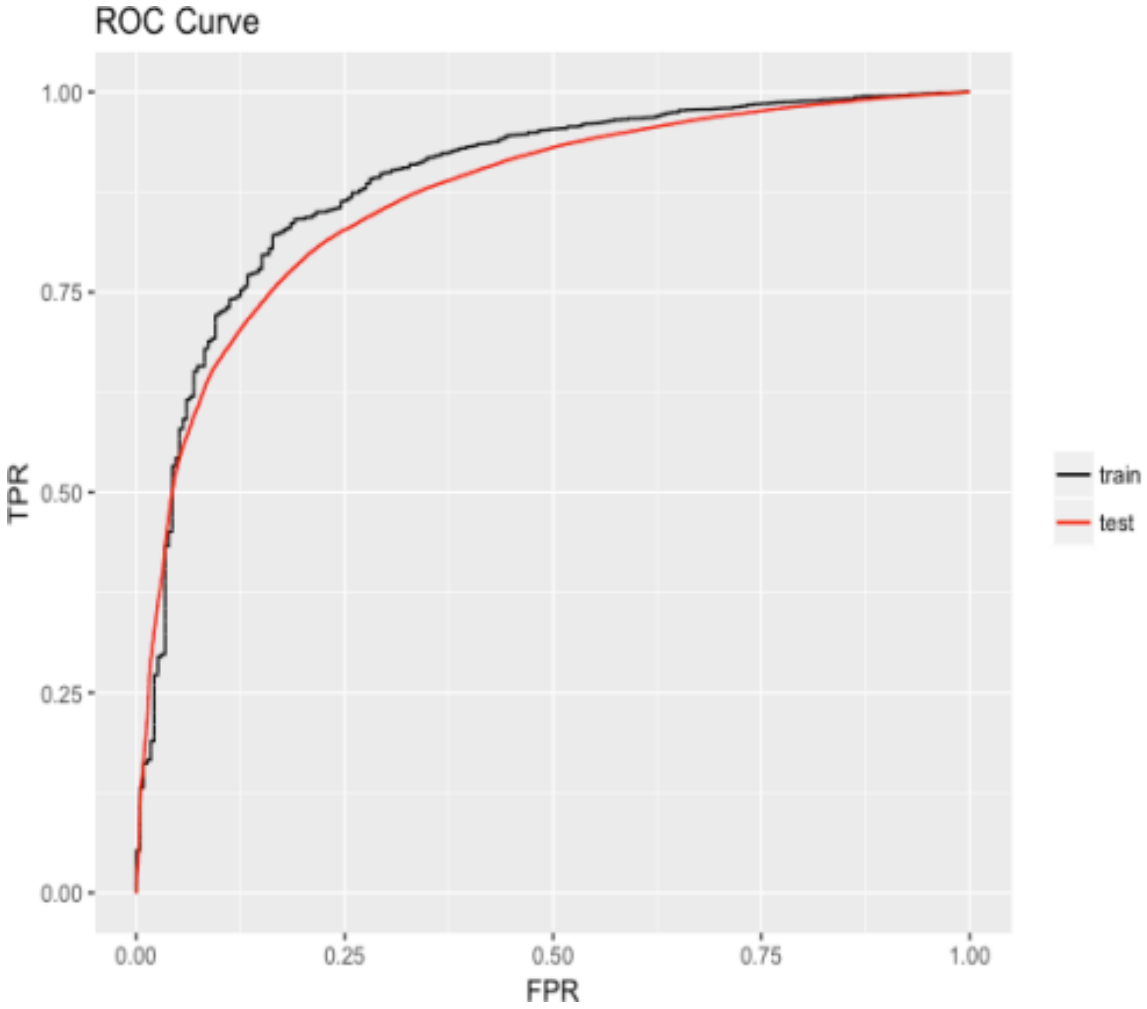

Figura 14: Curvas ROC da regressão logística (LR) para as bases de treino e teste.

Tabela 4: Acurácia, Sensibilidade e Especificidade (FAWCETT, 2006) considerando as respostas dos técnicos sêniores como padrão ouro, para as fatias (8-11). Os técnicos juniores possuem menos de 5 anos de experiência. O modelo LR teve seu limiar estabelecido com objetivo de minimizar a distância entre o gráfico ROC e o ponto (0, 1).

Experiência profissional	**Acurácia**	**Sensibilidade**	**Especificidade**
Técnicos juniores	0.824	0.343	0.844
Modelo LR	0.842	0.677	0.868

Tabela 5: Acurácia, Sensibilidade e Especificidade considerando as respostas dos técnicos sêniores como corretas, resultado aplicado para a fatia 8, a de menor contraste.

Experiência profissional	**Acurácia**	**Sensibilidade**	**Especificidade**
Técnicos juniores	0.583	0.560	0.584
Modelo LR	0.690	0.617	0.784

portanto, o mais difícil de visualizar. Ainda assim, nosso algoritmo é melhor do que os técnicos juniores (considerando as respostas dos técnicos seniores como corretas).

Notamos que mesmo técnicos seniores podem discordar entre si sobre a classificação de um furo.

4 MÉTODOS DE APRENDIZAGEM PROFUNDA

Este capítulo descreve de forma detalhada a segunda fase desta tese. Apesar dos resultados da primeira fase terem sido animadores, a acurácia resultante dos algoritmos não permitiu substituir completamente os técnicos, humanos, na execução do teste de baixo contraste do ACR. Como alternativa, avaliamos o desempenho da aprendizagem profunda (CNN) na detecção de pequenos furos de baixo contraste do *phantom*. Nesta etapa, também revisamos os rótulos (visíveis/invisíveis) atribuídos pelos técnicos, removendo aqueles com erros grosseiros.

4.1 Base de dados

4.1.1 Revisão dos rótulos

É humanamente compreensível que um técnico possa cometer erros ao rotular milhares de amostras. Desta forma, optamos por revisar os rótulos de um subconjunto de 400 imagens do conjunto original, e descartamos aquelas que continham erros grosseiros. Erros grosseiros incluem: declarar alguns furos como invisíveis em uma imagem I quando eles são claramente visíveis; rotular alguns furos em uma imagem I como visíveis quando a maioria dos outros técnicos e nós mesmos não podemos vê-los. Nestes casos, todos os rótulos dados por aquele técnico na imagem I foram descartados. Descartando todos estes erros, cada imagem foi rotulada por 2

Tabela 6: Número de imagens pelo número de técnicos que as rotularam.

							Total
Número de imagens	9	33	244	44	66	4	400
Número de técnicos	2	3	4	5	6	7	

a 7 técnicos, conforme tabela 6. Por exemplo, após descartar estes erros, 9 imagens foram rotuladas por apenas 2 técnicos, 33 imagens foram rotuladas por 3 técnicos e assim por diante. Definimos como padrão ouro a mediana das respostas dos técnicos, após descartar os erros. Este procedimento resultou em uma base de dados com 1.935 furos rotulados como invisíveis e 10.065 rotulados como visíveis.

4.2 Experimentos

4.2.1 Rede Neural Convolucional

Recentemente, houve uma verdadeira revolução na classificação de imagens com a introdução da rede neural convolucional (CNN) (LECUN et al., 1989; KRIZHEVSKY; SUTSKEVER; HINTON, 2012; LECUN; BENGIO; HINTON, 2015). Na CNN, o padrão de conectividade entre os neurônios é inspirado na organização do córtex visual dos animais (MATSUGU et al., 2003). A própria rede cria, automaticamente, a partir das imagens rotuladas contidas na amostra, filtros de baixo nível para extrair características úteis e os filtros de alto nível para concatenar adequadamente essas características. Em contrapartida, a extração de características de baixo nível é, normalmente, uma tarefa executada de forma manual nos algoritmos clássicos de aprendizado de máquina. Essa independência do conhecimento, a priori, e do esforço humano no desenvolvimento de sistemas de aprendizado de máquina, são as maiores vantagens da CNN sobre as técnicas clássicas.

Para justificar essa afirmação, basta demonstrarmos o que realizamos em cada etapa desta tese:

1. Usando aprendizagem de máquina clássica, tivemos que localizar precisamente onde estava o furo. A localização tinha que ser muito precisa, pois se considerássemos alguns pixels fora do furo como pertencendo ao furo, a média do sinal, por exemplo, ficaria errada. Em seguida, tivemos que desenvolver técnicas para fazer a extração de características, como: sinal dentro do furo, ruído dentro do furo, sinal em torno do furo e ruído em torno do furo. Extraímos também outras características e percebemos que elas eram pouco úteis para a o problema de classificação. Finalmente, testamos vários algoritmos de aprendizado de máquina com diferentes parâmetros, para verificar qual funcionava melhor. Tudo isso foi feito manualmente e o resultado ao qual chegamos foi inferior à CNN.

2. Usando a CNN, tivemos simplesmente que extrair as ROIs em torno do centro do furo. A localização do furo podia ser aproximada, pois um pequeno erro de localização no mesmo seria automaticamente ajustado pela CNN. Testamos algumas arquiteturas de CNN adequadas juntamente com parâmetros. Depois, a CNN fez tudo de forma independente para classificar os furos. Exigindo assim, menos esforço humano.

Nesta tese, desenvolvemos uma solução a partir de uma rede CNN implementada em *Keras/TensorFlow* (ABADI et al., 2015). A rede tem como objetivo prever a visibilidade de cada um dos furos. Nosso sistema faz a leitura de um arquivo do tipo *CSV* (*Comma-Separated Values*), que contém as coordenadas centrais dos furos e seus rótulos no padrão ouro, respectivamente. Além disso, o sistema recebe 400 imagens de 16 bits, com 256×256 *pixels* (com 12 bits significativos), correspondentes

a 100 aquisições de MRI. A partir destes dados, CSV e imagens, realizamos a extração de 12.000 pequenas regiões de interesse, com 17×17 *pixels*, ao redor do centro de cada furo. Cada região possui seu rótulo de visibilidade, conforme descrito em detalhes na Figura 15. O conjunto formado pelas 12.000 regiões de interesse (ROIs), com seus respectivos rótulos, é utilizado na construção das amostras de treino e teste do nosso problema.

Utilizamos a técnica de validação cruzada *5-fold* para obter métricas de desempenho mais robustas, juntamente com seus desvios-padrões. Desta forma, dividimos aleatoriamente as 12.000 ROIs em 5 subconjuntos de 2.400 cada. Em cada partição, selecionamos 4 subconjuntos como amostra de treinamento e o subconjunto restante como amostra de teste.

Conforme citado anteriormente, as regiões de interesse são imagens muito pequenas com 17×17 *pixels*. Portanto, optamos por utilizar uma arquitetura simplificada para a tarefa de classificação dos furos. Um processo de normalização foi realizado. Primeiro, calculamos a média μ e o desvio padrão σ dos *pixels* da base de treinamento para, em seguida, normalizar os *pixels* das bases de treinamento e teste $P_n = (P_o - \mu)/\sigma$, onde P_o é o valor do *pixel* original de 12 bits sem sinal e P_n é o valor de *pixel* normalizado de 32 bits em ponto flutuante.

Em seguida, realizamos um processo de *data augmentation* (SHORTEN; KHOSHGOFTAAR, 2019). Para cada uma das 9.600 regiões de interesse do conjunto de treinamento deslocamos um *pixel* nas direções norte, sul, leste e oeste, mantendo o original. Portanto, as 9.600 regiões de interesse se tornaram 5×9.600 = 48.000 regiões de interesse. Optamos por não utilizar deformações geométricas sofisticadas em função das imagens serem muito pequenas.

Nesta tese, testamos algumas arquiteturas e parâmetros, e apresentamos aquela

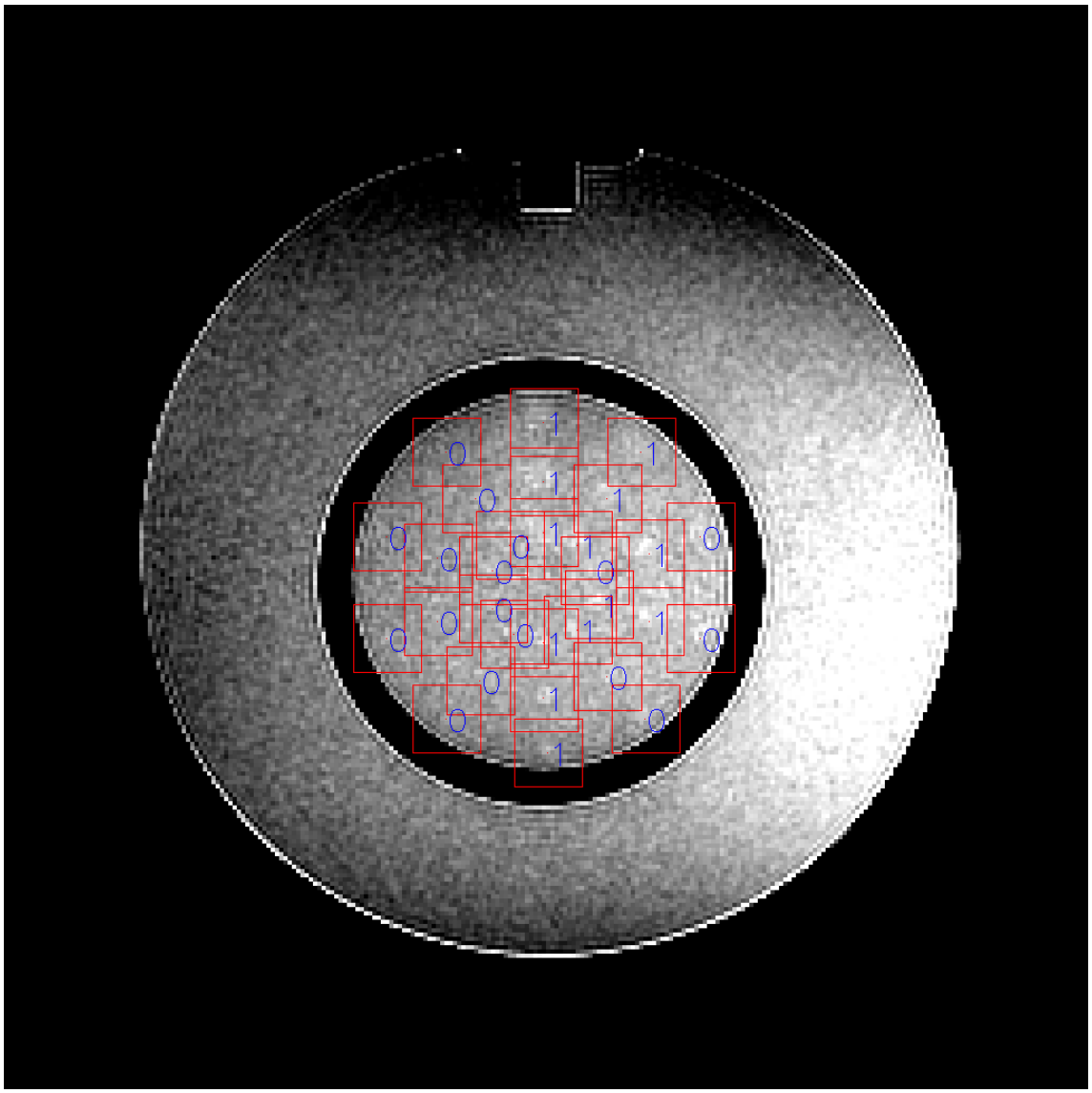

Figura 15: Extraímos ROIs (Regiões de Interesse) com 17×17 *pixels* ao redor do centro de cada furo. Os rótulos 0/1 indicam se o furo é invisível/visível.

que demonstrou o melhor desempenho. Testamos as arquiteturas mais usuais de CNN. Por exemplo, o mais comum nesse caso é usar *relu*, de forma que não testamos outras funções de ativação, inclusive em função dos resultados já obtidos. Vários parâmetros foram escolhidos empiricamente, descrevendo os que culminaram no melhor resultado. Assim, usamos uma rede CNN inspirada no modelo de arquitetura VGG (*Visual Geometry Group* (SIMONYAN; ZISSERMAN, 2014)), representada na Figura 16. A arquitetura é composta por três blocos (retângulos em azul na Figura 16) sequenciais, cada um com a seguinte estrutura interna:

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

Nesta estrutura, o número de camadas convolucionais é de n = 64, 96 e 128, no primeiro, segundo e terceiro blocos inspirados por VGG, respectivamente. Todas as camadas convolucionais são seguidas por uma função de ativação *relu*, e utilizam um regularizador L_2 de *kernel* com parâmetro $5{\times}10^{-4}$.

Todas as camadas convolucionais usam *kernel* 3×3, com o *padding* tipo *same* (para manter as resoluções de entrada e saída iguais), exceto a segunda camada convolucional do primeiro bloco, que usa tamanho de *kernel* 2×2, com *padding* tipo *valid*, para reduzir a resolução da imagem de 17×17 para 16×16, se tornando divisível por 2, adequado para uma sequência de várias camadas *max-poolings* 2×2.

As convoluções selecionam automaticamente os filtros adequados para extrair as características relevantes e combiná-las adequadamente. As camadas de norma-

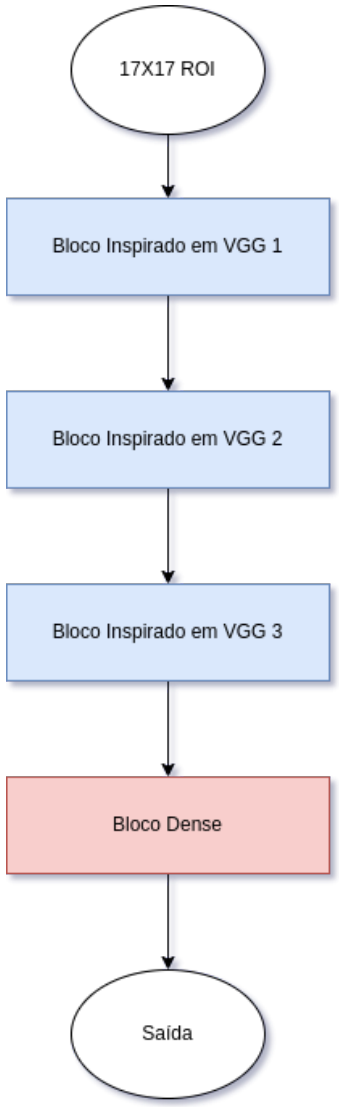

Figura 16: Arquitetura da rede CNN utilizada na classificação da visibilidade da região de interesse.

lização em lote auxiliam na convergência do processo de aprendizado e a camada de *dropout* ajuda a evitar o *overfitting*. As camadas *max-pooling* diminuem a resolução das características extraídas.

As saídas dos blocos VGG são 64 mapas de características com formato 8×8 (após o primeiro bloco), 92 mapas de características com formato 4×4 (após o segundo bloco) e 128 mapas de características com formato 2×2 (após o terceiro bloco). Esses 128 mapas são "achatados", ou seja, convertidos em um vetor 1-D com 2×2×128=512 valores, e passam por um bloco *dense* (retângulo vermelho na Figura 16) composto por duas camadas *dense* totalmente conectadas:

```
Flatten()
Dense(128, activation='relu')
BatchNormalization()
Dropout(0.3)
Dense(32, activation='relu')
Dense(1, activation='linear')
```

A saída do bloco denso é um número entre 0 e 1, de modo que quanto mais próximo de 1, maior a probabilidade de que o furo seja visível. Adotamos o *erro quadrático médio* como função de perda e ADAM (*Adaptive Moment Estimation*) como otimizador. O otimizador começa com sua taxa de aprendizado padrão de 0,001, que é reduzida pelo fator de 0,9 sempre que atinge um platô de acurácia no processo de treinamento do algoritmo. Usamos o tamanho de lote (*batch size*) de 32 e treinamos a CNN com 150 épocas (*epochs*). Executamos 150 épocas porque o desempenho do modelo parou de melhorar em torno deste ponto. Não usamos o desempenho nos dados de teste para decidir precisamente quando parar o treino, o que seria um "vazamento de informação" de treino para teste.

4.3 Resultados

Nosso sistema, como a maioria dos sistemas de classificação, não retorna uma resposta binária, ele retorna um valor entre 0 e 1. Quanto mais próximo de 1, maior a probabilidade do furo ser visível. Portanto, não é possível calcular a sensibilidade e especificidade de forma direta a partir da resposta obtida. Para isso, é necessário primeiro definir um limiar, para se obter em seguida uma resposta *booleana* e assim calcular a sensibilidade e especificidade (assim como os erros tipo I e II). A curva ROC traça as sensibilidades e especificidades obtidas, variando o limiar para todos os valores possíveis no intervalo entre 0 e 1, enquanto que, a área sob a curva ROC (AUC) mede o desempenho do sistema, independentemente do limiar escolhido. Usando a curva ROC é possível calcular a sensibilidade para uma determinada especificidade, ou ainda, a especificidade para uma determinada sensibilidade. Há um ponto especial na curva ROC, chamado de *Equal Error Rate* (EER) (POH; BENGIO, 2004), onde a acurácia, sensibilidade e a especificidade se tornam iguais. Nesse ponto é possível calcular acurácia, sensibilidade e especificidade sem escolher um valor para o limiar e, ao mesmo tempo, obter uma métrica que tenha uma interpretação intuitiva.

Sem a utilização da técnica de *Test-Time Augmentation* (TTA) (WANG et al., 2019), obtivemos os resultados descritos na Tabela 7. Resumindo, nosso sistema produziu um AUC médio de 0,981±0,003, com um *equal error rate* (EER) médio de 7.2±0.9%, ou seja, a sensibilidade, especificidade no ponto EER são todas iguais a 92.8%. A Figura 17 mostra as curvas ROC (*Receiver Operating Characteristic*) obtidas sem o uso da TTA.

Com a utilização da técnica TTA, obtivemos resultados ainda melhores. Como fizemos com as imagens de treinamento, deslocamos cada imagem de teste nas di-

Tabela 7: AUCs e EERs com validação cruzada *5-fold* e sem a utilização do *TTA*.

	fold1	fold2	fold3	fold4	fold5	mean±std
EER	0.063	0.066	0.068	0.074	0.087	0.072±0.009
Ac., sens., esp. em 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

Tabela 8: AUCs e EERs com validação cruzada *5-fold* com a utilização do *TTA*.

	fold1	fold2	fold3	fold4	fold5	mean±std
EER	0.063	0.066	0.068	0.074	0.087	0.068±0.007
Ac., sens., esp. em 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

reções norte, sul, leste e oeste. Assim, cada imagem de teste gerou 5 imagens (4 distorcidas mais a original). Todas essas imagens foram utilizadas como entrada ao sistema de IA e calculamos a média das 5 previsões. Os resultados são mostrados na Tabela 8. Todas as métricas de desempenho melhoraram ligeiramente: AUC aumentou de 0,981 para 0,9833; acurácia, sensibilidade e especificidade no ponto EER aumentaram de 92,8% para 93,2%. A Figura 18 mostra as curvas ROC obtidas com TTA. O AUC obtido é bastante alto e o desvio padrão é bastante baixo, o que significa que resultados semelhantes são obtidos ao repetir os experimentos.

Resumindo, nosso sistema produziu um AUC de 0,983±0,003 com um EER de 6,8±0,7%, ou seja, a sensibilidade, especificidade no ponto EER são todas iguais a 100-6,8=93,2%. A Figura 18 mostra as curvas ROC (*Receiver Operating Characteristic*) obtidas.

As medidas de desempenho dos 4 técnicos que rotularam todas as imagens do conjunto de dados são mostradas na Tabela 9 (outros técnicos rotularam apenas partes do conjunto de dados). Para calcular o desempenho de um técnico T, não

Tabela 9: Acurácia, sensibilidade e especificidade obtidas pelos quatro técnicos.

	Técnico 1	**Técnico 2**	**Técnico 3**	**Técnico 4**
Acurácia	**0.898**	**0.878**	0.924	0.883
Sensibilidade	0.940	0.873	0.961	0.878
Especificidade	0.690	0.904	0.757	0.905

podemos utilizar o mesmo padrão ouro que usamos para medir o desempenho do nosso sistema, pois as próprias respostas do técnico T estariam sendo consideradas no cálculo. Assim, para computar o desempenho de um técnico T, utilizou-se como padrão ouro a mediana das respostas dos técnicos, excluindo a resposta do próprio T. Como já descrito antes, os erros grosseiros foram descartados do cálculo do padrão ouro.

Podemos calcular o AUC do nosso sistema, já que ele gera um número entre 0 e 1. Os técnicos dão respostas binárias (visíveis ou invisíveis), a partir das quais podemos calcular a acurácia, sensibilidade e especificidade, mas não é possível calcular a AUC. A acurácia não é uma boa medida de desempenho para este problema porque nosso conjunto de dados é altamente desbalanceado. Temos muito mais furos visíveis (10.065) do que invisíveis (1.935), portanto, um sistema ou técnico com propensão a classificar os furos como visíveis obterá uma maior acurácia quando comparado a outro com propensão a classificar os furos como invisíveis. Sensibilidade e especificidade também não são boas medidas de desempenho neste caso, pois há um *trade-off* entre as duas, de forma que, aumentar uma faz com que a outra diminua. Portanto, para comparar de forma justa as respostas dos técnicos com a resposta do sistema, não usaremos acurácia, sensibilidade ou especificidade. Ao invés disso, marcamos os pontos no gráfico de especificidade-sensibilidade dos 4 técnicos sobre a curva ROC do nosso sistema. As quatro marcas "X" em vermelho, verde, azul e magenta nas Figuras 17 e 18 representam os desempenhos dos técnicos 1 a 4, respectivamente. Como todas as curvas ROC do sistema de IA estão acima

Tabela 10: Número de imagens pelo número de técnicos que as rotularam.

				Total
Número de Imagens	29	11	320	360
Número de Técnicos	1	2	3	

desses quatro pontos, podemos concluir que o sistema tem um desempenho melhor do que qualquer técnico individualmente.

4.3.1 Testes em uma base de dados independente

Para testar a robustez da solução, utilizamos o conjunto (*ensemble*) dos cinco modelos obtidos anteriormente, para classificar um conjunto de teste completamente independente. Quando se combinam vários modelos, precisamos de um método que faça a agregação das respostas dos algoritmos. Desta forma, estamos calculando a média das respostas dos modelos, antes do processo de limiarização. Outras técnicas de agregação são a moda e a mediana.

Este conjunto de dados é formado por 90 aquisições com o *phantom* ACR, totalizando 90×4=360 imagens com 360×30=10.800 furos. Três técnicos T_1, T_2 e T_3 classificaram cada furo como visível ou invisível. Como fizemos anteriormente, revisamos a rotulagem de cada imagem e descartamos aquelas que continham erros grosseiros. Descartando erros grosseiros, cada imagem foi rotulada por 1 a 3 técnicos, conforme a Tabela 10. Um modelo *ensemble* usa vários modelos para obter melhor desempenho preditivo do que poderia ser obtido por qualquer modelo individualmente. Neste caso, o modelo *ensemble* calcula a média das respostas dos cinco modelos.

Utilizamos como padrão ouro a mediana das respostas dos técnicos, após descartar os erros grosseiros, arredondando para 1 em caso de empate. Este procedimento

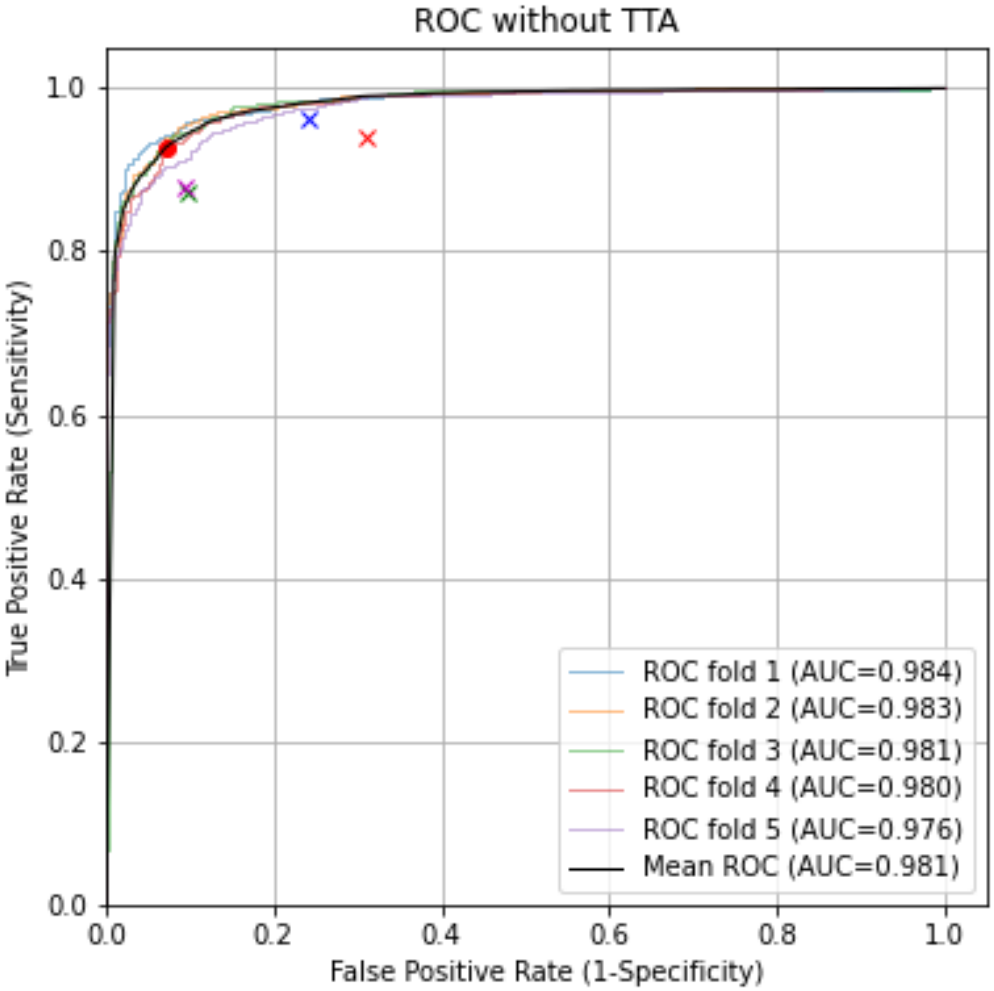

Figura 17: Curvas ROC obtidas na validação cruzada de *5-fold* sem *TTA* (cores esmaecidas) e curva ROC média (preto). O ponto vermelho indica o ponto de EER, onde a sensibilidade e especificidade se tornam iguais. As quatro marcas 'X' indicam os pontos de especificidade/sensibilidade dos quatro técnicos que rotularam todas as imagens do conjunto de dados.

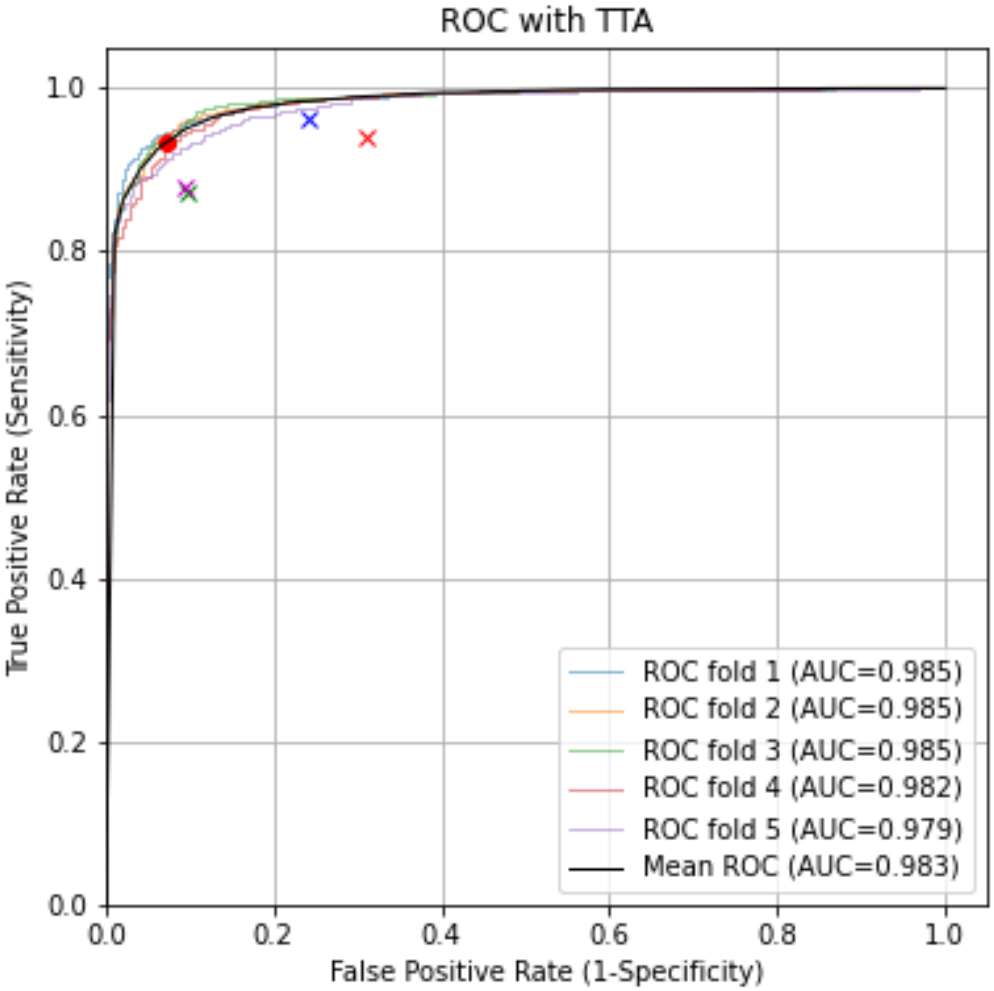

Figura 18: Curvas ROC obtidas na validação cruzada de *5-fold* com *TTA* (cores esmaecidas) e curva ROC média (preto). O ponto vermelho indica o ponto de EER, onde a sensibilidade e especificidade se tornam iguais. As quatro marcas 'X' indicam os pontos de especificidade/sensibilidade dos quatro técnicos que rotularam todas as imagens do conjunto de dados.

resultou em 1.821 furos rotulados como invisíveis e 8.979 rotulados como visíveis. Usando o modelo *ensemble* e TTA (4 imagens deslocadas mais o original), obtivemos a curva ROC representada na Figura 19. Marcamos os pontos no gráfico de especificidade-sensibilidade dos técnicos T_1 e T_2 na curva ROC do nosso sistema como marcas "X". Como nossas curvas ROC estão acima desses pontos, concluímos que nosso sistema tem um desempenho melhor que os dois técnicos. Para calcular o desempenho de um técnico T, foi utilizada como padrão ouro a mediana das respostas dos técnicos, excluindo a resposta do próprio técnico T. Não foi possível computar o desempenho do técnico T_3 porque havia algumas imagens que foram rotuladas apenas por ele mesmo (após descartarmos os erros).

De acordo com o manual do ACR (RADIOLOGY et al., 2015), um *spoke* é considerado visível se, e somente se, todos os seus três furos estiverem visíveis. Calculamos os erros médios de classificação de *spokes* por aquisição, obtendo as taxas de erro descritas na Tabela 11. A taxa de erro do nosso sistema usando um limiar adequado (6,17% de erro com limiar 0.68) é muito inferior aos dos técnicos T_1 e T_2 (12,31% e 14,67%). Nas duas últimas colunas, escolhemos limiares para resultar em casos falsos positivos (51 e 42) semelhantes aos dos técnicos (51 e 43). Mesmo nesta situação, as taxas de erro do nosso sistema (8,06% e 8,92%) são substancialmente inferiores às dos técnicos T_1 e T_2 (12,31% e 14,67%).

Em nossos dados de teste, 66 aquisições foram feitas em máquinas 1,5T e 24 em máquinas 3T. Usando o critério de que o número n de *spokes* visíveis deve ser $n \geq 28$ e $n \geq 37$ para aprovar respectivamente máquinas de 1,5T e 3T, o padrão ouro teria aprovado 49 (74%) de máquinas 1,5T e 10 (42%) de máquinas 3T .

Os dois técnicos e o sistema de IA discordaram do padrão ouro na aprovação ou reprovação das máquinas, conforme descrito na Tabela 12. Em todos os casos, o sis-

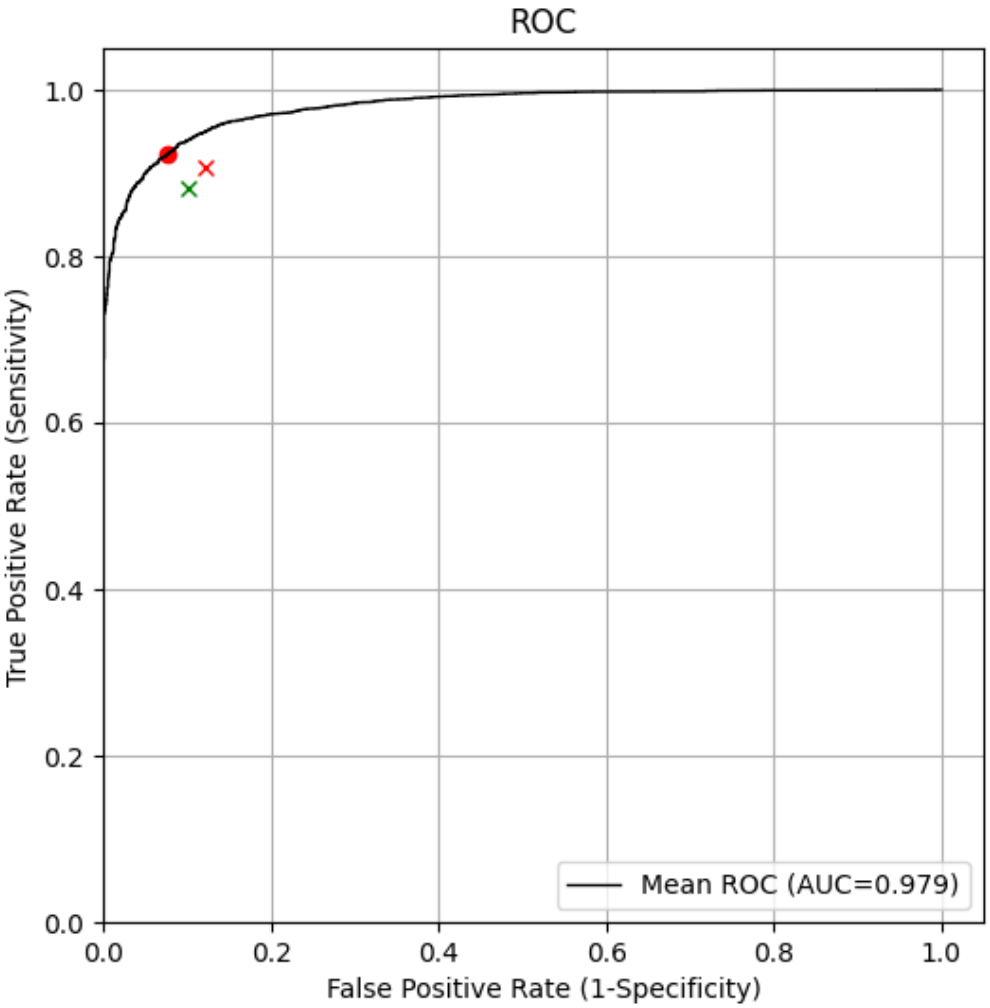

Figura 19: Curva ROC obtida do treinamento do modelo *ensemble* sobre o conjunto independente de teste com a utilização de *TTA*. Os dois símbolos ("X") marcam o ponto de sensibilidade/especificidade de cada técnico.

Tabela 11: Métricas da classificação dos *spokes* no conjunto de teste independente por técnicos T_1 e T_2, e por *ensemble* de modelos usando diferentes valores de limiar. De acordo com o padrão ouro, existem 2.779 *spokes* visíveis e 821 *spokes* invisíveis.

	T_1	T_2	Solução Proposta				
Limiar			0.66	0.68	0.70	0.84	0.87
VP	2516	2433	2654	2648	2638	2540	2500
VN	641	639	723	730	735	770	779
FP	51	43	98	91	86	51	42
FN	392	485	125	131	141	239	279
Erros	443	528	223	222	227	290	321
Taxa de Erro	12.31%	14.67%	6.19%	**6.17%**	6.31%	8.06%	8.92%

Tabela 12: Aprovação/reprovação dos equipamentos de MRI pelos dois técnicos e pelo modelo utilizando o limiar de 0.68.

	1.5T				3T			
	FP	FN	Erros	Taxa de Erro	FP	FN	Erros	Taxa de Erro
T_1	1	7	8	12%	0	8	8	33%
T_2	1	14	15	23%	0	9	9	38%
Solução Proposta	5	3	8	12%	1	3	4	17%

tema de IA discordou menos ou igualmente do padrão ouro do que os técnicos T_1 ou T_2. A maioria dos "erros" cometidos pelos técnicos é do tipo falso negativo, quando rejeitam uma máquina que teria sido aprovada pelo padrão ouro. Isto significa que T_1 e T_2 classificaram como invisíveis muitos furos que T_3 considerava visíveis. Há técnicos que tendem a considerar os furos como visíveis ou invisíveis.

Observe que o padrão ouro está longe de ser infalível, pois é apenas a mediana das opiniões dos técnicos, eliminando as respostas com erros grosseiros. Além disso, apenas 3 técnicos rotularam o conjunto de dados de teste e algumas imagens foram rotuladas por apenas 1 ou 2 técnicos.

5 DISCUSSÕES

5.1 Padrão Ouro

Na primeira fase, na abordagem com os métodos clássicos de aprendizado de máquina, consideramos as respostas dos técnicos seniores, com mais de 10 anos de experiência, como o padrão ouro. No entanto, analisando cuidadosamente nosso conjunto de dados, concluímos que os técnicos seniores cometem tantos erros grosseiros quanto os técnicos menos experientes. Portanto, anos de experiência parecem não garantir, por si só, maior precisão na classificação. Assim, na segunda fase, alteramos o padrão ouro de visibilidade dos furos em uma imagem I para a mediana das respostas de todos os técnicos (independente dos anos de experiência) que não cometeram erros grosseiros na classificação dos furos em I. Se T cometeu alguns erros grosseiros na classificação dos furos em I, todos os rótulos em I fornecidos por T foram descartados.

5.2 CNN com índices da Região de Interesse

Outro teste que realizamos foi colocar os índices da ROI como entrada da rede CNN, além da própria imagem. O índice é composto por três números: fatia (de 8 a 11), ângulo (ou *spoke*, de 1 a 10) e posição radial (1 a 3) - ver Figura 12. A lógica do teste reside no fato de que, intuitivamente, essas informações podem auxiliar na

classificação:

1. Fatia - o contraste da imagem depende desse número;
2. Ângulo (*spoke*) - o diâmetro do furo depende desse número;
3. Posição - geralmente, os furos externos são mais distorcidos e difíceis de visualizar quando comparados aos furos internos.

Esses números foram normalizados para variar no intervalo de -1 a +1 antes de entrar na rede, passando por uma camada densa e sendo concatenado com 512 características extraídas da imagem. Ao contrário do esperado, não obtivemos nenhuma melhora com essa modificação. Isso pode significar que a CNN é capaz de extrair essas informações da própria imagem.

5.3 Extração manual de características e métodos clássicos de aprendizado de máquina

Na primeira fase deste trabalho, extraímos manualmente algumas características das imagens da ROI e usamos algoritmos convencionais de aprendizado de máquina para atingir o AUC máximo de 0,878. Na segunda fase, testamos uma rede neural convolucional e conseguimos obter um AUC muito maior (0,983). No entanto, os dois trabalhos não são diretamente comparáveis porque usam diferentes rótulos para o padrão ouro. Para comparar de forma correta as duas abordagens, repetimos os experimentos anteriores usando o novo conjunto de dados.

Como na primeira fase (RAMOS; KIM; TANCREDI, 2018), usamos as quatro características extraídas das ROIs:

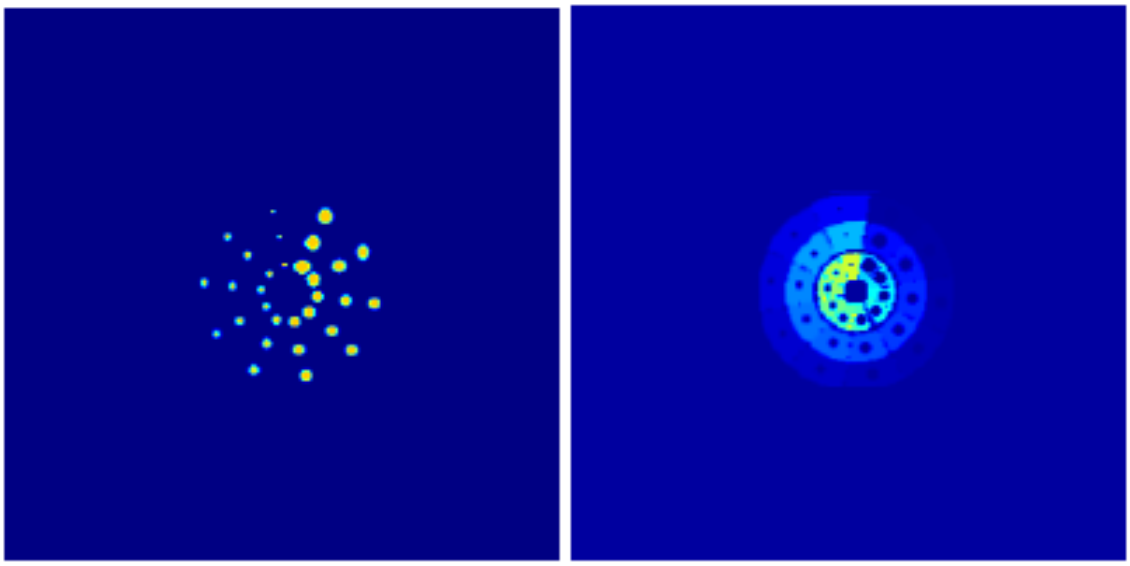

Figura 20: As máscaras utilizadas para calcular a média e o desvio padrão dentro dos furos (esquerda) e no entorno (direita).

- S_{in}: O sinal médio (valor médio) dentro do furo representado como uma variável (*float32*) normalizada para o intervalo entre 0 e 1.
- N_{in}: O ruído (desvio padrão) dentro do furo normalizado para o intervalo entre 0 e 1.
- S_{out}: O sinal médio na área do entorno ao furo, normalizado para o intervalo entre 0 e 1.
- N_{out}: O ruído na área do entorno ao furo, normalizado para o intervalo entre 0 e 1.

A Fig. 20 mostra as máscaras usadas para calcular essas variáveis. Assim como em (RAMOS; KIM; TANCREDI, 2018), também utilizamos como características os três índices ROI: fatia (de 8 a 11), ângulo (1 a 10) e posição radial (1 a 3). Testamos os cinco algoritmos clássicos de aprendizado de máquina usados na fase 1: Regressão Logística, *SVM*, *Random Forest*, *Multilayer Perceptron* e *Extreme Gradient Boosting*, fornecidos pela biblioteca *Scikit-Learn*.

Tabela 13: Média de resultados de validação cruzada com *5-fold* usando algoritmos clássicos de aprendizado de máquina com e sem os três índices de ROI (fatia, ângulo e posição).

	Com índices da ROI				**Sem índices**
	MSE	**EER**	**Ac.em EER**	**AUC**	**AUC**
Reg. Logística	**0.098±0.006**	**0.166±0.008**	**0.834±0.008**	**0.906±0.006**	**0.716±0.020**
SVM	**0.099±0.005**	0.175±0.013	0.825±0.013	0.892±0.006	0.724±0.032
Random Forest	0.088±0.004	0.126±0.005	0.874±0.005	0.942±0.002	**0.888±0.011**
Mult. Perceptron	0.101±0.009	**0.164±0.009**	**0.836±0.009**	**0.908±0.006**	**0.747±0.016**
Extr. Grad. Boost.	0.091±0.004	**0.126±0.004**	0.874±0.004	0.943±0.003	**0.884±0.009**

As médias e os desvios dos resultados utilizando validação cruzada com *5-fold* são descritos na Tabela 13.

Random Forest e *Extreme Gradient Boosting* produziram bons resultados (AUCs de 0,942±0,002 e 0,943±0,003), mas substancialmente inferiores ao CNN (AUC de 0,983±0,003). É possível tirar a mesma conclusão da precisão, sensibilidade e especificidade no ponto EER: *Random Forest* e *Extreme Gradient Boosting* alcançaram 87,4±0,5%, enquanto que CNN alcançou 93,2±0,7%. Resultados mencionados acima se encontram no capítulo 4, página 81. Ao contrário da CNN, os algoritmos clássicos de aprendizado de máquina parecem depender fortemente dos três índices de ROI. Resultados consideravelmente piores são obtidos quando essas informações são retiradas do modelo, que pode ser visto comparando as duas últimas colunas da Tabela 13.

6 CONCLUSÕES

Nesta tese, demonstramos que é possível automatizar o teste de resolução de baixo contraste do programa do Colégio Americano de Radiologia (ACR). Aparentemente, este é o primeiro trabalho que realmente consegue emular a percepção de um observador humano neste contexto utilizando aprendizado de máquina.

Na primeira fase do trabalho, para treinar os algoritmos de classificação, uma base de dados foi estruturada a partir de 620 aquisições de imagens do *phantom* ACR realizadas ao longo de 12 meses, onde se podem observar 74.400 furos de baixo contraste, capturados em uma ampla variedade de condições. Extraímos alguns atributos dessas imagens, nível médio do sinal e ruído dentro e fora dos furos. Com estes dados, alimentamos cinco algoritmos clássicos de aprendizado de máquina e com rótulos (detectáveis/indetectáveis) atribuídos por técnicos seniores, com mais de 10 anos de experiência. Nesta fase consideramos as respostas dos técnicos seniores, com mais de 10 anos de experiência, como nosso padrão ouro.

Entre os cinco métodos clássicos testados, a Regressão Logística apresentou a maior área sob a curva ROC (0,878) e o maior alfa de *Krippendorff* (0,995). Os resultados alcançados nesta fase já são substancialmente melhores do que os relatados anteriormente descritos na literatura. Além disso, os resultados são melhores do que aqueles obtidos quando técnicos juniores, com menos de cinco anos de experiência, rotulam manualmente os furos da imagem.

Para a segunda fase, criamos um conjunto de dados com 100 aquisições *phantom* ACR, totalizando 12.000 furos. Da mesma forma que na fase anterior, os técnicos rotularam cada furo como visível ou invisível. Desta vez, consideramos a mediana das respostas dos técnicos como padrão ouro, pois analisando cuidadosamente nosso conjunto de dados, concluímos que os técnicos seniores cometem tantos erros quanto os técnicos menos experientes. Dividimos o conjunto de dados em 5 subconjuntos e usamos validação cruzada *5-fold*, para treinar e testar o sistema baseado em rede neural convolucional. Obtivemos um AUC médio, área sob curva ROC, de 0,983±0,003 e uma acuracidade média de 93,2±0,7% no ponto de EER, que é melhor do que qualquer um dos resultados obtidos pelos técnicos individualmente.

Repetimos os experimentos usando um conjunto de teste independente, obtendo um AUC de 0,979. A classificação dos *spokes* pelo sistema de IA concorda com o padrão ouro mais do que com qualquer técnico individualmente. As decisões do sistema de aprovar ou rejeitar mais máquinas de ressonância magnética também concordam mais ou igualmente com o padrão ouro do que com as decisões tomadas por qualquer técnico individualmente. Esses resultados mostram que esse teste pode ser automatizado com confiança usando a CNN.

Também usamos a mesma metodologia da primeira fase para classificar os dados da segunda fase, obtendo um AUC médio de 0,943±0,003 e uma acuracidade média de 87,4±0,4% no ponto de EER. Esses resultados mostram que a rede CNN proposta é superior aos algoritmos clássicos de aprendizado de máquina, usando características extraídas manualmente a partir de máscaras.

Portanto, a solução proposta neste trabalho atingiu os objetivos inicialmente sugeridos.

REFERÊNCIAS

ABADI, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Software available from tensorflow.org. Disponível em: <https://www.tensorflow.org/>.

ALAYA, I. B.; MARS, M. Automatic analysis of acr phantom images in mri. Current Medical Imaging, Bentham Science Publishers, v. 16, n. 7, p. 892–901, 2020.

BONACCORSO, G. Machine learning algorithms. [S.l.]: Packt Publishing Ltd, 2017.

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

CHEN, T. et al. xgboost: Extreme gradient boosting. R package version 0.4-4. [S.l.]: Accessed, 2016.

DAVIDS, M. et al. Fully-automated quality assurance in multi-center studies using mri phantom measurements. Magnetic resonance imaging, Elsevier, v. 32, n. 6, p. 771–780, 2014.

EHMAN, M. O. et al. Automated low-contrast pattern recognition algorithm for magnetic resonance image quality assessment. Medical physics, Wiley Online Library, v. 44, n. 8, p. 4009–4024, 2017.

FAWCETT, T. An introduction to roc analysis. Pattern recognition letters, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FITZPATRICK, A. O. Automated Quality Assurance for Magnetic Resonance Imaging with Extensions to Diffusion Tensor Imaging. Tese (Doutorado) — Virginia Tech, 2005.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology, v. 143, n. 1, p. 29–36, 1982.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. Applied logistic regression. [S.l.]: John Wiley & Sons, 2013. v. 398.

KIMPE, T.; TUYTSCHAEVER, T. Increasing the number of gray shades in medical display systems—how much is enough? Journal of digital imaging, Springer, v. 20, n. 4, p. 422–432, 2007.

KRIPPENDORFF, K. Computing krippendorff's alpha-reliability. 2011.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, v. 25, 2012.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. Communications of the ACM, AcM New York, NY, USA, v. 60, n. 6, p. 84–90, 2017.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. biometrics, JSTOR, p. 159–174, 1977.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. nature, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LECUN, Y. et al. Backpropagation applied to handwritten zip code recognition. Neural computation, MIT Press, v. 1, n. 4, p. 541–551, 1989.

LIAW, A.; WIENER, M. Classification and regression based on a forest of trees using random inputs. R Package, 2020.

MATLAB, M.; SIMULINK. MATLAB. (acessado: 07.11.2022). Disponível em: <https://www.mathworks.com/products/matlab.html>.

MATSUGU, M. et al. Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Networks, Elsevier, v. 16, n. 5-6, p. 555–559, 2003.

MEYER, D. et al. Package 'e1071': Misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien. R package version 1.6–7, 2015. 2015.

PANYCH, L. 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, Wiley Online Library, v. 43, n. 4, p. 843–852, 2016.

POH, N.; BENGIO, S. Evidences of equal error rate reduction in biometric authentication fusion. [S.l.], 2004.

R-PROJECT. The R Project for Statistical Computing. (acessado: 07.11.2022). Disponível em: <https://www.r-project.org/>.

RADIOLOGY, A. C. of et al. Magnetic resonance imaging quality control manual. Reston, VA: American College of Radiology, 2015.

RAMOS, J. E.; KIM, H. Y.; TANCREDI, F. Automation of the acr mri low-contrast resolution test using machine learning. In: IEEE. 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). [S.l.], 2018. p. 1–6.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. Encyclopedia of database systems, Springer, v. 5, p. 532–538, 2009.

RIPLEY, B.; VENABLES, W.; RIPLEY, M. B. Package 'nnet'. R package version, v. 7, n. 3-12, p. 700, 2016.

ROSE, A. A unified approach to the performance of photographic film, television pickup tubes, and the human eye. Journal of the Society of Motion Picture Engineers, SMPTE, v. 47, n. 4, p. 273–294, 1946.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. Journal of big data, SpringerOpen, v. 6, n. 1, p. 1–48, 2019.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

SUN, J. et al. An open source automatic quality assurance (osaqa) tool for the acr mri phantom. Australasian physical & engineering sciences in medicine, Springer, v. 38, n. 1, p. 39–46, 2015.

TERAMOTO, A. et al. Estimating subjective evaluation of low-contrast resolution using convolutional neural networks. Physical and Engineering Sciences in Medicine, Springer, v. 44, n. 4, p. 1285–1296, 2021.

WANG, G. et al. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing, Elsevier, v. 338, p. 34–45, 2019.

WARING, J.; LINDVALL, C.; UMETON, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial intelligence in medicine, Elsevier, v. 104, p. 101822, 2020.