UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO
DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA

RAFAEL SILVA DEL LAMA

Algoritmos Genéticos e Redes Neurais Convolucionais para Auxílio ao Diagnóstico de Fraturas Vertebrais por Compressão

Ribeirão Preto–SP

2020

RAFAEL SILVA DEL LAMA

Algoritmos Genéticos e Redes Neurais Convolucionais para Auxílio ao Diagnóstico de Fraturas Vertebrais por Compressão

Versão Original

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) da Universidade de São Paulo (USP), como parte das exigências para a obtenção do título de Mestre em Ciências.

Área de Concentração: Computação Aplicada.

Orientador: Renato Tinós

Ribeirão Preto–SP

2020

Agradecimentos

Agradeço à minha família por todo o carinho, amor e força. Sou grato, especialmente, aos meus pais, Carlos e Eliana, que tanto lutaram pela minha educação e sempre me apoiaram proporcionando a tranquilidade e o conforto que tanto precisava para vencer esta etapa. Agradeço também à minha irmã Gabriela e minha namorada Carol por me apoiarem nos momentos difíceis. Sem a força de vocês eu não conseguiria seguir em frente.

A todos os amigos, meu muito obrigado pelos inúmeros conselhos, frases de motivação e puxões de orelha. As risadas, que vocês compartilharam comigo nessa etapa tão desafiadora da vida acadêmica, também fizeram toda a diferença.

Ao Prof. Dr. Zhao Liang, Prof. Dr. Marcello Henrique Nogueira-Barbosa, Prof. Dr. Paulo Mazzoncini de Azevedo Marques e a Natália Santana Chiari Correia pelo apoio ao longo deste projeto.

Obrigado Prof. Dr. Renato Tinós, grande professor e orientador. Agradeço por sua confiança, paciência, incansável dedicação e por esclarecer tantas dúvidas durante essa minha trajetória.

Agradeço à Universidade de São Paulo pela oportunidade d estudar nesta renomada instituição. Obrigado por proporcionar um ambiente saudável para todos os alunos, além de estimular a criatividade, a interação e a participação nas atividades acadêmicas. Sou grato a todo corpo docente, funcionários, à direção e administração dessa instituição.

Agradeço à FAPESP (processo 2019/01219-2) pelo financiamento do projeto de pesquisa.

"Há verdadeiramente duas coisas diferentes: saber e crer que se sabe. A ciência consiste em saber; em crer que se sabe reside a ignorância."

(Hipócrates)

Resumo

A *Fratura Vertebral por Compressão* (FVC) é uma fratura do corpo vertebral relacionada a forças compressivas, com colapso parcial do corpo vertebral. As FVCs podem ocorrer secundariamente ao trauma, mas as FVCs não traumáticas podem ser secundárias à fragilidade causada por osteoporose (FVCs benignas) ou tumores (FVCs malignas). No caso de FVCs não traumáticas, a investigação da etiologia é geralmente necessária, uma vez que o tratamento e o prognóstico são dependentes do tipo da FVC. Atualmente, tem havido grande interesse no uso de Redes Neurais Convolucionais (CNNs) para a classificação de imagens médicas, pois essas redes permitem a extração automática de características interessantes para a classificação em um determinado problema. No entanto, as CNNs geralmente exigem grandes bancos de dados que muitas vezes não estão disponíveis. Além disso, essas redes geralmente não usam informações adicionais que podem ser importantes para a classificação. Uma abordagem diferente é classificar a imagem com base em um grande número de características predefinidas, uma abordagem conhecida como radiômica. Neste trabalho, propomos um método híbrido de classificação de FVCs que utiliza características de três fontes distintas: i) camadas intermediárias de CNNs; ii) radiômica; iii) informações adicionais dos pacientes e histograma de imagens. No método híbrido proposto aqui, características externas extraídas das imagens são inseridas como entradas adicionais para a primeira camada densa de uma CNN. Um Algoritmo Genético (AG) foi empregado para i) selecionar um subconjunto de características visuais, radiômicas e clínicas relevantes para a classificação de FVCs; ii) selecionar hiper-parâmetros que definem a arquitetura do modelo híbrido proposto para classificação. Experimentos usando diferentes abordagens para as entradas indicam que combinar informações pode ser interessante para melhorar o desempenho do classificador.

Palavras-chave: Rede Neural Convolucional. Algoritmo Genético. Diagnóstico Auxiliado por Computador. Fratura de Coluna Vertebral por Compressão.

Abstract

Vertebral Compression Fracture (VCF) is a vertebral body fracture related to compressive forces, with vertebral body partial collapse. VCFs may occur secondary to trauma, but non- traumatic VCFs may be secondary to osteoporosis fragility (benign VCFs) or tumors (malignant VCFs). In the case of non-traumatic VCFs, the investigation of etiology is usually necessary, since treatment and prognosis are dependent on the type of VCF. Currently, there has been great interest in using Convolutional Neural Networks (CNNs) for the classification of medical images because these networks allow the automatic extraction of interesting features for the classification in a given problem. However, CNNs usually require large databases that are often not available. Besides, these networks generally do not use additional information that may be important for classification. A different approach is to classify the image based on a large number of predefined features, an approach known as radiomics. In this work, we propose a hybrid method for classifying VCFs that uses features from three different sources: i) intermediate layers of CNNs; ii) radiomics; iii) additional information from the patients and image histogram. In the hybrid method proposed here, external features extracted from the images are inserted as additional inputs to the first dense layer of a CNN. A Genetic Algorithm (GA) was used to i) select a subset of visual, radiomic and clinical characteristics relevant to the classification of FVCs; ii) select hyper parameters that define the architecture of the proposed hybrid model for classification. Experiments using different approaches for the inputs indicate that combining information can be interesting to improve the performance of the classifier.

Keywords: Convolutional Neural Network. Genetic Algorithm. Computer Aided Diagnosis. Vertebral Compression Fracture.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

AG	Algoritmo Genético
AM	Aprendizado de Máquina
CNN	Rede Neural Convolucional
FVC	Fratura vertebral por compressão
IRM	Imagem de ressonância magnética
RM	Ressonância Magnética
CAD	Diagnóstico auxiliado por computador
IA	Inteligência Artificial
RNAs	Redes Neurais Artificiais
ReLu	Unidade de retificação linear
MLP	Rede neural multicamadas
PACS	Picture Archiving and Communication System
DICOM	Digital Imaging and Communications in Medicine
TC	Tomografia Computadorizada
TIFF	Tagged Image File Format
ROC	Receiver Operating Characteristic
AUC	Area under the ROC curve
ACC	Acurácia
SENS	Sensibilidade
SPEC	Especificidade
V_p	Verdadeiro positivo
V_n	Verdadeiro negativo
F_p	Falso positivo
F_n	Falso negativo

Sumário

1

Introdução

Fratura vertebral por compressão (FVC) é uma fratura do corpo vertebral relacionada à compressão, representada em imagens médicas como colapso parcial do corpo vertebral. As FVCs podem ocorrer secundariamente ao trauma, mas as FVCs não traumáticas podem ser secundárias à fragilidade causada por osteoporose (FVCs benignas) ou tumores (FVCs malignas). No caso de FVCs não traumáticas, a investigação da etiologia é geralmente necessária, uma vez que o tratamento e o prognóstico dependem do tipo da FVC. O clínico e o radiologista não terão dúvidas sobre a etiologia de uma FVC de trauma de alta energia, dada a história recente do paciente [Tehranzadeh and Tao, 2004]. No entanto, quando o paciente desenvolve um colapso vertebral recente e doloroso sem trauma, podem ocorrer dificuldades no diagnóstico, uma vez que as queixas clínicas de pacientes com FVCs benignas e malignas podem ser semelhantes [Tehranzadeh and Tao, 2004]. Esse cenário clínico desafiador é comum principalmente na população idosa em que a osteoporose e o câncer são mais comuns [Taberner et al., 2007]. Além do que, múltiplas fraturas vertebrais podem ocorrer e o mesmo paciente pode ter diagnóstico simultâneo de osteoporose e câncer.

Imagens de Tomografia Computadorizada e Imagens de Ressonância Magnética (IRMs) são úteis na detecção e classificação de FVCs [Pereira et al., 2015, Taberner et al., 2007, Cuenod et al., 1996]. A Tomografia Computadorizada e as IRMs podem auxiliar no diagnóstico diferencial com base em anormalidades morfológicas e de intensidade de sinal detectadas especialmente nos corpos vertebrais, mas também a partir de elementos posteriores vertebrais e tecidos moles paravertebrais.

A distinção entre causas benignas e malignas do colapso vertebral é um problema clínico comum. O colapso osteoporótico no quadro agudo pode ser difícil de diferenciar das causas patológicas. Além disso, vários lesões podem ocorrer.

De acordo com Cuenod et al. [1996], Jung et al. [2003], Tehranzadeh and Tao [2004], os achados que sugerem FVC benigna quando a IRM é empregada são: (a) Imagens

ponderadas em T1 [1] demonstraram preservação da medula normal em pelo menos algumas áreas da corpo vertebral; (b) As imagens ponderadas em T1 com gadolínio e T2 [1] exibiram intensidade de sinal grosseiramente normal (isointensa com vértebras adjacentes normais); (c) Presença de linha de fratura; (d) Banda de baixa intensidade de sinal na imagem ponderada em T1 e T2; (e) Retropulsão de um fragmento ósseo posterior; (f) Múltiplas fraturas por compressão. Os achados que sugerem colapso vertebral maligno são: (a) Imagens ponderadas em T1 geralmente não exibiam intensidade residual normal do sinal da medula; (b) As imagens ponderadas em T1 e em T2, com gadolínio, demonstraram áreas de intensidade de sinal anormalmente aumentadas e, às vezes, irregulares; (c) Um córtex posterior convexo do corpo vertebral; (d) áreas difusas de baixa intensidade de sinal no corpo vertebral e pedículos em imagens ponderadas em T1; (e) Massa epidural, massa epidural encapsulada, massa paraespinhal focal e outras metástases da coluna vertebral.

A Figura 1 ilustra três IRM contendo exemplos de corpos vertebrais normais e com as fraturas estudadas.

Devido a diversos motivos, tais como escassez de profissionais experientes e aumento da facilidade em obter imagens médicas, existe um grande interesse na utilização de técnicas de Aprendizado de Máquina para auxílio ao diagnóstico médico [Azevedo-Marques, 2001, Shen et al., 2017].

Neste projeto, imagens de corpos vertebrais obtidas por ressonância magnética são classificadas utilizando-se métodos de Aprendizado de Máquina. O sistema de auxílio ao diagnóstico resultante classifica os corpos vertebrais em normais ou em FVCs benignas ou malignas. Poucos trabalhos da literatura investigaram o uso de Aprendizado de Máquina para o auxílio ao diagnóstico de FVCs utilizando-se IRM.

[1] A ressonância magnética produz imagens de cortes finos de tecidos utilizando um campo magnético e ondas de rádio. Através do controle das radiofrequências de pulso e das ondas de gradientes, programas de computador determinam como uma imagem é obtida (ponderada) e como os vários tecidos aparecem.

Por exemplo, gordura aparece brilhante (alta intensidade de sinal) nas imagens ponderadas em T1 e relativamente escura (baixa intensidade de sinal) nas imagens ponderadas em T2, água e líquidos aparecem relativamente escuros em imagens ponderadas em T1 e brilhantes nas imagens ponderadas em T2. As imagens ponderadas em T1 mostram de forma ideal a anatomia de partes moles e gordura (p. ex., para confirmar uma massa que contém gordura). Imagens ponderadas em T2 mostram, idealmente, líquidos e patologias (p. ex., tumores, inflamação, trauma). Na prática, as imagens ponderadas em T1 e T2 fornecem informações complementares, de forma que ambas são importantes para caracterizar a patologia.

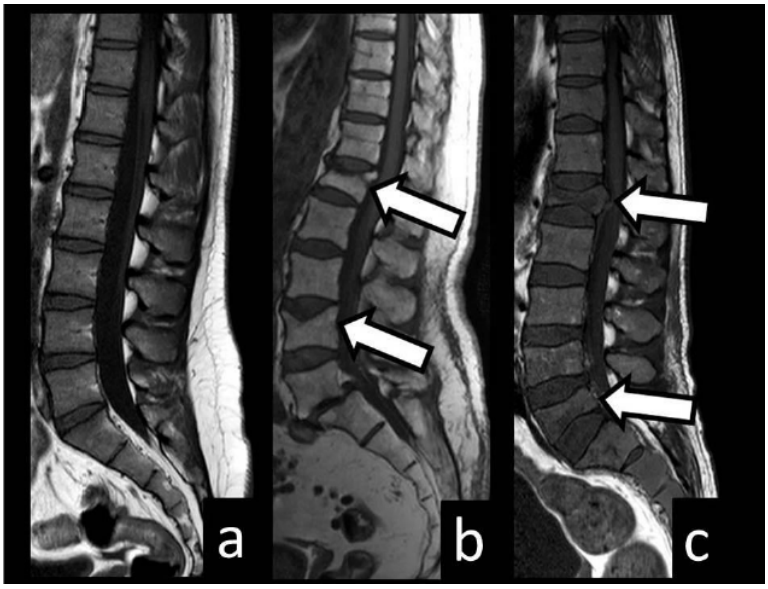


Figura 1 – Ilustração de três IRMs da coluna vertebral de diferentes pacientes em cortes sagitais medianos. Em (a) os corpos vertebrais estão íntegros não havendo presença de fraturas vertebrais. (b) apresenta duas FVCs benignas nos corpos vertebrais L1 e L4, onde o índice após "L" indica o número do corpo vertebral. (c) apresenta duas FVCs malignas nos corpos vertebrais L1 e L5.

Fonte: [Pereira, 2016]

1.1 Revisão bibliográfica

1.1.1 Diagnóstico auxiliado por computador

As imagens médicas são parte integrante do registro eletrônico de saúde de um paciente e são analisadas por radiologistas humanos, que são limitados pela velocidade, fadiga e experiência [Ker et al., 2017]. Leva anos e um grande custo financeiro para treinar um radiologista qualificado, e alguns sistemas de assistência médica terceirizam os relatórios de radiologia para países de baixo custo, como a Índia, por telerradiologia. Um diagnóstico atrasado ou errôneo pode causar danos ao paciente.[Ker et al., 2017].

Devido a estas e outras causas, como por exemplo as incertezas causadas por grandes variações na patologia e potencial fadiga de especialistas humanos, pesquisadores e médicos começaram a se beneficiar de intervenções assistidas por computador [Shen et al., 2017]. Embora, em comparação com os avanços nas tecnologias de imagens médicas, os avanços na análise de imagens médicas computacionais sejam tardios, a área vem melhorando recentemente com a ajuda de técnicas de aprendizado de máquina [Shen et al., 2017].

Diagnóstico auxiliado por computador (*Computer-aided Diagnosis system* - CAD) é definido como um diagnóstico realizado pelo especialista, utilizando o resultado de

análises quantitativas automatizadas como uma segunda opinião para a tomada de decisão [Azevedo-Marques, 2001, Seixas and Saade, 2005].

A finalidade do CAD é melhorar a acurácia e a consistência do diagnóstico utilizando a resposta do computador como referência. O auxílio por computador pode ser útil visto que o diagnóstico é baseado em avaliação subjetiva, estando sujeito a variações intrapessoais [Azevedo-Marques, 2001, Seixas and Saade, 2005].

Azevedo-Marques [2001] define e discute os conceitos básicos relacionados ao diagnóstico auxiliado por computador e apresenta uma revisão bibliográfica sobre o assunto, na qual apresenta as duas principais aplicações do CAD (auxílio na detecção de lesões e auxílio ao diagnóstico), que utilizam-se de técnicas provenientes de duas áreas do conhecimento: visão computacional e inteligência artificial. Também são abordadas medidas de desempenho que podem ser utilizadas na avaliação de sistemas de auxílio ao diagnóstico.

Atualmente, tem havido um grande interesse em utilizar Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNNs) para a análise e classificação de imagens médicas [Greenspan et al., 2016, Shen et al., 2017, Litjens et al., 2017]. A abordagem tradicional para a classificação de imagens médicas é a extração de características por meio de filtros pré-definidos, tal como foi feito no trabalho para a classificação de FVCs [Pereira, 2016]. A definição dos filtros mais adequados para uma determinada aplicação tem geralmente grande impacto no processo de classificação. A grande vantagem das CNNs é que estes algoritmos descobrem de forma automática as representações (características) interessantes para a classificação em um determinado problema [LeCun et al., 2015]. Isto ocorre devido ao uso de camadas de processamento de entradas anteriormente não presentes em Redes Neurais Artificiais tradicionais; as principais camadas de processamento de entradas nas CNNs são as camadas convolucionais e de *pooling*.

Entretanto, tais redes geralmente exigem grandes bases de dados que muitas vezes não são disponíveis em aplicações médicas. Além disso, geralmente, tais redes não utilizam informações adicionais que podem ser importantes para a classificação. No caso de imagens médicas, tais informações são importantes para o auxílio ao diagnóstico médico. Por exemplo, informações sobre a idade do paciente e histórico recente de traumas são importantes para a classificação das FVCs, como já discutido anteriormente.

Em análise de imagens médicas, cada diagnóstico médico é relacionado a uma amostra e os bancos de dados típicos são relativamente pequenos quando comparados com aqueles de outras áreas em que CNNs têm sido utilizadas. Este é o caso para o banco de dados que é utilizado neste trabalho (ver Seção 3.1).

Desta forma, várias pesquisas envolvendo CNNs na classificação de imagens médicas

têm utilizado a técnica de transferência de aprendizado [2]. A ideia principal da transferência de aprendizado é utilizar CNNs pré-treinadas, geralmente com grandes bancos de dados genéricos, para a classificação em um problema específico. Em [Litjens et al., 2017], duas estratégias são identificadas: i) o uso de CNNs pré-treinadas como extratores de características das imagens; ii) o ajuste de CNNs pré-treinadas para o problema específico.

A pesquisa por Sistemas CAD baseado em inteligência artificial (IA) vem se desenvolvendo rapidamente nos últimos anos. Recentemente, começaram a ser investigados sistemas CAD que utilizam redes neurais convolucionais (CNNs). Este modelo de rede neural artificial é amplamente utilizado em diversas aplicações relacionadas à análise de imagens, como para reconhecer rostos humanos no Facebook e na ferramenta de imagens do Google. Um sistema CAD utilizando CNN pode, muitas vezes, alcançar uma maior precisão diagnóstica do que o CAD convencional devido as suas características que permitem identificar relações na imagem [Komeda et al., 2017].

Shen et al. [2017] apresentaram uma revisão sobre a análise assistida por computador e como os avanços na aprendizagem de máquina, especialmente no que diz respeito à aprendizagem profunda, estão ajudando a identificar, classificar e quantificar padrões em imagens médicas. Devido ao sucessos, pesquisadores no campo da imagiologia médica computacional começaram a investigar o potencial da aprendizagem profunda em imagens médicas adquiridas com, por exemplo, tomografia computadorizada (TC), IRM, tomografia por emissão de pósitrons (PET) e raios-X. Os autores discutem as aplicações práticas da aprendizagem profunda no registro e localização de imagens, detecção de estruturas anatômicas e celulares, segmentação de tecidos e prognóstico e diagnóstico da doença auxiliado por computador.

Neste trabalho, combinamos informações extraídas da imagem por uma CNN com características radiômicas extraídas da imagem e informações extraídas do prontuário eletrônico paciente. A ideia de combinar informações de imagens com outras informações médicas não é nova. Dimitrovski et al. [2015] propuseram um método para melhorar a classificação de modalidade de imagem médica usando uma combinação de recursos visuais e textuais. Os bancos de dados de imagens médicas usados continha imagens de muitas modalidades diferentes, como Raio-X, TC, ultra-som, etc. Afim de melhorar o desempenho da recuperação das imagens, as características visuais e textuais foram combinadas, aumentando o desempenho preditivo dos classificadores.

---

[2] As redes neurais profundas demoraram dias para serem treinadas do zero, mesmo utilizando GPU no seu treinamento. No intuito de diminuir o tempo de treinamento de novos modelos entra o *Transfer Learning*. A transferência de aprendizado consiste em utilizar o conhecimento adquirido por redes neurais profundas previamente treinadas para o reconhecimento de grandes conjuntos multi-classe de dados ao invés de iniciar o treinamento do zero, com o objetivo de tornar o processo de aprendizagem muito mais rápido.

1.1.2 CNNs Aplicadas na Análise de Imagens Médicas

Li et al. [2014] propuseram utilizar CNNs para completar e integrar dados de neuroimagem de multimodalidade. Projetaram uma CNN tridimensional, composta por duas camadas convolucionais e uma camada totalmente conectada, que usa uma modalidade de dados volumétricos como entrada e outra modalidade de dados volumétricos como sua saída. Quando treinados com ambas as modalidades de dados, a rede conseguiu capturar a relação não linear entre as duas modalidades. Esses experimentos demonstraram que algumas informações contidas em imagens PET poderiam ser previstas e estimadas, usando dados de ressonância magnética de entrada.

Greenspan et al. [2016] apresentam uma revisão bibliográfica de 18 artigos em uma edição especial do IEEE-Transactions on Medical Imaging (IEEE-TMI) sobre aprendizagem profunda em imagens médicas. A revisão apresenta conquistas recentes de CNNs, desde detecção à categorização (por exemplo, detecção de lesões, segmentação de imagens, modelagem de formas, registro de imagens), bem como abertura de novos domínios de aplicação. Também estão incluídos vários trabalhos que se concentram na exploração das redes e fornecem informações sobre as arquiteturas a serem escolhidas para várias tarefas, parâmetros, conjuntos de treinamento e muito mais. A revisão também aborda a transferência de aprendizagem e ajuste fino que são componentes-chave no uso de CNNs profundas em aplicações de imagens médicas, na qual a base de dados normalmente é limitada.

Litjens et al. [2017] propuseram uma pesquisa bibliográfica envolvendo mais de 300 artigos, a maioria deles recentes, sobre uma ampla variedade de aplicações de aprendizagem profunda em análise de imagens médicas. Os autores analisaram os principais conceitos de aprendizagem profunda pertinentes à análise de imagens médicas. Foram abordados o uso de aprendizado profundo para classificação de imagens, detecção de objetos, segmentação e outras tarefas envolvendo vários campos de aplicações da área de imagens médicas.

1.1.3 Estudos em FVCs

Genant et al. [1993] propuseram um método de classificação baseado nas alturas da porção anterior, central e posterior dos corpos vertebrais. Corpos vertebrais normais são classificadas como grau 0; levemente deformados, com uma redução de aproximadamente 20-25% em qualquer altura, são classificadas com grau 1; moderadamente deformados, com redução de aproximadamente 25-40% em qualquer altura, foram classificadas como grau 2; e gravemente deformados, com redução de 40% de redução em qualquer altura (grau 3).

Além disso, uma valor 0.5 foi definido como limiar de deformação no corpo vertebral. O corpo vertebral foi considerado fraturado se recebeu a classificação de grau 1 ou maior, e foi considerado normal se graduou 0 ou 0.5.

A fim de auxiliar os radiologistas na interpretação de imagem e consequentemente permitir o diagnóstico precoce da osteoporose, Kasai et al. [2006] desenvolveram um método computadorizado para a detecção de fraturas vertebrais em radiografias laterais de tórax. O corpo vertebral fraturado era detectado por meio da comparação da altura vertebral medida com a altura vertebral esperada. A sensibilidade obtida foi de 95%, enquanto a acurácia foi de 70.9% - 76.6%.

Ribeiro et al. [2012] propuseram um método para extração e análise de corpos vertebrais para o diagnóstico de fraturas por compressão utilizando imagens de radiografias laterais da coluna lombar. Filtros de Gabor e uma rede neural artificial foram aplicados para extrair os platôs superior e inferior de cada corpo vertebral. Em seguida, as alturas das vértebras foram analisadas usando a classificação semiquantitativa proposta por Genant et al. [1993]. Com o CAD proposto, foram obtidas uma sensibilidade de 78% e uma especificidade de 95%.

Ghosh et al. [2011] propuseram um método automatizado para a detecção, localização e segmentação de FVCs em imagens de TC. Foi utilizado um ensemble de cinco classificadores, dentre eles SVM (Support VectorMachine), kNN (k Nearest Neighbor), LDA (Linear DiscriminantAnalysis), QDA (Quadratic Discriminant Analysis) e Naive Bayes. Para combinar a saída dos classificadores, foi utilizado um classificador de votação majoritária, obtendo uma precisão de 97.33%.

Al-Helo et al. [2013] propuseram um sistema CAD automatizado para o diagnóstico de fratura vertebral por compressão a partir de imagens de TC. Após realizar a localização e rotulagem dos corpos vertebrais, foi realizada a segmentação e então os corpos vertebrais foram diagnosticados. A extração de características foi baseada no conjunto de pontos resultantes da aplicação do Modelo Ativo de Forma (ASM) durante a etapa de segmentação. Foram utilizadas duas soluções de aprendizado de máquina para classificar os corpos vertebrais em fraturados e não fraturados: I) modelo supervisionado (Redes Neurais Artificiais - RNAs); II) modelo não supervisionado (K-Means) onde a acurácia foi calculada baseado no label de cada corpo vertebral em cada um dos dois clusters (cluster fratura e cluster não fratura). A precisão usando RNAs foi de 93.2%, em média, enquanto a precisão de diagnóstico usando o K-Means foi de 98%. O K-Means resultou em uma especificidade de 87.5% e sensibilidade acima de 99%.

Pereira [2016] propôs um método para detectar a presença de FVCs e de classificá-las como FVC maligna ou FVC benigna utilizando técnicas de processamento de imagens e aprendizado de máquina aplicadas em IRM. Os corpos vertebrais foram segmentados manualmente e na sequência foram extraídos os vetores de características. Após a aplicação

de métodos de seleção de atributos o autor utilizou diferentes classificadores tendo como entradas características de forma e textura selecionadas. Os classificadores utilizados foram: k-nearest-neighbor (K-NN), uma rede neural artificial com funão de base radial (RBF), Naive Bayes, J48 e Support Vector Machine (SVM). Os autores obtiveram bons resultados, alcançando uma acurácia de 95.8%, sensibilidade de 94.1% e especificiade de 97.8% utilizando o classificador Naive Bayes.

Lama [2018], em seu trabalho de conclusão de curso, utilizou a mesma base de dados de Pereira [2016] para estudar o uso de CNN para detectar a presença de fraturas vertebrais por compressão em IRM e classificá-las como malignas ou benignas. Os corpos vertebrais foram segmentados manualmente e utilizados para treinar a CNN. Ou seja, utilizou como entrada apenas as IRMs, sem que atributos de radiômica (ex.: atributos de textura e forma) e clínicos fossem investigados. A CNN obteve bons resultados, diferenciando corretamente imagens sem (normal) ou com FVCs, porém apresentou dificuldades para diferenciar FVCs benigna e FVCs maligna.

Raineri [2018] em seu trabalho de conclusão de curso utilizou a mesma base de dados de [Pereira, 2016] e [Lama, 2018] para estudar métodos de redução de dimensionalidade visando ajudar na diferenciação entre corpos vertebrais sem fratura e corpos vertebrais com FVCs benignas e malignas em IRM. Os atributos passíveis de seleção eram características (intensidade de cinza, forma e textura) extraídas de imagens segmentadas. Para a redução do número de atributos foi utilizados o método *wrapper* com dois métodos de busca: algoritmos genéticos e *sequential forward selection*. A autora conclui que a utilização de técnicas de seleção de atributos pode ser um importante recurso para o aprimoramento de análises de imagens médicas, sendo importante para o auxílio ao diagnóstico e para melhor compreensão de problemas da área médica.

1.2 Objetivos

Nas seções anteriores, foram citadas algumas observações relevantes: i) a importância de se utilizar um sistema de auxílio ao diagnóstico de FVCs; ii) a importância do uso de CNNs para a descoberta de representações (características visuais) na classificação de imagens médicas; iii) as dificuldades advindas do uso de bases pequenas para o treinamento de CNNs em aplicações médicas específicas; iv) o não uso, por parte de CNNs, de características radiômicas e informações do paciente (características clínicas) relevantes para a classificação das imagens. Dadas estas observações, podemos definir a hipótese investigada neste mestrado:

HIPÓTESE: *"um classificador eficiente que utiliza como entradas características visuais, radiômicas e clínicas relevantes deverá ter, em geral, desempenho superior ao das*

técnicas computacionais atualmente utilizadas para a classificação de FVCs".

As caraterísticas relevantes citadas na hipótese são um subconjunto do conjunto de características visuais, radiômicas e clínicas disponíveis. A seleção de características é importante para reduzir a dimensionalidade dos dados e para melhorar o desempenho de classificadores. Algoritmos Evolutivos têm recebido grande atenção de pesquisadores que investigam técnicas de seleção de atributos [Xue et al., 2016]. A partir da hipótese, podemos formular o objetivo principal deste projeto:

OBJETIVO PRINCIPAL: *"Investigar o uso de Algoritmos Genéticos (AG) para selecionar hiper-parâmetros e características relevantes em uma CNN na qual informações adicionais das imagens e dos pacientes são inseridas na primeira camada densa. O AG é empregado para i) selecionar um subconjunto de características visuais, radiômicas e clínicas relevantes para a classificação de FVCs; ii) selecionar um subconjunto de hiper-parâmetros que definem a arquitetura do modelo híbrido proposto para classificação. As características visuais são provenientes do processamento das camadas intermediárias da CNN. As informações adicionais (radiômicas e clínicas) são inseridas na primeira camada densa da CNN em conjunto com as informações provenientes das camadas intermediárias."*

Vale ressaltar que a análise dos atributos selecionados pode trazer informações relevantes do ponto de vista médico. Assim, um objetivo secundário deste trabalho é analisar os atributos selecionados, em busca de informações relevantes que possam auxiliar ou ajudar a entender a classificação de FVCs.

1.3 Organização do trabalho

O trabalho está organizado em capítulos, devidamente referenciados no Sumário. Uma introdução sobre Rede Neural Convolucional, Algoritmo Genético e sobre a biblioteca *pyRadiomics* são encontrados na Seção 2. A Metodologia proposta é apresentada na Seção 3. Os resultados estão na Seção 4, seguido pela Discussão na Seção 5 e Conclusão na Seção 6.

2

Referencial teórico

2.1 Rede Neural Convolucional (CNN)

Atualmente, as CNNs são uma das técnicas mais populares nas tarefas de reconhecimento de imagens. CNNs são as técnicas mais comuns utilizadas na abordagem de Aprendizado Profundo (*Deep Learning*). Técnicas de Aprendizado Profundo são exemplos de métodos de aprendizado de representação. Estes são métodos que permitem que uma máquina, ao ser alimentada com dados brutos, descubra automaticamente as melhores representações para a classificação e análise destes dados [LeCun et al., 2015].

Métodos de Aprendizado Profundo são: i) métodos de aprendizado de representação com múltiplos níveis de representação; ii) compostos por módulos simples não-lineares, que transformam a representação dos dados sucessivamente em níveis ligeiramente mais abstratos. Para tarefas de classificação, camadas superiores de representação amplificam os aspectos das entradas que são importantes para a discriminação e suprimem variações que são irrelevantes.

O aspecto chave do Aprendizado Profundo é que as características interessantes para a classificação e as transformações de representação não são diretamente projetadas por especialistas. Elas são aprendidas a partir da inferência utilizando dados brutos, criando-se assim um procedimento de aprendizado de propósito geral.

CNNs são modelos de Aprendizado Profundo baseados em Redes Neurais Artificiais (RNAs). As RNAs têm sido aplicadas na solução de uma infinidade de problemas [Haykin, 2003]. O adjetivo "neural" é usado porque que RNAs são compostas por neurônios artificiais, que são módulos descritos por funções matemáticas simples. Os módulos são dispostos em camadas de processamento, sendo conectados por números representando pesos sinápticos. Com a composição de funções não-lineares simples, funções complexas podem ser estimadas. Geralmente, o aprendizado se dá pela otimização dos pesos sinápticos, dado um conjunto de treinamento para uma determinada tarefa.

A CNN é um tipo específico de rede neural que usa variações de RNAs tradicionais, principalmente no que diz respeito ao processamento dos dados de entrada nas camadas iniciais da rede. As CNNs foram desenvolvidas tomando como base o córtex visual de animais, que é composto por milhões de agrupamentos celulares complexos, sensíveis a pequenas sub-regiões do campo visual, chamadas de campos receptivos. Diferentemente das camadas totalmente conectadas utilizadas em RNAs tradicionais, as camadas convolucionais das CNNs processam as entradas considerando pequenos campos receptivos. Em seguida, são geralmente utilizadas camadas que incluem operações conhecidas como *pooling*, responsáveis por reduzir a dimensionalidade das representações. As camadas finais das redes CNN são camadas tradicionais utilizadas em RNAs, i.e., camadas totalmente conectadas com funções de ativação não-lineares.

As CNNs usam relativamente pouco pré-processamento em comparação com outros algoritmos de classificação de imagem. Isso significa que a rede aprende os filtros que, nos algoritmos tradicionais, são manipulados de forma manual. Essa independência do conhecimento prévio e do esforço humano no design de recursos é uma grande vantagem [Liu et al., 2015].

A Figura 2 mostra um exemplo de CNN utilizada na classificação de imagens. Uma imagem geralmente vem na forma de uma matriz de valores de pixel. Tipicamente em uma CNN, as características descobertas na primeira camada indicam a presença ou ausência de bordas em orientações e locais específicos da imagem. As camadas seguintes normalmente detectam pequenos padrões particulares de bordas, independentemente de pequenas variações nas posições e orientações destas. As camadas posteriores podem então combinar pequenos padrões que resultam em padrões maiores que correspondem a partes de objetos familiares. Camadas subseqüentes detectariam objetos por meio da combinação de padrões obtidos nas camadas anteriores.

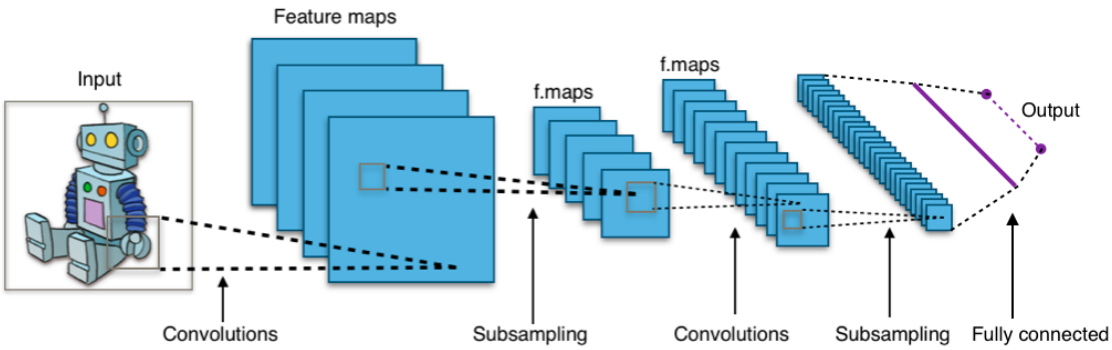


Figura 2 – Arquitetura padrão de uma rede neural convolucional (CNN)

Fonte: https://en.wikipedia.org/wiki/Convolutional_neural_network/media/File:Typical$_c nn.png$

2.2 Algoritmo Genético

Os Algoritmos Genéticos (AGs) são métodos adaptativos inspirados nos processos genéticos de organismos biológicos e na teoria da evolução por seleção natural. AGs são frequentemente utilizados para resolver problemas de busca e otimização [Mitchell, 1996].

No AG padrão, um conjunto de indivíduos (ou cromossomos) representando soluções do problema é sujeito a operadores de seleção e transformação inspirados em mecanismos encontrados na evolução por seleção natural e na genética. A solução x_i (também chamada de indivíduo ou cromossomo), para i=1,...,N, sendo N o tamanho da população, é avaliada através de uma função de custo, ou *fitness*, f (x_i).

Os operadores de transformação mais utilizados são o *crossover* e a mutação. No primeiro, dois indivíduos da população corrente escolhidos por meio do operador de seleção têm algumas das variáveis de decisão trocadas. Um exemplo de operador de seleção é o método da roleta, no qual a probabilidade de um indivíduo ser selecionado é proporcional ao seu *fitness* relativo (i.e., ao *fitness* normalizado pela soma de *fitness* dos indivíduos da população atual).

A probabilidade de se aplicar *crossover* é definida por uma taxa p_c , chamada de taxa de *crossover*. Na mutação, indivíduos têm alguns de seus elementos alterados por meio de uma regra pré-definida. Por exemplo, quando ocorre mutação no i-ésimo elemento do cromossomo para o caso binário, este elemento é negado. O número de variáveis de decisão alteradas por mutação é definido por uma taxa p_m , chamada de taxa de mutação. O pseudo-código simplificado do AG padrão é apresentado no Algoritmo 1.

Algoritmo 1: Pseudo-código do AG padrão

```
1 início
2     inicialize a população
3     avalie a população inicial
4     enquanto critério de convergência não for satisfeito faça
5         selecione indivíduos para a nova população
6         aplique mutação e cruzamento nos indivíduos selecionados
7         avalie os indivíduos da nova população
8     fim
9 fim
```

O uso de Algoritmos Evolutivos para seleção de características na abordagem *wrapper* tem atraído bastante atenção [Xue et al., 2016]. Na abordagem *wrapper*, o problema de seleção de características é visto como um problema de otimização, na qual o objetivo é encontrar um subconjunto de características ótimo para um determinado classificador, dada uma base de dados. Nestes problemas, os indivíduos do AG representam

combinações de características para o classificador utilizado.

2.3 *pyRadiomics*

A biblioteca *pyRadiomics* é uma biblioteca de código aberto escrita em Python para extração de características radiométricas de imagens médicas [Van Griethuysen et al., 2017].

Com a *pyRadiomics* é possível extrair das imagens atributos de estatísticas de primeira ordem, que estão relacionadas à distribuição de níveis de cinza na imagem. Alguns exemplos são: média, desvio-padrão e entropia. São oferecidos também recursos para a extração de características de forma (e.g., volume, área) e textura (e.g., correlação, contraste), os quais estão distribuídos da seguinte forma:

- First Order Statistics: 19 atributos
- Shape-based (3D): 14 atributos
- Gray Level Co-occurence Matrix: 23 atributos
- Gray Level Size Zone Matrix: 16 atributos
- Gray Level Run Length Matrix: 16 atributos
- Neighbouring Gray Tone Difference Matrix: 4 atributos
- Gray Level Dependence Matrix: 14 atributos

3

Metodologia

No método proposto neste trabalho, uma CNN classifica FVCs usando imagens segmentadas de corpos vertebrais como entradas e atributos externos são inseridos como entradas adicionais para a primeira camada densa. A seguir, é apresentado o método investigado, bem como o conjunto de dados. Em seguida, a forma de análise de resultados é discutida.

3.1 Base de Dados

O conjunto de imagens que foi utilizado neste mestrado é composto por IRMs da coluna vertebral de 61 pacientes diagnosticados com FVC, que passaram por tratamento no HCFMRP/USP. As imagens estão no plano sagital mediano (corte central) ponderadas na sequência de contraste T1. Esse conjunto de imagens foi extraído do PACS do HCFMRP/USP em formato Digital Imaging and Communications in Medicine (DICOM) com 16-bits/pixel (65536 níveis de cinza). Todo o conjunto foi posteriormente convertido para o formato Tagged Image File Format (TIFF) sem compressão e convertido para a escala de 8-bits/pixel (256 níveis de cinza).

A imagem resultante de um exame de ressonância magnética é uma imagem 3D. O radiologista responsável pela base de dados sugeriu que se utilizasse apenas o slicer central da IRM, para a classificação das FVCs. Assim, apenas o slice central dos 61 pacientes foi selecionado para ser utilizado neste trabalho, resultando em uma imagem 2D.

As imagens dos 61 pacientes foram segmentadas manualmente em janelas retangulares contendo um corpo vertebral (amostra) cada. Um total de 189 janelas (amostras) foi obtido pelo processo de segmentação. Um exemplo do processo de segmentação é mostrado na Figura 3. O radiologista responsável pela base de dados rotulou cada um dos corpos vertebrais considerando três classes: FVC benigna (54 exemplos), FVC maligna (46 exemplos) e normal (89 exemplos). Os corpos vertebrais considerados normais foram

extraídos da coluna lombar de pacientes com FVCs benignas[1]. Exemplos das imagens podem ser vistos na Figura 1. A Tabela 1 mostra a distribuição de todas as amostras da base de dados de acordo com os corpos vertebrais [2].

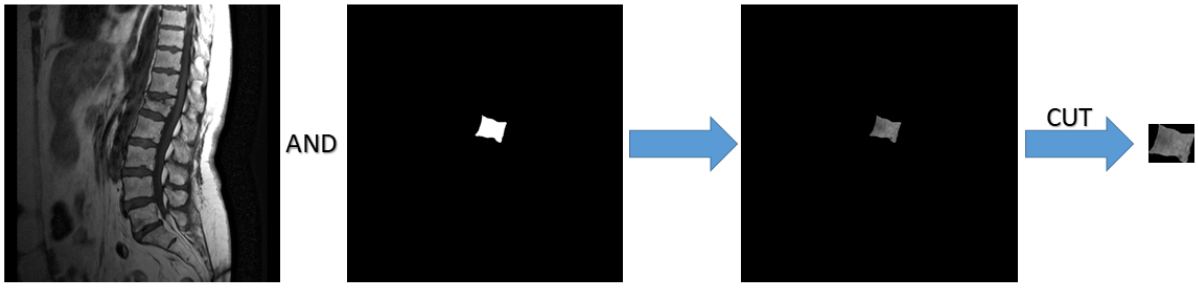


Figura 3 – Etapas do processo de segmentação aplicado a ressonâncias magnéticas. A segunda imagem mostra uma máscara aplicada à imagem original (primeira imagem). A imagem resultante (terceira imagem) é então cortada para formar um exemplo do conjunto de dados (última imagem).

Fonte: [Lama, 2018]

Tabela 1 – Base de dados do trabalho quantificando os corpos vertebrais lombares de acordo com a classificação padrão-ouro.

Classificação	**L1**	**L2**	**L3**	**L4**	**L5**	**Total**
FVC maligna	8	9	10	11	8	46
FVC benigna	21	10	9	9	5	54
Normal	9	18	19	20	23	89
Total	38	37	38	40	36	189

O conjunto de dados, composto por 61 IRMs de diferentes pacientes, foi dividido em conjuntos de treinamento e teste. O conjunto de treinamento é composto por imagens de 55 pacientes, totalizando 169 corpos vertebrais. O conjunto de testes consiste em imagens de 6 pacientes, totalizando 20 corpos vertebrais.As Tabelas 2 e 3 mostram a distribuição dos dados de treinamento e teste de acordo com os corpos vertebrais.

No conjunto de treinamento, a validação cruzada de 10-folds foi utilizada para ajuste dos hiper-parâmetros e avaliação do modelo. Na validação cruzada, 9 subconjuntos (folds) de dados são utilizados para treinamento e um para avaliação da qualidade do classificador (ou seja, este conjunto não é utilizado no treinamento). O classificador é treinado 10 vezes, mudando os subconjuntos para treinamento e avaliação. Apesar de não conter os mesmos corpos vertebrais (exemplos) nos dois subconjuntos, corpos vertebrais de um mesmo paciente podem ser utilizados para treinamento e avaliação. Assim, utilizou-se

[1] O conjunto de dados é o mesmo utilizado em trabalhos anteriores [Raineri, 2018], [Lama, 2018] e com pequenas mudanças em relação ao conjunto utilizado em [Pereira, 2016]. O conjunto tem sido também utilizado, com diversos fins, no Doutoramento de Natália Santana Chiari Correia. A utilização deste conjunto de dados para fins de pesquisa foi aprovada pelo Comitê de Ética em Pesquisa do Hospital da Clinicas da Faculdade de Medicina de Ribeirão Preto da USP (HCFMRP/USP).

[2] O índice após "L" indica o número do corpo vertebral

uma abordagem no qual exemplos de um mesmo paciente são separados para uma etapa posterior de teste do modelo. O conjunto de testes é utilizado para simular um cenário real encontrado em um hospital, onde novos pacientes chegariam para serem atendidos e seriam tratados por um médico utilizando o sistema CAD no apoio ao diagnóstico.

Tabela 2 – Conjunto de treinamento quantificando os corpos vertebrais lombares de acordo com a classificação padrão-ouro.

Classificação	**L1**	**L2**	**L3**	**L4**	**L5**	**Total**
FVC maligna	7	7	9	10	8	41
FVC benigna	18	10	8	8	4	48
Normal	9	15	17	18	21	80
Total	34	32	34	36	33	169

Tabela 3 – Conjunto de teste quantificando os corpos vertebrais lombares de acordo com a classificação padrão-ouro.

Classificação	**L1**	**L2**	**L3**	**L4**	**L5**	**Total**
FVC maligna	1	2	1	1	0	5
FVC benigna	3	0	1	1	1	6
Normal	0	3	2	2	2	9
Total	4	5	4	4	3	20

3.1.1 Dados clínicos adicionais e atributos extraídos do histograma da imagem

As informações clínicas e demográficas utilizadas neste trabalho foram extraídas do prontuário eletrônico dos pacientes. As informações extraídas foram:

1. Idade do paciente;
2. Sexo do paciente;
3. Número correspondente ao corpo vertebral (L1 a L5);

O histograma de uma imagem é um conjunto de números que indica a quantidade de pixels em cada um dos níveis de cinza da imagem. Os histogramas correspondentes ao slice central das IRM de cada pacientes podem ser divergentes, o que provavelmente dificultaria o aprendizado por parte da rede neural. Por isso, foram extraídas também as seguintes informações do histograma resultante da imagem do slice central das IRM originais:

1. Valor de nível de cinza máximo da imagem;
2. Média dos valores de nível de cinza da imagem;

3. Valor de nível de cinza mínimo da imagem;

4. Desvio padrão dos valores de nível de cinza da imagem;

No Apêndice A encontra-se uma análise dos dados clínicos disponíveis no prontuário eletrônico dos pacientes e aprovados para uso em pesquisas.

3.1.2 Vetor de características radiômicas

As informações contidas dentro das imagens radiológicas já vem sendo estudadas, além da representação gráfica usual, por muitos trabalhos na literatura médica, dentro do novo campo da ciência médica chamado Radiômica [Santos et al., 2019]. A Radiômica significa extração de dados mensuráveis de imagens radiológicas e sua integração em modelos preditivos, ou seja, usar informações contidas nas imagens para fazer uma análise ampla e complexa para definir o tratamento mais apropriado para cada paciente e prever a evolução da doença. A ideia básica da área de radiômica é coletar informação de um gigantesco banco de dados para treinamento de modelos preditivos, por exemplo, que possibilitem a análise de imagens de tumores, evitando, assim, procedimentos invasivos desnecessários.

Neste mestrado, as características radiômicas que compõem o vetor de características foram extraídas usando a biblioteca *pyRadiomics*. Como a biblioteca funciona apenas para imagens 3D, para a extração do vetor de características foram utilizadas as IRM originais (imagens 3D) ao invés de utilizar apenas o slicer central (imagens 2D). Mais informações sobre a biblioteca *pyRadiomics* estão disponíveis na Seção 2.3.

3.1.3 Aumento de dados

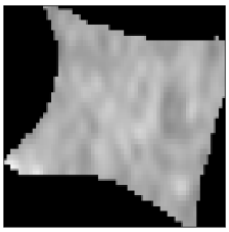
Redes neurais profundas possuem uma grande quantidade de parâmetros para serem otimizados, fazendo com que seja necessário dispor de uma grande quantidade de dados de para treinamento. Como a quantidade de imagens disponíveis neste trabalho é pequena, foi testado um método de *data augmentation*, i.e. geração de uma base de treinamento aumentada.

Para gerar a base de dados aumentada, foi utilizada a biblioteca *ImageDataGenerator* [3] do Keras. Foram aplicadas operações sobre as imagens originais, dentre elas translação, rotação e espelhamento. Achatamento, alongamento e alteração brilho não foram utilizados pois alteram características relevantes que são observadas no diagnóstico da FVC.

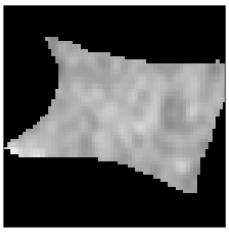
[3] `https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator`

3.1.4 Leitura das imagens

Para realizar o treinamento de uma CNN, todas as imagens da base de dados devem possuir a mesma dimensão. Para isso, foi necessário desenvolver um método customizado para deixar todas as imagens com a mesma dimensão de forma a preservar a proporção (altura x largura) da imagem original. O método desenvolvido consiste em ler a imagem no tamanho original, redimensionar a imagem de forma a ficar o mais próximo possível da dimensão esperada pela CNN mantendo a proporção original (a imagem foi redimensionada proporcionalmente em relação a altura e a largura até alguma delas ser igual a dimensão esperada pela rede), e caso necessário, a imagem foi complementada com fundo preto. A Figura 4 ilustra um exemplo de uma imagem redimensionada sem utilizar e utilizando o método desenvolvido.



(a) Exemplo de uma imagem redimensionada sem utilizar o método desenvolvido.



(b) Exemplo de uma imagem redimensionada utilizando o método desenvolvido.

Figura 4 – Comparação de imagens sendo redimensionadas pelo método padrão x método desenvolvido.

Como é possível observar, o método padrão ao redimensionar a imagem altera o formato do corpo vertebral. Como a forma é uma característica importante na classificação do corpo vertebral, o método customizado foi desenvolvido.

3.2 Sistema de Apoio ao Diagnóstico de FVCs baseado em Algoritmos Genéticos e Redes Neurais Artificiais

3.2.1 Sistema Híbrido

O sistema híbrido proposto neste trabalho utiliza um CNN para classificação. Os atributos extraídos das imagens (radiômica) e histograma e informações do paciente são inseridos

como entradas adicionais para a primeira camada densa da CNN. Assim, três tipos de informações são utilizadas pela CNN. As imagens brutas são fornecidas como entradas da rede convolucional, como nas CNNs tradicionais. As entradas da primeira camada densa da CNN são fornecidas por um vetor composto pelos atributos radiômicos, informações do paciente e dos histogramas e saídas das camadas intermediárias. Na prática, uma CNN é construída incluindo as entradas dos atributos radiômicos e informações do paciente e dos histogramas nas camadas densas (ver Figura 5(c)). Desta forma, as características encontrados pelas camadas convolucional e max-pooling (camadas iniciais do CNN) são usados como entradas, juntamente com atributos externos, nas camadas densas. A CNN é treinada por Backpropagation.

As imagens (brutas) e as informações adicionais usados aqui foram descritas nas seções anteriores. Como o conjunto de dados é pequeno, a ideia é usar a CNN para descobrir novos recursos que combinados com as as informações adicionais melhoram a classificação dos FVCs.

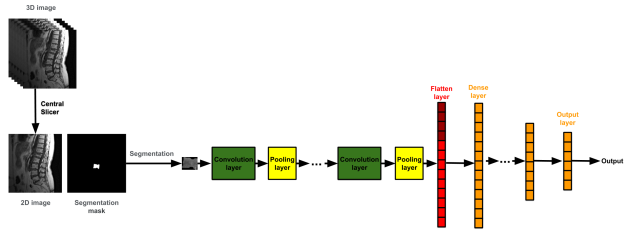
A fim de testar nossa hipótese de que adicionar informações adicionais a CNN pode melhorar a classificação, o modelo híbrido proposto (Figura 5(c)) é comparado a duas outras abordagens: i) um CNN usando apenas imagens (brutas) (Figura 5(a)). Esta é a mesma abordagem usada em [Lama, 2018]; ii) um MLP usando apenas atributos radiômicos e histograma e informações do paciente (Figura 5(b)). A abordagem para o MLP é semelhante àquela usada em [Pereira, 2016, Pereira et al., 2015], mas adicionando histograma e informações do paciente às entradas do classificador.

Todos os códigos usados neste trabalho foram desenvolvidos em Python usando a API do Keras (versão 2.3.1) com o back-end TensorFlow (versão 2.1.0). Experimentos preliminares foram realizados para encontrar os melhores modelos para cada abordagem. Diferentes arquiteturas e parâmetros (por exemplo, número de camadas, número de neurônios e parâmetros CNN) foram testados usando o conjunto de validação cruzada. Experimentos com aumento de dados e CNNs pré-treinadas também foram feitos.

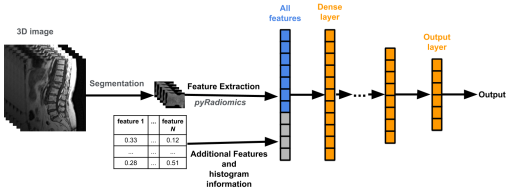
3.2.2 CNN pré-treinada

Com o objetivo de obter a saída das camadas intermediárias de uma CNN, torna-se necessário um modelo pré-treinado com uma arquitetura passível de ser alterada, ou seja, remover as camadas totalmente conectadas. Após a revisão bibliográfica, o autor desconhece um modelo que tenha sido pré-treinado utilizando-se imagens médicas e que tenha uma arquitetura passível de ser alterada. Por esse motivo, foram utilizadas os modelos pré-treinados do Keras [4]. Dentre todos os modelos disponíveis, foram investigadas neste trabalho 3 modelos: *Xception*, *InceptionV3* e *VGG16*. Estes modelos foram escolhidos

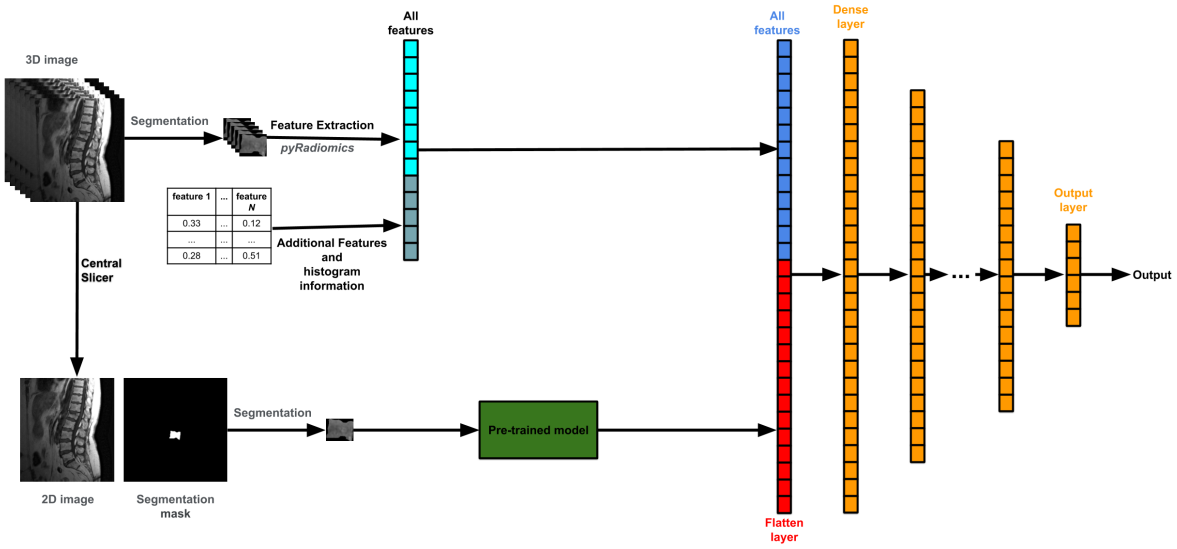
[4] `https://keras.io/applications/`



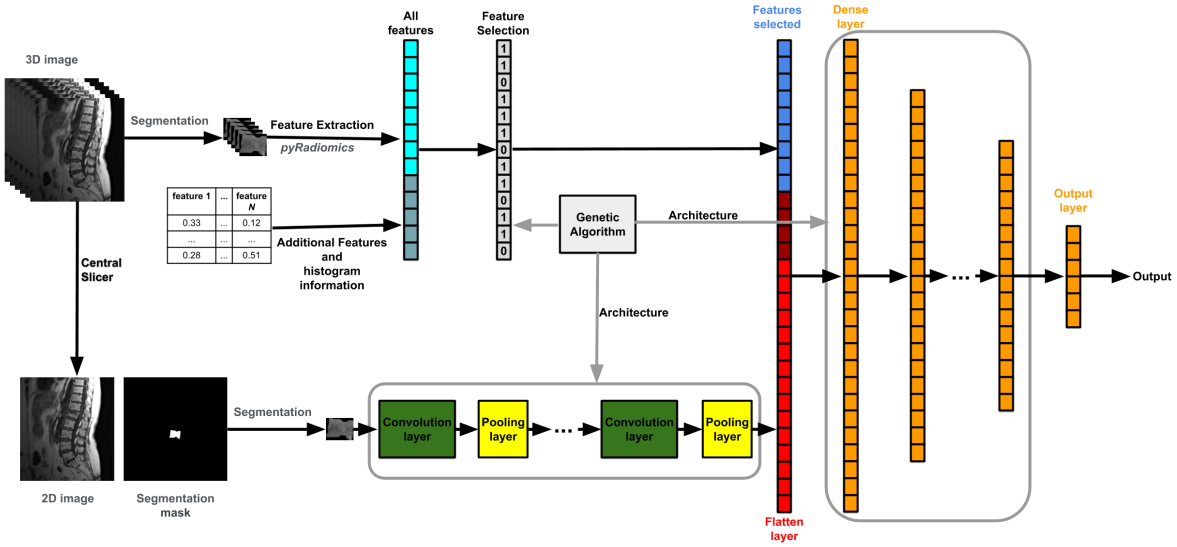
(a) Arquitetura da CNN utilizada para classificação de FVCs utilizando como entrada o slicer central (imagem 2D) da IRM (imagem 3D).



(b) Arquitetura da MLP utilizada para classificação de FVCs utilizando como entrada atributo de radiômica (extraídos da imagem 3D resultando do exame de RM), dados clínicos do paciente e informações extraídas do histograma do slicer central.



(c) Arquitetura do modelo híbrido proposto, utilizando todas as fontes de informação. O modelo híbrido é treinado de forma única, utilizando uma CNN pré-treinada.



(d) Arquitetura do modelo híbrido proposto, utilizando todas as fontes de informação. O modelo híbrido é treinado de forma única, sem utilizar modelos pre-treinados em partes. O AG, representado pela cor cinza, é responsável pela seleção de atributos e pela otimização dos hiper-parâmetros tanto da CNN utilizada para extração de características, quanto do MLP utilizado para realizar a classificação combinando todas as fontes de dados.

Figura 5 – Arquitetura dos modelos utilizados no trabalho. Mais informações sobre a arquitetura dos modelos estão disponíveis no Apêndice C

por terem obtido os melhores desempenhos nos 5 testes de desempenho do *ImageNet* [5] e por serem modelos possíveis de serem executados com o poder computacional disponível para o projeto (ver Apêndice B). Dentre os modelos pré-treinados testados, a *VGG16* foi o modelo que obteve o melhor resultado nos experimentos preliminares realizado com o conjunto de treinamento (utilizando cross validation), e por isso foi o modelo pré-treinado utilizado.

3.2.3 Impacto do Aumento de Dados

Nos experimentos preliminares usando o conjunto de validação cruzada, a CNN treinada utilizando aumento de dados obteve uma acurácia de 67.4% contra 67.8% obtido pela CNN treinada sem aumento de dados (resultados apresentados na próxima sessão). Além disso, o tempo médio para o treinamento da CNN sem aumento de dados foi de 6 min, enquanto para a CNN utilizando aumento de dados foi de 13 min. Como não houve melhora no desempenho da CNN utilizando aumento de dados e o tempo de treinamento é maior, não foi utilizado aumento de dados para o modelo híbrido.

3.3 Algoritmo Genético

O AG foi empregado para selecionar subconjuntos de características a partir do conjunto total de características disponíveis dos três tipos e para encontrar os melhores hiper-parâmetros para construção do modelo. Em um AG utilizado para seleção de atributos, cada indivíduo representa um subconjunto de características, sendo que cada elemento do cromossomo indica a presença ou não de cada característica disponível. Já em um AG aplicado na otimização de hiper-parâmetros, cada indivíduo representa um conjunto de hiper-parâmetros que determinam a arquitetura do modelo. Neste trabalho, cada indivíduo representa um subconjunto de características e um conjunto de hiper-parâmetros que determina a arquitetura do modelo híbrido.

Cada indivíduo do AG é avaliado treinando o modelo correspondente utilizando a base de treinamento. A divisão treino(80%)/teste(20%) é utilizada nesta etapa. O resultado da acurácia balanceada nos exemplos de teste é utilizado para a avaliação do indivíduo. No final da execução do AG, a validação cruzada do tipo *10-folds* é utilizada para avaliar o melhor indivíduo encontrado pelo AG. Ou seja, a base de treinamento é dividida em 10 subconjuntos, sendo o MLP treinado 10 vezes considerando diferentes subconjuntos para treinamento e teste. A Figura 6 mostra o fluxograma simplificado do AG utilizado.

[5] O projeto *ImageNet* é um grande banco de dados de imagens, possuindo mais de 14 milhões de imagens de mais de 20.000 categorias.

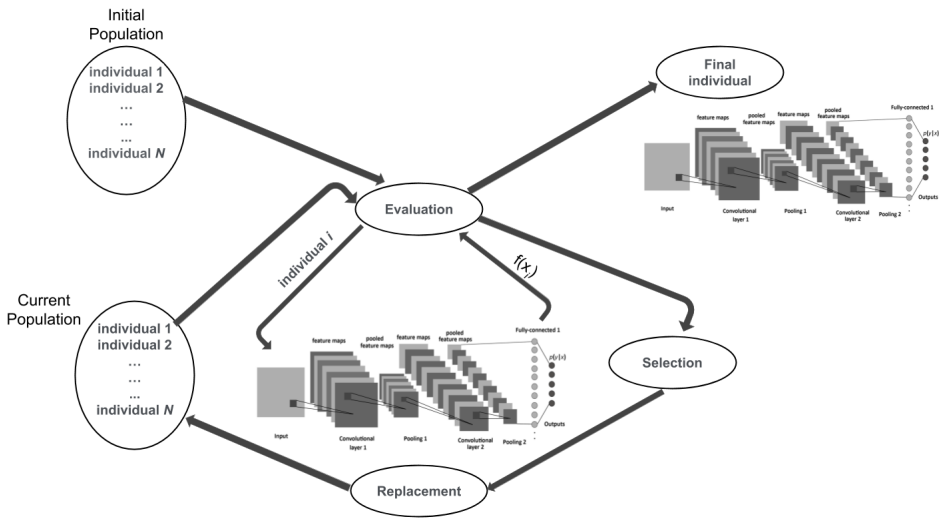


Figura 6 – Fluxograma simplificado do AG utilizado para seleção de características e seleção dos hiper-parâmetros do modelo.

Cada indivíduo gerado pelo AG é responsável por determinar os hiper-parâmetros do modelo híbrido proposto e quais atributos (características adicionais) são selecionados. Um exemplo da codificação utilizada é apresentado na Figura 7. As arquiteturas codificadas pelo AG podem ter até 5 camadas de convolução e até 5 camadas densas. A Tabela 4 apresenta os parâmetros que podem ser selecionados pelo AG em cada posição da codificação, sendo eles do tipo booleano, inteiro ou string. Assim, além dos 3 modelos para classificação anteriores (mostrados nas Figuras 5(a)-5(c)) um quarto foi testado (Figura 5(d))

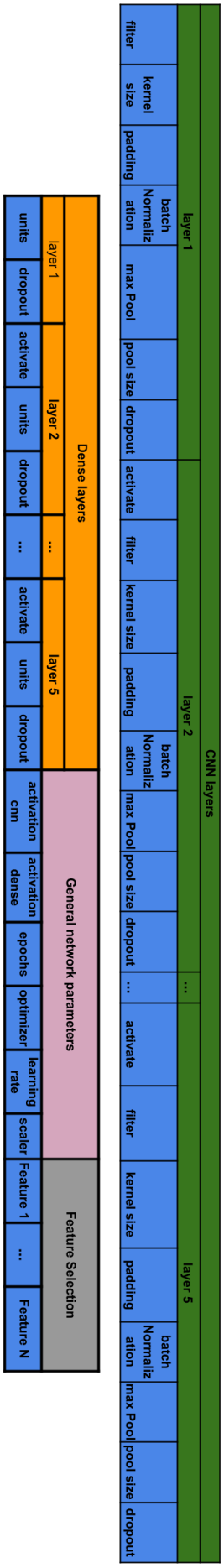


Figura 7 – Exemplo de codificação representada por um indivíduo (indivíduo é representado pela cor azul) no AG.

Tabela 4 – Codificação do AG. Lista de parâmetros que poderiam ser utilizados em cada posição da codificação representada por cada indivíduo.

Parâmetro	Valores	Tipo	Descrição
activation_cnn	['relu', 'tanh']	string	Função de ativação a ser utilizada nas camadas convolucionais.
activation_dense	['relu', 'tanh']	string	Função de ativação a ser utilizada nas camadas densas.
dropout	[0.0, 0.1, 0.2, 0.3, 0.4]	real	Taxa de *dropout* a ser utilizado na camada.
epochs	[30, 45, 60, 90, 120, 150, 200]	inteiro	Número de épocas que o modelo deve ser treinado.
filters	[4, 5, 6, 7, 8, 9]	inteiro	Inteiro que define a dimensionalidade do espaço de saída (ou seja, o número de filtros de saída na convolução). O número de filtros é definido pela função: filtros = 2^i, onde i é o valor fornecido.
kernel_size	[2, 3, 4, 5, 6]	inteiro	Inteiro, especificando a altura e largura da janela de convolução 2D. O mesmo valor é usado para ambas as dimensões.
learning_rate	[0.0001, 0.001, 0.01, 0.05, 0.1]	real	*Learning rate* utilizada no treinamento do modelo.
optimizer	['adam', 'sgd']	string	Otimizador utilizado no treinamento do modelo.
padding	['valid', 'same']	string	
batchNormalization	['true', 'false']	booleano	True: normaliza as ativações da camada anterior, ou seja, aplica uma transformação que mantém a ativação média próxima a 0 e o desvio padrão da ativação próximo a 1. False: não faz nada.
max_pool	['true', 'false']	booleano	True: Aplica a operação de max-pooling. False: não faz nada.
pool_size	[2, 3, 4, 5]	inteiro	Fator inteiro pelo qual reduzir (vertical, horizontal). O mesmo comprimento de janela é usado para ambas as dimensões.
scaler	['standard', 'minmax']	string	Método a ser utilizado para deixar os dados na mesma escala.
units	[20, 50, 80, 100, 150, 200, 250, 500]	inteiro	Número de neurônios na camada densa.
feature	['true', 'false']	booleano	True: A Feature é selecionada. False: A Feature não é selecionada.
activate	['true', 'false']	booleano	True: A camada é adicionada ao modelo. False: A camada não é adicionada ao modelo.

3.3.1 Mutação

Devido ao fato do cromossomo ser formado por diferentes tipos de dados, são implementados diversos tipos de mutação, que são utilizados de acordo com o tipo de dado presente na posição escolhida aleatoriamente para sofrear a mutação. Na sequência cada uma delas é apresentada.

3.3.1.1 Mutação booleana

A mutação booleana é aplicada no caso da mutação ocorrer no tipo de dado booleano, negando o valor presente no cromossomo. A Tabela 5 apresenta exemplos de mutações booleanas.

Tabela 5 – Exemplo de mutação booleana.

Valor	Valor após mutação
True	False
False	True

3.3.1.2 Mutação em janela

A mutação em janela é aplicada no caso da mutação ocorrer no tipo de dado inteiro ou real. Como neste caso os valores são ordinais, o valor após a mutação depende do valor presente no cromossomo. A mutação consiste em pegar 2 elementos maiores e 2 elementos menores mais próximos do valor presente no cromossomo e escolher aleatoriamente entre eles o novo valor. A Tabela 6 apresenta exemplos de mutações em janela.

Tabela 6 – Exemplo de mutação em janela.

Posição	Valor	Possíveis valores	Valores da janela	Valor após mutação
dropout	0.1	[0.0, 0.1, 0.2, 0.3, 0.4]	[0.0, 0.2, 0.3]	0.2
dropout	0.2	[0.0, 0.1, 0.2, 0.3, 0.4]	[0.0, 0.1, 0.3, 0.4]	0.4
epochs	60	[30, 45, 60, 90, 120, 150, 200]	[30, 45, 90, 120]	90
epochs	30	[30, 45, 60, 90, 120, 150, 200]	[45, 60]	60
epochs	200	[30, 45, 60, 90, 120, 150, 200]	[120, 150]	120

3.3.1.3 Mutação nominal

A mutação nominal é aplicada no caso da mutação ocorrer no tipo de dado string, ou seja, em uma variável categórica nominal. Por não haver ordem, um valor diferente do presente no cromossomo é sorteado aleatoriamente. A Tabela 7 apresenta exemplos de mutações nominal.

Tabela 7 – Exemplo de mutação nominal.

Posição	Valor	Possíveis valores	Valor após mutação
optimizer	'adam'	['adam', 'sgd']	'sgd'
padding	'same'	['valid', 'same']	'valid'

Note que quando estão disponíveis apenas dois valores para um elemento do tipo string, a mutação aleatória é semelhante a mutação booleana. Porém, como podem existir mais de dois valores, este tipo de mutação se faz necessária.

3.3.2 Crossover

O método de crossover utilizado foi o crossover de dois pontos. A recombinação em dois pontos consiste em selecionar dois pontos nas cromossomos dos indivíduos. Tudo que está entre estes dois pontos nos cromossomos pais é trocado para gerar os novos cromossomos filhos. A Figura 8 ilustra um crossover de dois pontos.

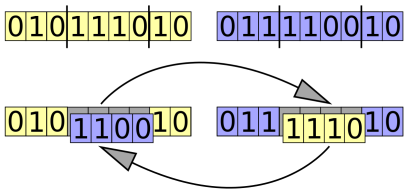


Figura 8 – Exemplo Crossover de dois pontos.

Fonte: https://pt.wikipedia.org/wiki/Recombinação(*computação_evolutiva*)

A Figura 5 mostra a arquitetura dos modelos utilizados neste trabalho. A classificação obtida por cada modelo foi comparada com a classificação padrão-ouro baseada no diagnóstico final no prontuário eletrônico.

3.4 Forma de Análise dos Resultados

Como dito na Seção 1.1, poucos trabalhos da literatura investigaram o uso de Aprendizado de Máquina (AM) para o auxílio ao diagnóstico de FVC utilizando-se IRM, e todos os trabalhos relacionados investigaram a utilizaram de métodos de AM para detecção de fraturas não diferenciando sua etiologia. Neste mestrado, três modelos baseados em RNAs são comparados. Os modelos diferem na maneira como as entradas são consideradas. O modelo híbrido proposto é comparado com RNAs treinadas com características de cada um dos tipos isoladamente. São utilizadas medidas padrão para comparação de classificadores na área médica, como acurácia balanceada, sensibilidade, especificidade e área sob a curva ROC.

4

Resultados

Nesta seção são apresentados os resultados experimentais do sistema híbrido proposto. O sistema híbrido é comparado a duas abordagens: i) abordagem tradicional de aprendizado de máquina utilizando um modelo de MLP treinado com atributos extraídos das imagens, dados clínicos adicionais e artibutos extraídos do histograma da imagem como entradas; ii) CNN treinada utilizando apenas imagens (brutas) como entrada. Conforme discutido na Seção 3, outras configurações são também testadas. No total, 6 modelos são comparados: 1) MLP (usando apenas radiômicas e outras informações adicionais); 2) CNN (usando apenas imagens brutas) otimizada manualmente; 3) CNN (usando apenas imagens brutas) otimizada manualmente considerando aumento de dados; 4) CNN (usando apenas imagens brutas) pré-treinada; 5) Modelo Híbrido utilizando CNN pré-treinada; 6) Modelo Híbrido otimizado pelo AG. Os resultados da validação cruzada e do conjunto de testes são mostrados considerando o melhor modelo obtido em cada abordagem. No Apêndice B se encontram os tempos médios de execução de cada modelo e a configuração da máquina utilizada nos experimentos.

A última camada de cada CNN é uma camada Softmax. Ou seja, a camada Softmax atribui probabilidades decimais a cada classe em um problema de várias classes. Essas probabilidades decimais devem somar 1.0. A classe prevista corresponde à classe com o maior valor de probabilidade.

Por empregar soluções iniciais aleatórias e por utilizar operadores estocásticos, o AG foi executado 5 vezes com diferentes sementes pseudo-aleatórias. Os parâmetros dos algoritmos foram ajustados interativamente realizando-se diversas simulações. A arquitetura e os parâmetros do melhor modelo encontrado em cada abordagem estudada estão disponíveis no Apêndice C.

O Algoritmo Genético foi executado utilizando uma população de tamanho 30, durante 50 gerações. A taxa de crossover utilizada foi de 60% enquanto a taxa de mutação foi de 2%. Foi utilizado o método de elitismo, onde os 2 melhores indivíduos de cada geração foram passados automaticamente para a próxima geração. O método de seleção

utilizado foi o Torneio de tamanho 3. Os resultados para cada modelo são apresentados a seguir.

4.1 Modelo usando apenas atributos radiômicos e histograma e informações do paciente

4.1.1 MLP

A Tabela 8 mostra os resultados de precisão (para o conjunto de testes) obtidos pelo melhor modelo MLP encontrado usando apenas vetor de características radiômicas, dados clínicos adicionais e atributos extraídos do histograma da imagem. A previsão da MLP para cada exemplo do conjunto de testes é mostrada. Também são mostrados os resultados da MLP, multiplicados por 100%, para cada classe. Na primeira coluna, o índice após "P" indica o número do paciente e o índice após "L" indica o número do corpo vertebral. A acurácia balanceada média obtida no conjunto de treinamento (usando a validação cruzada) foi de 73.43% (esse resultado não é mostrado na tabela), enquanto o melhor resultado para o conjunto de testes foi de 58.8%. Na Tabela 8, os exemplos do conjunto de testes classificados erroneamente são mostrados em negrito. Vale ressaltar que os conjuntos de treinamento e teste são compostos por corpos vertebrais de diferentes pacientes, ou seja, os pacientes no conjunto de testes são diferentes dos pacientes no conjunto de validação cruzada (treinamento).

Tabela 8 – Resultado do conjunto de testes para o MLP que usa apenas vetor de características radiômicas, dados clínicos adicionais e atributos extraídos do histograma da imagem como entradas. A rede neural artificial obteve uma acurácia balanceada no conjunto de teste de 58.8%.

Corpo vertebral	**Classe Real**	**Predito**	**Benigna**	**Maligna**	**Normal**
P2L1	benigna	benigna	84.61%	3.35%	12.04%
P2L2	normal	normal	10.31%	1.24%	88.45%
P2L3	benigna	**normal**	41.99%	3.09%	54.92%
P2L4	normal	normal	4.25%	1.78%	93.97%
P2L5	benigna	**normal**	34.37%	4.80%	60.83%
P16L2	maligna	maligna	0.06%	99.90%	0.04%
P16L4	maligna	maligna	0.04%	99.95%	0.01%
P20L1	benigna	**maligna**	18.87%	76.43%	4.7%
P20L2	normal	normal	3.30%	1.45%	95.25%
P20L3	normal	normal	5.62%	1.06%	93.32%
P20L4	normal	normal	2.28%	2.49%	95.23%
P20L5	normal	normal	1.40%	0.42%	98.18%
P24L1	maligna	**benigna**	71.58%	8.63%	19.79%
P24L2	maligna	**benigna**	66.72%	20.43%	12.85%
P33L1	benigna	**normal**	13.40%	2.32%	84.28%
P33L2	normal	normal	0.44%	0.4%	99.16%
P33L3	normal	normal	1.10%	1.78%	97.12%
P33L4	benigna	**normal**	9.98%	2.82%	87.20%
P33L5	normal	normal	1.76%	1.39%	96.85%
P50L3	maligna	maligna	32.77%	62.89%	4.34%

A Tabela 9 mostra a matriz de confusão obtida pelo modelo, enquanto a Tabela 10 apresenta outras métricas usadas para avaliar o classificador. A Figura 9 mostra as curvas ROC obtidas pelo modelo. Todos esses resultados foram calculados para o conjunto de testes.

Tabela 9 – Matriz de confusão obtida pela MLP.

		Classe obtida		
	Classificação	**Benigna**	**Maligna**	**Normal**
Classe real	Benigna	1	1	4
	Maligna	2	3	0
	Normal	0	0	9

Tabela 10 – Resultados de métricas utilizadas para avaliação da MLP.

	Precisão	**Recall**	**F1-score**	**Support**	**Especificidade**	**Sensibilidade**
Benigna	0.33	0.17	0.22	6	0.85	0.16
Maligna	0.75	0.60	0.67	5	0.93	0.6
Normal	0.69	1.00	0.82	9	0.63	1.0

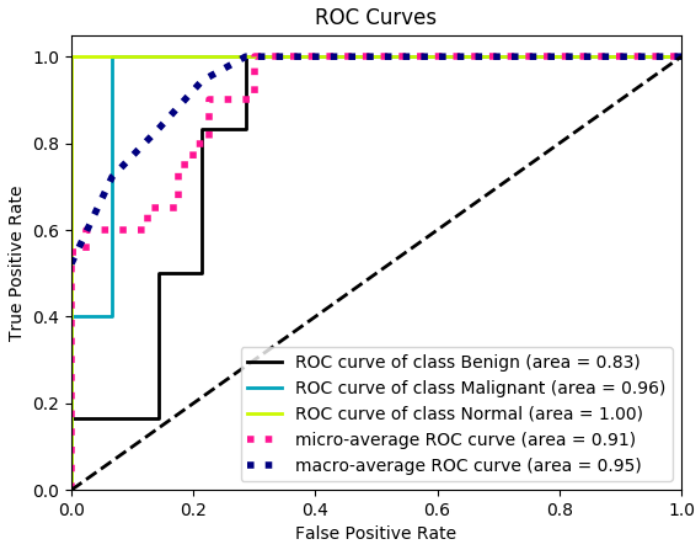


Figura 9 – Curva ROC obtida pela MLP.

4.2 Modelos usando apenas imagens (brutas)

4.2.1 CNN otimizada manualmente

Este modelo de CNN foi otimizado manualmente. Diversos conjuntos de parâmetros foram testados. A melhor configuração encontrada durante o cross-validation (utilizando o conjunto de treinamento) foi selecionada. A Tabela 11 apresenta os resultados de precisão (para o conjunto de testes) obtidos pela melhor CNN otimizada manualmente que usa apenas imagem como entradas. A acurácia balanceada média obtida no treinamento (usando a validação cruzada) foi de 67.8% (resultados não mostrados na tabela), enquanto a acurácia balanceada do conjunto de testes foi de 77.33%. A matriz de confusão é mostrada na Tabela 12, enquanto a Tabela 13 apresenta os resultados de métricas adicionais usadas para avaliar o classificador. A Figura 10 mostra as curvas ROC obtidas pelo modelo.

Tabela 11 – Resultado do conjunto de testes para a CNN otimizada manualmente utilizando apenas imagem como entrada. A CNN obteve uma acurácia balanceada no conjunto de teste de 77.3%.

Corpo vertebral	**Classe Real**	**Predito**	**Benigna**	**Maligna**	**Normal**
P2L1	benigna	benigna	95.88%	4.07%	0.05%
P2L2	normal	normal	0.00%	0.00%	100%
P2L3	benigna	**maligna**	8.17%	52.67%	39.16%
P2L4	normal	normal	39.81%	6.54%	53.65%
P2L5	benigna	benigna	99.98%	0.02%	0.00%
P16L2	maligna	maligna	0.19%	99.78%	0.01%
P16L4	maligna	maligna	0.1%	99.90%	0.00%
P20L1	benigna	**maligna**	0.16%	99.83%	0.01%
P20L2	normal	normal	0.00%	0.00%	100%
P20L3	normal	normal	0.00%	0.00%	100%
P20L4	normal	normal	0.12%	0.02%	99.86%
P20L5	normal	normal	0.00%	0.00%	100%
P24L1	maligna	maligna	28.80%	70.50%	0.70%
P24L2	maligna	**benigna**	92.20%	7.25%	0.55%
P33L1	benigna	benigna	100%	0.00%	0.00%
P33L2	normal	normal	0.00%	0.00%	100%
P33L3	normal	normal	0.01%	0.30%	99.59%
P33L4	benigna	**normal**	0.16%	5.39%	94.45%
P33L5	normal	normal	0.00%	0.00%	100%
P50L3	maligna	maligna	1.85%	98.15%	0.00%

Tabela 12 – Matriz de confusão obtida pela CNN otimizada manualmente.

		Classe obtida		
	Classificação	**Benigna**	**Maligna**	**Normal**
Classe real	Benigna	3	2	1
	Maligna	1	4	0
	Normal	0	0	9

Tabela 13 – Resultados de métricas utilizadas para avaliação da CNN otimizada manualmente.

	Precisão	**Recall**	**F1-score**	**Support**	**Especificidade**	**Sensibilidade**
Benigna	0.75	0.50	0.60	6	0.91	0.5
Maligna	0.67	0.80	0.73	5	0.87	0.8
Normal	0.90	1.00	0.95	9	0.91	1.0

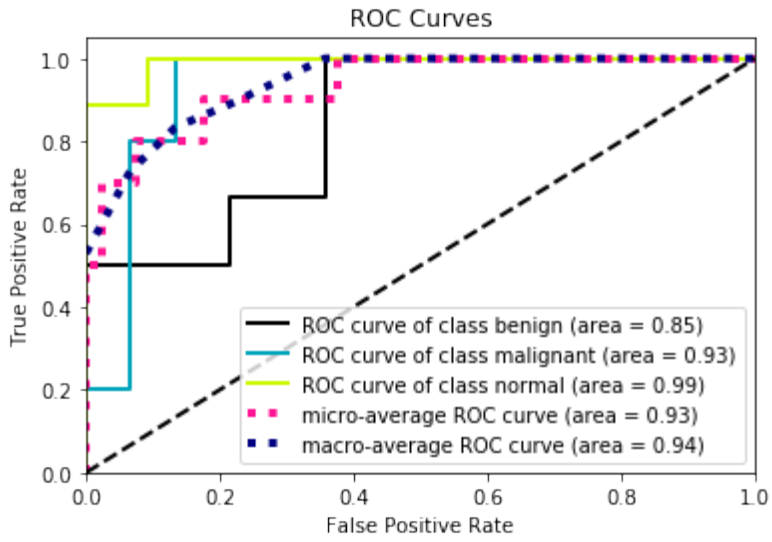


Figura 10 – Curva ROC obtida pela CNN otimizada manualmente.

4.2.2 CNN otimizada manualmente utilizando aumento dos dados

Este modelo de CNN também foi otimizado manualmente, na qual diversos conjuntos de parâmetros foram testados. Este foi o único modelo treinado utilizando aumento da base de dados. A melhor configuração encontrada durante o cross-validation foi selecionada.

A Tabela 14 apresenta os resultados de precisão (para o conjunto de testes) obtidos pela melhor CNN otimizada manualmente que usa apenas imagem como entradas. A acurácia balanceada média obtida no treinamento (usando a validação cruzada) foi de 67.4% (resultados não mostrados na tabela), enquanto a acurácia balanceada do conjunto de testes foi de 75.5%. A matriz de confusão é mostrada na Tabela 15, enquanto a Tabela 16 apresenta os resultados de métricas adicionais usadas para avaliar o classificador. A Figura 11 mostra as curvas ROC obtidas pelo modelo.

Tabela 14 – Resultado do conjunto de testes para a CNN otimizada manualmente utilizando aumento de dados. A CNN obteve uma acurácia balanceada no conjunto de teste de 75.5%.

Corpo vertebral	**Classe Real**	**Predito**	**Benigna**	**Maligna**	**Normal**
P2L1	benigna	benigna	76.56%	23.40%	0.04%
P2L2	normal	normal	0.20%	0.87%	98.93%
P2L3	benigna	**maligna**	19.07%	80.55%	0.38%
P2L4	normal	normal	20.28%	2.34%	77.38%
P2L5	benigna	benigna	94.31%	5.69%	0.00%
P16L2	maligna	maligna	22.46%	77.24%	0.03%
P16L4	maligna	maligna	0.01%	99.39%	0.6%
P20L1	benigna	benigna	52.41%	99.83%	0.01%
P20L2	normal	normal	4.22%	5.38%	90.4%
P20L3	normal	normal	5.04%	0.00%	94.96%
P20L4	normal	normal	1.15%	0.00%	98.85%
P20L5	normal	normal	0.05%	0.11%	99.84%
P24L1	maligna	maligna	41.98%	51.69%	6.33%
P24L2	maligna	**normal**	1.01%	1.33%	97.66%
P33L1	benigna	benigna	81.87%	17.14%	0.99%
P33L2	normal	normal	0.21%	0.19%	99.6%
P33L3	normal	normal	9.13%	0.00%	90.87%
P33L4	benigna	**normal**	3.76%	6.35%	89.89%
P33L5	normal	normal	0.27%	6.74%	92.99%
P50L3	maligna	**benigna**	57.73%	41.86%	0.41%

Tabela 15 – Matriz de confusão obtida pela CNN otimizada manualmente utilizando aumento de dados.

		Classe obtida		
	Classificação	**Benigna**	**Maligna**	**Normal**
Classe real	Benigna	4	1	1
	Maligna	1	3	1
	Normal	0	0	9

Tabela 16 – Resultados de métricas utilizadas para avaliação da CNN otimizada manualmente utilizando aumento de dados.

	Precisão	**Recall**	**F1-score**	**Support**	**Especificidade**	**Sensibilidade**
Benigna	0.8	0.50	0.73	6	0.92	0.666
Maligna	0.75	0.60	0.67	5	0.92	0.6
Normal	0.82	1.00	0.90	9	0.84	1.0

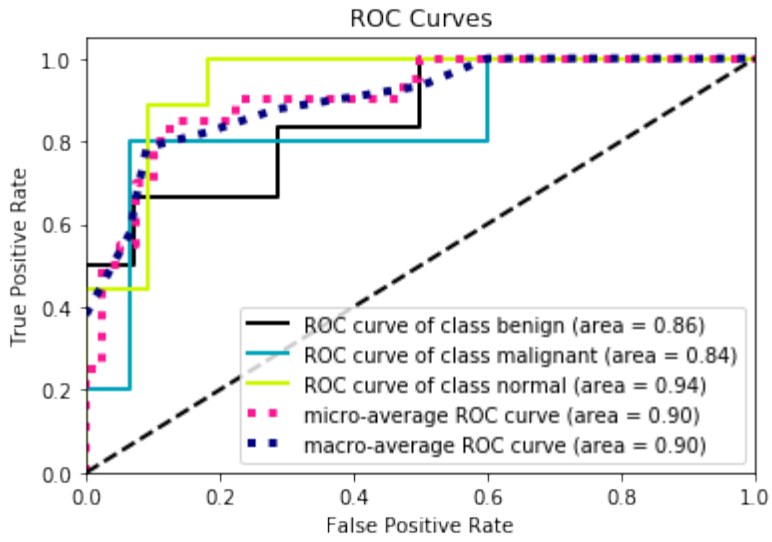


Figura 11 – Curva ROC obtida pela CNN otimizada manualmente utilizando aumento de dados.

4.2.3 CNN Pré-treinada

A Tabela 17 apresenta os resultados de precisão (para o conjunto de testes) obtidos pela *VGG16* que usa apenas imagem como entrada. A acurácia balanceada média obtida no treinamento (usando a validação cruzada) foi de 71.9%, e no conjunto de imagens de teste foi de 82.96%.

Tabela 17 – Resultado do conjunto de testes para a *VGG16* utilizando apenas imagem como entrada. A CNN pré-treinada obteve uma acurácia balanceada no conjunto de teste de 82.96%.

Corpo vertebral	Classe Real	Predito	Benigna	Maligna	Normal
P2L1	benigna	benigna	99.66%	0.06%	0.28%
P2L2	normal	normal	0.02%	0.01%	99.97%
P2L3	benigna	benigna	91.83%	2.84%	5.33%
P2L4	normal	normal	0.25%	0.05%	99.70%
P2L5	benigna	benigna	75.94%	0.26%	23.80%
P16L2	maligna	**benigna**	43.95%	43.80%	12.25%
P16L4	maligna	maligna	8.00%	87.80%	4.20%
P20L1	benigna	benigna	57.03%	35.74%	7.23%
P20L2	normal	normal	0.23%	3.10%	96.67%
P20L3	normal	normal	0.20%	0.18%	99.62%
P20L4	normal	normal	1.40%	2.62%	95.98%
P20L5	normal	normal	0.01%	0.01%	99.98%
P24L1	maligna	**benigna**	77.16%	29.74%	3.10%
P24L2	maligna	maligna	35.51%	58.56%	5.93%
P33L1	benigna	benigna	95.25%	3.17%	1.58%
P33L2	normal	normal	0.14%	0.50%	99.36%
P33L3	normal	**maligna**	8.56%	71.26%	20.18%
P33L4	benigna	benigna	73.99%	14.34%	11.67%
P33L5	normal	normal	0.04%	0.12%	99.84%
P50L3	maligna	maligna	9.29%	86.72%	3.99%

A Tabela 18 mostra a matriz de confusão obtida pelo modelo no conjunto de teste,

enquanto a Tabela 19 apresenta outras métricas utilizadas para avaliar o classificador, e finalmente a Figura 12 apresenta a curva ROC obtida pelo modelo.

Tabela 18 – Matriz de confusão obtida pela *VGG16*.

		Classe obtida		
	Classificação	**Benigna**	**Maligna**	**Normal**
Classe real	Benigna	6	0	0
	Maligna	2	3	0
	Normal	0	1	8

Tabela 19 – Resultados de métricas utilizadas para avaliação da *VGG16*.

	Precisão	**Recall**	**F1-score**	**Support**	**Especificidade**	**Sensibilidade**
Benigna	0.75	1.00	0.86	6	0.85	1.0
Maligna	0.75	0.60	0.67	5	0.93	0.6
Normal	1.00	0.89	0.94	9	1.00	0.88

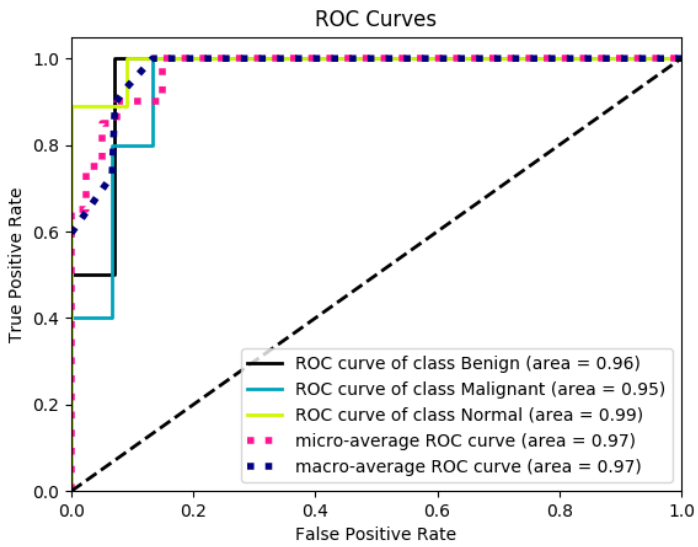


Figura 12 – Curva ROC obtida pela *VGG16*.

4.3 Modelos usando todas as fontes de informações

4.3.1 Modelo Híbrido Proposto utilizando CNN pré-treinada

O resultado obtido pelo modelo híbrido proposto utilizando *VGG16* como modelo de CNN pré-treinada são apresentados aqui. O modelo utiliza como entrada as imagens (brutas) e

o vetor de características radiômicas, dados clínicos adicionais e atributos extraídos do histograma da imagem são inseridos na primeira camada densa.

A Tabela 20 mostra os resultados de precisão para o conjunto de testes. A acurácia balanceada média obtida no conjunto de treinamento (usando a validação cruzada) foi de 71.53% (resultados não mostrados na tabela), enquanto a acurácia do conjunto de testes foi de 87.77%.

Tabela 20 – Resultado do conjunto de testes para o modelo híbrido utilizando todas as fontes de informações. O modelo obteve uma acurácia balanceada no conjunto de teste de 87.77%.

Corpo vertebral	**Classe Real**	**Predito**	**Benigna**	**Maligna**	**Normal**
P2L1	benigna	benigna	96.60%	3.22%	0.19%
P2L2	normal	normal	0.30%	0.05%	99.65%
P2L3	benigna	benigna	74.01%	24.16%	1.83%
P2L4	normal	normal	0.08%	0.01%	99.91%
P2L5	benigna	benigna	78.83%	0.76%	20.41%
P16L2	maligna	maligna	0.96%	98.35%	0.69%
P16L4	maligna	maligna	0.36%	99.52%	0.12%
P20L1	benigna	**maligna**	14.83%	84.90%	0.27%
P20L2	normal	normal	0.12%	0.64%	99.24%
P20L3	normal	normal	0.15%	0.16%	99.69%
P20L4	normal	normal	0.06%	0.1%	99.84%
P20L5	normal	normal	0.00%	0.00%	1.00%
P24L1	maligna	**benigna**	51.11%	48.35%	0.54%
P24L2	maligna	maligna	39.80%	59.06%	1.14%
P33L1	benigna	benigna	90.52%	6.74%	2.74%
P33L2	normal	normal	0.05%	0.08%	99.87%
P33L3	normal	normal	26.70%	16.50%	56.80%
P33L4	benigna	benigna	76.69%	4.72%	18.59%
P33L5	normal	normal	0.01%	0.01%	99.98%
P50L3	maligna	maligna	33.19%	66.57%	0.24%

A Tabela 21 apresenta a matriz de confusão, enquanto a Tabela 22 apresenta os resultados de métricas adicionais. A Figura 13 mostra as curvas ROC obtidas pelo modelo.

Tabela 21 – Matriz de confusão obtida pelo Modelo Híbrido.

		Classe obtida		
	Classificação	**Benigna**	**Maligna**	**Normal**
Classe real	Benigna	5	1	0
	Maligna	1	4	0
	Normal	0	0	9

Tabela 22 – Resultados de métricas utilizadas para avaliação do Modelo Híbrido.

	Precisão	**Recall**	**F1-score**	**Support**	**Especificidade**	**Sensibilidade**
Benigna	0.83	0.83	0.83	6	0.93	0.83
Maligna	0.80	0.80	0.80	5	0.93	0.80
Normal	1.00	1.00	1.00	9	1.00	1.00

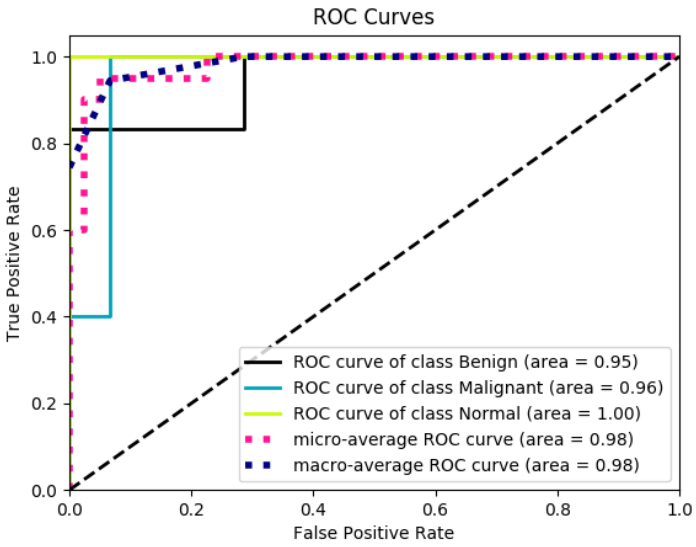


Figura 13 – Curva ROC obtida pelo Modelo Híbrido.

4.3.2 Modelo Híbrido Proposto otimizado pelo Algoritmo Genético

A Tabela 23 ilustra o resultado obtido pelo melhor indivíduo de cada uma das 5 execuções do AG. Os atributos selecionados pelos indivíduos estão disponíveis no Apêndice D.

Tabela 23 – Resultado do melhor indivíduo de cada uma das 5 execuções do AG, avaliada no conjunto de treinamento (utilizando cross-validation) e no conjunto de teste.

Exec	Cross-validation	Test set
1	0.757	0.644
2	0.787	0.729
3	0.792	0.700
4	0.748	0.644
5	0.770	0.755
Média	0.770	0.694
std	0.018	0.049

Como a execução 3 obteve a maior acurácia balanceada durante o cross-validation, os resultados desta execução que foram considerados. A Tabela 24 mostra os resultados de precisão obtidos pelo modelo híbrido otimizado pelo AG no conjunto de teste. A acurácia balanceada média obtida no conjunto de treinamento (utilizando validação cruzada) foi de 79.25%, enquanto a acurácia balanceada obtida no conjunto de teste foi de 70.00%.

Tabela 24 – Resultado do conjunto de testes para o modelo híbrido otimizado pelo AG utilizando todas as fontes de informações. O modelo obteve uma acurácia balanceada no conjunto de teste de 70.00%.

Corpo vertebral	**Classe Real**	**Predito**	**Benigna**	**Maligna**	**Normal**
P2L1	benigna	benigna	99.34%	0.25%	0.41%
P2L2	normal	normal	0.00%	0.01%	99.99%
P2L3	benigna	**normal**	30.22%	1.14%	68.64%
P2L4	normal	normal	0.16%	0.02%	99.82%
P2L5	benigna	**normal**	00.80%	0.26%	98.94%
P16L2	maligna	maligna	0.04%	99.94%	0.02%
P16L4	maligna	maligna	0.01%	99.98%	0.01%
P20L1	benigna	benigna	56.18%	42.85%	0.97%
P20L2	normal	normal	0.43%	0.10%	99.47%
P20L3	normal	normal	15.20%	0.02%	84.78%
P20L4	normal	normal	4.68%	0.01%	95.31%
P20L5	normal	normal	0.73%	0.01%	99.26%
P24L1	maligna	**benigna**	99.55%	00.42%	0.03%
P24L2	maligna	**benigna**	96.60%	3.03%	0.37%
P33L1	benigna	benigna	78.13%	0.10%	21.77%
P33L2	normal	normal	0.41%	0.00%	99.59%
P33L3	normal	normal	4.86%	0.02%	95.12%
P33L4	benigna	**normal**	8.22%	0.06%	91.72%
P33L5	normal	normal	0.25%	0.01%	99.74%
P50L3	maligna	maligna	22.09%	77.87%	0.04%

A Tabela 25 ilustra a matriz de confusão obtida pelo conjunto de teste, na sequência, a Tabela 26 apresenta os resultados de métricas utilizadas para avaliar o classificador, e finalmente a Figura 14 apresenta a curva ROC obtida pelo modelo.

Tabela 25 – Matriz de confusão obtida pelo Modelo Híbrido otimizado pelo AG.

		Classe obtida		
	Classificação	**Benigna**	**Maligna**	**Normal**
Classe real	Benigna	3	0	3
	Maligna	2	3	0
	Normal	0	0	9

Tabela 26 – Resultados de métricas utilizadas para avaliação do Modelo Híbrido otimizado pelo AG.

	Precisão	**Recall**	**F1-score**	**Support**	**Especificidade**	**Sensibilidade**
Benigna	0.6	0.50	0.55	6	0.85	0.5
Maligna	1.00	0.60	0.75	5	1.00	0.6
Normal	0.75	1.00	0.86	9	0.72	1.00

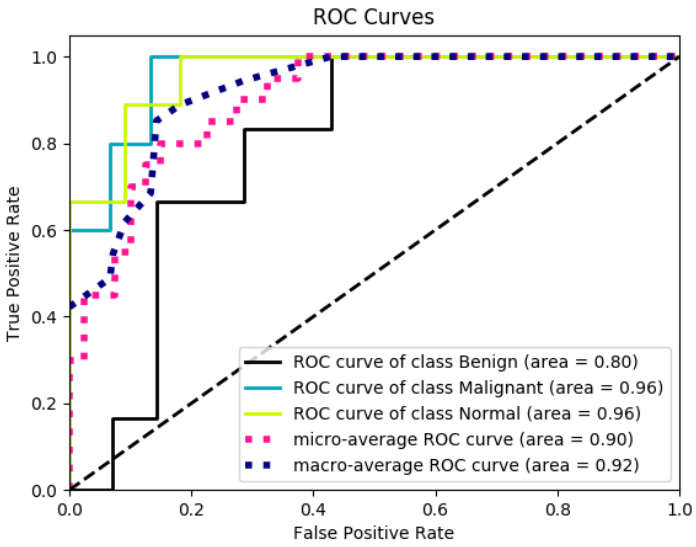


Figura 14 – Curva ROC obtida pelo Modelo Híbrido otimizado pelo AG.

4.4 Comparação dos Resultados

Para comparar estatisticamente os classificadores, foram utilizados as médias na 10-fold stratified cross validation das 10 execuções.

A Tabela 27 apresenta o resultado obtido em cada execução para cada um dos modelos estudados para separação entre as classes Benigna x Maligna x Normal. A CNN otimizada manualmente utilizando apenas as imagens como entrada foi considerada modelo referência.

Após verificar que os critérios de normalidade e variância, foi aplicado o *Teste t de Student*, concluindo-se que, para o nível de significância de 5%, existe diferença significativa entre o Modelo Híbrido otimizado pelo AG e a CNN otimizada manualmente.

A Tabela 28 resume os resultados obtidos pelos modelos que utilizaram diferentes fontes de informação como entrada. A Tabela 29 mostra o(s) modelo(s) com maior(es) sensibilidade e especificidade para cada classe.

Tabela 27 – Acurácia obtida em cada uma das execuções com 10-fold stratified cross validation para separação das classes Benigna x Maligna x Normal. A letra S indica que o resultado é estatisticamente significante, considerando-se o *Teste t*, se comparado com o modelo referência. Os símbolos + e - significam respectivamente que médias obtidas pelo modelo foi maior ou menor que a média do modelo referência.

Exec	**MLP**	**CNN otimizada manualmente**	**CNN otimizada manualmente aumento de dados**	**CNN Pré-treinada**	**Modelo Híbrido utilizando CNN pré-treinada**	**Modelo Híbrido otimizado pelo AG**
1	0.666	0.640	0.656	0.944	0.833	0.833
2	0.411	0.697	0.669	0.705	0.588	0.705
3	0.882	0.676	0.686	0.647	0.764	0.882
4	0.941	0.650	0.728	0.647	0.882	0.764
5	0.764	0.698	0.703	0.941	0.823	0.823
6	0.823	0.705	0.628	0.470	0.823	0.882
7	0.529	0.626	0.686	0.882	0.529	0.823
8	0.823	0.698	0.644	0.705	0.470	0.647
9	0.812	0.681	0.680	0.750	0.812	0.750
10	0.687	0.662	0.662	0.500	0.625	0.812
Média	0.734 (+)	0.678	0.674 (-)	0.719 (+)	0.715 (+)	0.792 (S+)
std	0.163	0.029	0.017	0.166	0.147	0.075

Tabela 28 – Resumo dos resultados obtidos por cada um dos modelos utilizando diferentes fontes de informação.

Métrica	**Conjunto**	**MLP**	**CNN otimizada manualmente**	**CNN otimizada manualmente aumento de dados**	**CNN pré-treinada**	**Modelo Híbrido utilizando CNN pré-treinada**	**Modelo Híbrido otimizado pelo AG**
Acurácia balanceada (cross-validation)	Treinamento	73.4%	67.8%	67.4%	71.9%	71.5%	**79.2%**
Acurácia balanceada	Teste	58.8%	77.3%	75.5%	82.9%	**87.7%**	70.0%
F1-Score	Teste	57%	76%	76.6%	76.6%	**87.6%**	70%
Especificidade	Teste	80.3%	89.6%	89.3%	92.6%	**95.3%**	85.6%
Sensibilidade	Teste	58.6%	76.6%	75.3%	82.6%	**87.6%**	70%
Macro-average ROC	Teste	95%	94%	90%	97%	**98%**	92%

Tabela 29 – Modelo com maior sensibilidade e especificidade para cada classe de FVC.

	Especificidade	**Sensibilidade**
Benigno	Modelo Híbrido com CNN pré-treinada (0.93)	CNN pré-treinada (1.0)
Maligno	Modelo Híbrido otimizado pelo AG (1.0)	CNN otimizada manualmente e Modelo Híbrido com CNN pré-treinada (0.8)
Normal	CNN pré-treinada e Modelo Híbrido com CNN pré-treinada (1.0)	MLP CNN otimizada manualmente CNN otimizada manualmente utilizando aumento de dados Modelo Híbrido com CNN pré-treinada Modelo Híbrido otimizado pelo AG (1.0)

5

Discussão

Algumas observações podem ser feitas sobre os resultados apresentados nas Tabelas 27, 28 e 29.

5.1 Impacto do aumento de dados

Primeiro, a precisão da CNN com aumento de dados (Seção 4.2.2) foi um pouco pior do que a precisão da CNN sem aumento de dados (Seção 4.2.1). Nos experimentos para ambos os modelos, apenas imagens (brutas) foram usadas como entradas. Na Tabela 28, pode-se observar que tanto para o conjunto de validação cruzada quanto para o conjunto de teste o modelo sem aumento de dados obteve um resultado melhor. No entanto, a CNN usando o aumento de dados foi ligeiramente melhor no conjunto de teste se analisarmos o F1-Score. O conjunto de dados foi aumentado fazendo modificações nas imagens do corpo vertebral. Essas modificações artificiais no conjunto de dados não melhoraram significativamente a generalização do modelo CNN. Isso pode ser explicado porque o conjunto de dados é muito pequeno e as imagens originais no conjunto de dados têm algumas propriedades, por exemplo, rotação, que são semelhantes.

Além disso, o tempo médio para executar o modelo com aumento de dados é muito mais longo: 6 min para o modelo sem aumento de dados versus 13 min para o modelo com aumento de dados (Apêndice B).

5.2 Impacto de se utilizar arquiteturas pré-treinadas

Apesar do pequeno conjunto de dados combinado a uma grande complexidade das CNNs (um grande número de pesos) no modelo pré-treinado, a CNN pré-treinada (Seção 4.2.3)

obteve resultados superiores ao resultados obtidos pelas CNNs otimizadas manualmente (Seções 4.2.1 e 4.2.2) tanto para o conjunto de treino (validação cruzada) quanto para o conjunto de teste. A CNN pré-treinada era composta por 24 camadas, enquanto as CNNs otimizadas manualmente possuíam apenas 13 camadas. Quanto mais profunda a CNN, maior o poder computacional necessário para treina-la. Logo, os modelos otimizados manualmente tiveram sua profundidade limitada devido ao poder computacional disponível.

Também é importante observar que os conjuntos de dados originalmente usados para treinar a CNN pré-treinada eram compostos por imagens muito diferentes das ressonâncias magnéticas do conjunto de treinamento aqui empregado. Logo, utilizar um modelo pré-treinado um dataset genérico como o *ImageNet*, permitiu o modelo descobrir características interessantes para as imagens genéricas, não consumindo muito tempo no re-treino e ficando especializada no dataset aqui empregado.

5.3 Impacto de se utilizar informações radiômicas e adicionais

Ao analisar a Tabela 28, pode-se observar que a CNN pré-treinada treinada com imagens (brutas) (Seção 4.2.3) obteve resultados ligeiramente melhores para o conjunto de validação cruzada (treinamento) do que o modelo híbrido proposto utilizando CNN pré-treinada (Seção 4.3.1). No entanto, o modelo híbrido proposto utilizando CNN pré-treinada obteve resultados melhores para o conjunto de teste. Esses resultados indicam que os atributos de radiômica, histograma da imagem e as características do paciente foram importantes para a classificação das ressonâncias magnéticas neste conjunto de dados. É importante observar que as características radiômicas foram extraídos da IRM 3D usando a biblioteca *pyRadiomics*. Consequentemente o vetor de características pode conter atributos relevantes presentes apenas na imagem 3D, não sendo passíveis de serem extraídos da imagem 2D utilizando CNNs. Além disso, o vetor contém características de forma e textura semelhantes às características analisadas pelo radiologista ao classificar FVCs benignos e malignos. A CNN pré-treinada obteve precisão de 75% para as classes benigna e maligna (Tabela 19), enquanto o modelo híbrido obteve precisão de 83% para a classe benigna e 80% para a classe maligna (Tabela 22). A estratégia de combinar diferentes fontes de informações fornece diferentes características que podem potencializar o resultado obtido pelo classificador, ajudando a diferenciar as FVCs.

A Tabela 28 mostra que quando os atributos de radiômica e adicionais foram incorporados à CNN no sistema híbrido proposto (Seções 4.3.1 e 4.3.2), os modelos obtiveram acurácia balanceada superior na validação cruzada. O modelo híbrido utilizando a CNN pré-treinada também obteve acurácia balanceada superior para o conjunto de

teste em relação a CNN treinada apenas com as imagens (brutas) (Seções 4.2.1 e 4.2.2). Isso mostra que a combinação de todas as fontes de informação foi útil para classificar FVCs neste conjunto de dados. Em suma, a CNN que usa apenas imagens (brutas) como input obteve bons resultados, conseguindo detectar a presença de fraturas vertebrais por compressão na ressonância magnética, diferenciando corpos vertebrais normais de corpos fraturados. No entanto, ele teve dificuldades em diferenciar FVCs malignos de FVCs benignos. O uso de atributos de radiômica e adicionais no sistema híbrido aqui proposto melhorou o desempenho do classificador. Ao usar as informações da radiômica, por exemplo, características de forma e textura, combinadas com imagens (brutas) e características adicionais, o sistema híbrido empregou uma estratégia semelhante àquela adotada por um radiologista, que usa o máximo de informações possível para classificar uma ressonância magnética.

5.4 Impacto de se utilizar o AG para seleção de atributos e hiper-parâmetros

A Tabela 37 (Apêndice D) apresenta quais foram os atributos selecionados pela melhor solução encontrada em cada uma das 5 execuções pelo Algoritmo Genético. O atributo clínico idade foi selecionado nas 5 execuções, o número correspondente ao corpo vertebral foi selecionado 3 vezes, enquanto o atributo sexo foi selecionado apenas 1 vez. Em relação aos atributos extraídos do histograma da imagem, o atributo max foi selecionado 4 vezes e os atributos mean, min e STD foram selecionados 3 vezes. Portanto os atributos adicionais se mostraram relevantes para o problema estudado.

Além da idade, o atributo de intensidade de nível de cinza *firstorder RobustMeanAbsoluteDeviation* foi selecionado nas 5 execuções do AG. Dois atributos de forma (*shape Compactness1* e *shape Maximum2DDiameterRow*), três atributos de intensidade de cinza (*firstorder Energy, firstorder InterquartileRange* e *firstorder Skewness*), sete atributos de textura (*gldm DependenceVariance, glrlm ShortRunEmphasis, glrlm RunEntropy, glszm ZoneEntropy, glszm LargeAreaEmphasis, glszm HighGrayLevelZoneEmphasis* e *glszm LargeAreaLowGrayLevelEmphasis*) e o atributo adicionar Max foram selecionados em 4 das 5 execuções. A descrição de cada um dos atributos está disponível na documentação do *pyRadiomics* [1]. Esse atributos podem ser analisados pelo radiologista afim de auxiliar na tomada de decisão.

Como dito anteriormente, a Tabela 28 mostra que, quando os atributos radiômicos e adicionais foram incorporados à CNN pré-treinada no sistema híbrido proposto, o modelo alcançou resultados superiores. Isso mostra que a combinação de todas as fontes de

[1] https://pyradiomics.readthedocs.io/en/latest/features.html

informação foi útil para classificar as FVCs, melhorando o desempenho do classificador. O sistema híbrido empregou uma estratégia semelhante à adotada por um radiologista, que utiliza o máximo de informações possível para classificar uma ressonância magnética.

A acurácia balanceada obtida na validação cruzada pelo modelo híbrido otimizado pelo AG (Seção 4.3.2) foi estatisticamente superior a obtida pela CNN otimizada manualmente (Seções 4.2.1 e 4.2.2) e pelo modelo de MLP (Seção 4.1.1). O resultado demostra que o AG consegue encontrar bons hiper-parâmetros e selecionar atributos relevantes para a construção do classificador com eficiência e de forma automática. Além disso, os atributos selecionados pelo AG podem ser analisados e posteriormente serem utilizados na prática clínica.

A CNN no modelo híbrido proposto utilizando CNN pré-treinada (Seção 4.3.1) é mais profunda que a CNN no modelo híbrido otimizado pelo AG (Seção 4.3.2), o que explica o pior resultado para o conjunto de teste. Entretanto, o modelo otimizado pelo AG obteve os melhores resultados para a validação cruzada. Isso mostra que o método proposto é eficiente, apesar de a generalização ter sido pior para o conjunto de teste devido ao uso do pré-treinamento e de uma rede mais profunda.

Como pode ser visto nas Tabelas 28 e 29, a CNN pré-treinada obteve a mais alta sensibilidade para a classe benigna, e o modelo híbrido com CNN pré-treinada obteve a maior especificidade para esta classe. Para a classe maligna, o modelo com maior especificidade foi o modelo híbrido otimizado pelo AG, e com maior sensibilidade os modelos CNN otimizada manualmente e o modelo híbrido utilizando CNN pré-treinada. Em relação à classe normal, os modelos com maior especificidade foram a CNN pré-treinada e o modelo híbrido também utilizando a CNN pré-treinada. Em relação a sensibilidade da classe normal, a MLP, as CNNs otimizadas manualmente, e ambos os modelos híbridos obtiveram sensibilidade máxima. Nenhum classificador conseguiu obter a maior especificidade e nem a maior sensibilidade para todas as classes, mas em média o modelo híbrido utilizando CNN pré-treinada foi o modelo que obteve o melhor resultado para ambas as métricas.

6

Conclusão

Neste trabalho, é proposto um método híbrido para o problema de classificação de FVcs em IRMs: i) a incorporação de informações radiômicas, clínicas e de histograma na CNN, inserindo-as na primeira camada densa; ii) a utilização do AG para selecionar os atributos e os hiperparâmetros da CNN. Foram realizados testes com vários modelos, inclusive com aumento da base de dados e com redes pré-treinadas. Quando as informações radiômicas, clínicas e de histograma foram incorporadas à CNN no sistema híbrido os modelos obtiveram acurácia superior na validação cruzada. Além disso, o uso de modelos pré-treinados trouxe vantagens significativas em relação a CNNs otimizadas manualmente.

A extração de características por meio de filtros pré-definidos utilizando a biblioteca *pyRadiomics* fornece características já conhecidas e relevantes para problemas da área médica. Combinar esses filtros com novos filtros provenientes da CNN potencializam o resultado do classificador.

Em resumo, o modelo híbrido é muito promissor. A estratégia de combinar várias fontes de informações, semelhante à adotada por um radiologista na análise de caso, apresentou resultado superior às abordagens tradicionais aplicadas separadamente, como realizar a extração de características por meio de filtros pré-definidos ou utilizando apenas uma CNN.

No futuro, outras abordagens para o treinamento do modelo híbrido podem ser investigadas, como por exemplo um ensemble de classificadores treinados individualmente com uma junção tardia no intuito de aproveitar os benefícios de cada classificador e melhorar o resultado do modelo final. Outro trabalho futuro relevante é projetar outras estratégias para combinar os atributos radiômicos e imagens (brutas).

O sistema aqui proposto usa como entradas as imagens de corpos vertebrais isolados, enquanto um radiologista usa a imagem inteira, podendo comparar o corpo vertebral com outros corpos vertebrais do mesmo paciente. Assim, um trabalho futuro é usar informações de corpos vertebrais vizinhos para classificar um determinado exemplo. Atualmente, o número de imagens é pequeno, o que é uma realidade em muitos conjuntos de dados do

mundo real relacionados à Medicina que são investigados por meio do Aprendizado de Máquina. Melhorias no AG e otimizações para diminuir o tempo de execução são também trabalhos futuros adicionais. Também será investigado o AG empregado para otimizar o modelo pré-treinado, deixando os parâmetros das camadas iniciais e intermediárias fixas e otimizando apenas os parâmetros das camadas densas e os atributos de radiômica e adicionais.

A metodologia proposta neste trabalho é genérica e pode ser aplicada para outros problemas de análise de imagens médicas. Como exemplo, esta metodologia foi aplicada no diagnóstico de COVID-19 utilizando raio-x de tórax (Apêndice F), conseguindo alcançar resultados similares ao da *COVID-Net* [1] utilizando menos de 4% do dataset. Um trabalho futuro é aplicar os métodos propostos em outras áreas do conhecimento.

Todos os códigos desenvolvidos durante o projeto são disponibilizados gratuitamente na plataforma de compartilhamento e hospedagem de código-fonte GitHub. Mais detalhes sobre o código-fonte estão disponíveis no Apêndice E

[1] Iniciativa de código aberto, em fase de pesquisa, que está sendo muito utilizado como modelos de referência. A COVID-Net está disponível em: `https://github.com/lindawangg/COVID-Net`

Referências

S. Al-Helo, R. S. Alomari, S. Ghosh, V. Chaudhary, G. Dhillon, M. B. Al-Zoubi, H. Hiary, and T. M. Hamtini. Compression fracture diagnosis in lumbar: a clinical cad system. *International Journal of Computer Assisted Radiology and Surgery*, 8(3):461–469, May 2013. ISSN 1861-6429. doi: 10.1007/s11548-012-0796-0. URL `https://doi.org/10.1007/s11548-012-0796-0`.

P. M. d. Azevedo-Marques. Diagnóstico auxiliado por computador na radiologia. *Radiologia Brasileira*, 34(5):285–293, 2001.

C. Barreto. Whitebook: como diagnosticar a pneumonia por covid-19?, Apr 2020. URL `https://pebmed.com.br/whitebook-como-diagnosticar-a-pneumonia-por-covid-19/`.

L. F. C. S. Bastos. Opas/oms brasil - folha informativa – covid-19 (doença causada pelo novo coronavírus): Opas/oms, Feb 2020. URL `https://www.paho.org/bra/index.php?option=com_content&view=article&id=6101:covid19&Itemid=875`.

C. A. Cuenod, J.-D. Laredo, S. Chevret, B. Hamze, J.-F. Naouri, X. Chapaux, J.-M. Bondeville, and J.-M. Tubiana. Acute vertebral collapse due to osteoporosis or malignancy: Appearance on unenhanced and gadolinium-enhanced mr images. *Radiology*, 199(2):541–549, 1996.

I. Dimitrovski, D. Kocev, I. Kitanovski, S. Loskovska, and S. Džeroski. Improved medical image modality classification using a combination of visual and textual features. *Computerized Medical Imaging and Graphics*, 39:14 – 26, 2015. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2014.06.005. URL `http://www.sciencedirect.com/science/article/pii/S0895611114000986`. Medical visual information analysis and retrieval.

H. K. Genant, C. Wu, C. van Kuijk, and M. C. Nevitt. Vertebral fracture assessment using a semiquantitative technique. *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*, 8 9:1137–48, 1993.

S. Ghosh, R. Alomari, V. Chaudhary, and G. Dhillon. Automatic lumbar vertebra segmentation from clinical ct for wedge compression fracture diagnosis. volume 796303-796309, 02 2011.

H. Greenspan, B. van Ginneken, and R. M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, May 2016. ISSN 0278-0062. doi: 10.1109/TMI.2016.2553401.

S. Haykin. *Redes Neurais - Princípios e prática*. Bookman, 2rd edition, 2003.

H.-S. Jung, W.-H. Jee, T. R. McCauley, K.-Y. Ha, and K.-H. Choi. Discrimination of metastatic from acute osteoporotic compression spinal fractures with mr imaging. *Radiographics*, 23(1):179–187, 2003.

S. Kasai, F. Li, J. Shiraishi, Q. Li, and K. Doi. Computerized detection of vertebral compression fractures on lateral chest radiographs: Preliminary results with a tool for early detection of osteoporosis. *medical physics*, 33(12):4664–74, 2006.

J. Ker, L. Wang, J. Rao, and T. Lim. *Deep Learning Applications in Medical Image Analysis*, volume 6. IEEE, 2017.

Y. Komeda, H. Handa, T. Watanabe, T. Nomura, M. Kitahashi, T. Sakurai, A. Okamoto, T. Minami, M. Kono, T. Arizumi, M. Takenaka, S. Hagiwara, S. Matsui, N. Nishida, H. Kashida, and M. Kudo. *Computer-Aided Diagnosis Based on Convolutional Neural Network System for Colorectal Polyp Classification: Preliminary Experience*, volume 93. Oncology, 2017.

R. S. D. Lama. Uso de redes neurais convolucionais para classificação de fraturas vertebrais por compressão, 2018.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–44, 05 2015. doi: 10.1038/nature14539.

R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji. Deep learning based imaging data completion for improved brain disease diagnosis. 17(3):305–312, 2014.

Z. Q. L. Linda Wang and A. Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images, 2020.

G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. der Laak, B. Ginneken, and C. I.Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

T. Liu, S. Fang, Y. Zhao, P. Wang, and J. Zhang. *Implementation of Training Convolutional Neural Networks*. 2015.

M. Mitchell. *An introduction to genetic algorithms*. MIT Press, 1996.

W. H. Organization. Coronavirus disease (covid-19), 2020. URL `https://www.who.int/emergencies/diseases/novel-coronavirus-2019`.

L. F. Pereira. Classificação semiautomática de fraturas vertebrais e malignas em imagens de ressonância magnética. Master's thesis, Universidade de São Paulo, Ribeirão Preto, Brasil, 2016.

L. F. Pereira, R. Menezes-Reis, G. A. Metzner, R. M. Rangayyan, M. H. Nogueira-Barbosa, and P. M. Azevedo-Marques. Classification of vertebral compression fractures in magnetic resonance images using shape analysis. In *2015 E-Health and Bioengineering Conference (EHB)*, pages 1–4. IEEE, 2015.

B. B. Practice. Doença do coronavírus 2019 (covid-19), 2020. URL `https://bestpractice.bmj.com/topics/pt-br/3000168`.

L. T. Raineri. Seleção de atributos baseada em algoritmos genéticos para o problema de predição de fraturas vertebrais por compressão, 2018.

E. Ribeiro, M. Nogueira-Barbosa, R. Rangayyan, and P. Azevedo-Marques. Detection of vertebral compression fractures in lateral lumbar x-ray images. In *Proc. XXIII Congresso Brasileiro em Engenharia Biomedica-XXIII CBEB*, pages 1136–1139, 2012.

M. K. Santos, J. R. Ferreira Júnior, D. T. Wada, A. P. M. Tenório, M. H. N. Barbosa, and P. M. d. A. Marques. Inteligência artificial, aprendizado de máquina, diagnóstico auxiliado por computador e radiômica: avanços da imagem rumo à medicina de precisão. *Radiologia Brasileira*, 52(6):387–396, 2019.

F. L. Seixas and D. C. M. Saade. *Diagnóstico Auxiliado por Computador*. 2005.

D. Shen, G. Wu, and H.-I. Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017. doi: 10.1146/annurev-bioeng-071516-044442. URL `https://doi.org/10.1146/annurev-bioeng-071516-044442`. PMID: 28301734.

G. S. Taberner, J. Natour, and A. da Rocha Fernandes. Contribuição da tomografia computadorizada e da ressonância magnética na diferenciação entre fraturas agudas benignas e malignas da coluna vertebral. *Revista Brasileira de Reumatologia*, 47(1), 2007.

J. Tehranzadeh and C. Tao. Advances in mr imaging of vertebral colapse. *Seminars in Ultrasound, CT, and MR*, 25(6):440–461, 2004.

J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

B. Xue, M. Zhang, W. N. Browne, and X. Yao. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*, 20 (4):606–626, Aug 2016. ISSN 1089-778X. doi: 10.1109/TEVC.2015.2504420.

Apêndices

A

Análise da base de dados

Esta seção apresenta uma análise dos dados clínicos extraídos do prontuário eletrônico dos pacientes. A Figura 15 apresenta a distribuição por sexo e corpo vertebral de todos os pacientes presentes na base de dados, enquanto as Figuras 16, 17 e 18 apresentam a distribuição dos dados por classe. Já a Tabela 30 e a Figura 19 apresentam a distribuição da idade dos pacientes presentes na base de dados.

De acordo com a base de dados utilizada neste trabalho, é possível observar que no sexo masculino ocorre uma maior incidência de fraturas malignas (Figura 18), enquanto no sexo feminino ocorre uma maior incidência de fraturas benignas (Figura 17). Como os corpos vertebrais normais foram retiradas de exames com fraturas benignas, era esperado que a maioria dos corpos vertebrais normais seriam provenientes do sexo feminino 16).

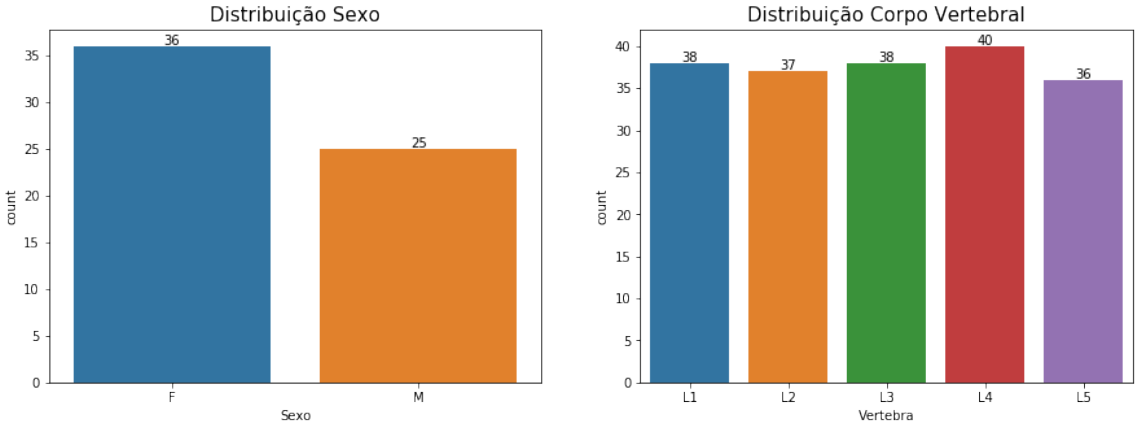


Figura 15 – Distribuição de todos os pacientes presentes na base de dados. a) Distribuição por sexo. b) Distribuição por corpo vertebral.

Tabela 30 – Idade mínima, média e máxima por grupo de pacientes.

Idade	**Base de dados**	**Classe Benigna**	**Classe Maligna**	**Classe Normal**	**Sexo Fem**	**Sexo Masc**
Min	27	45	27	53	33	27
Mean	63	68	57	69	63	62
Max	88	88	79	84	88	84

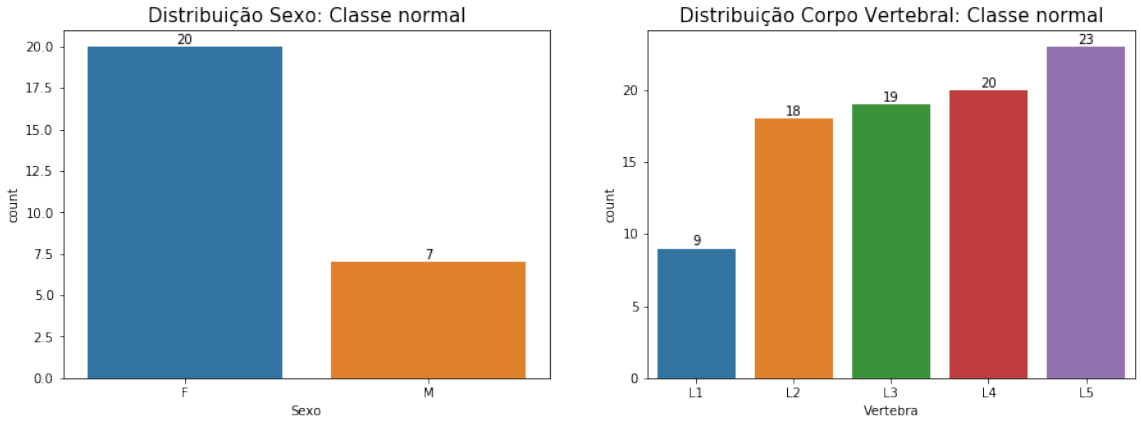


Figura 16 – Distribuição de todos os pacientes presentes na base de dados diagnosticados com corpos vertebrais normais. a) Distribuição por sexo. b) Distribuição por corpo vertebral.

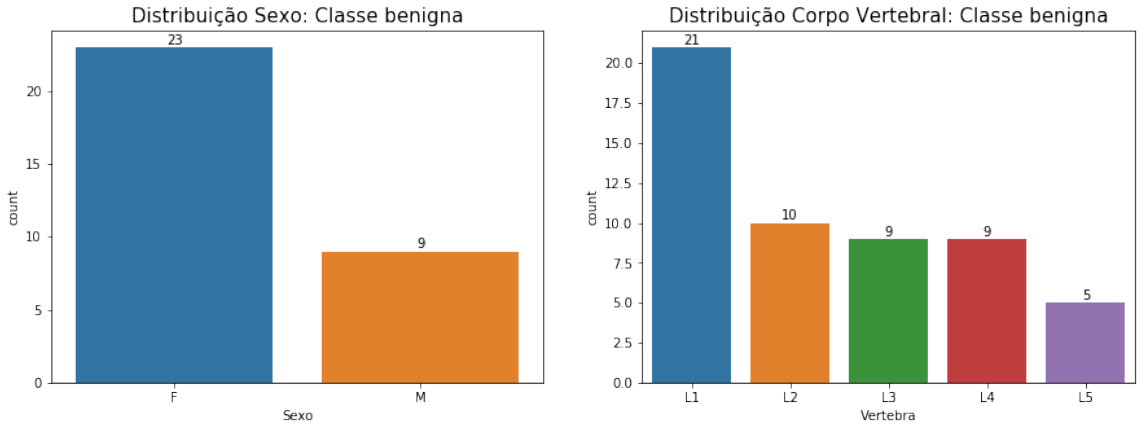


Figura 17 – Distribuição de todos os pacientes presentes na base de dados diagnosticados com fratura vertebral benigna. a) Distribuição por sexo. b) Distribuição por corpo vertebral.

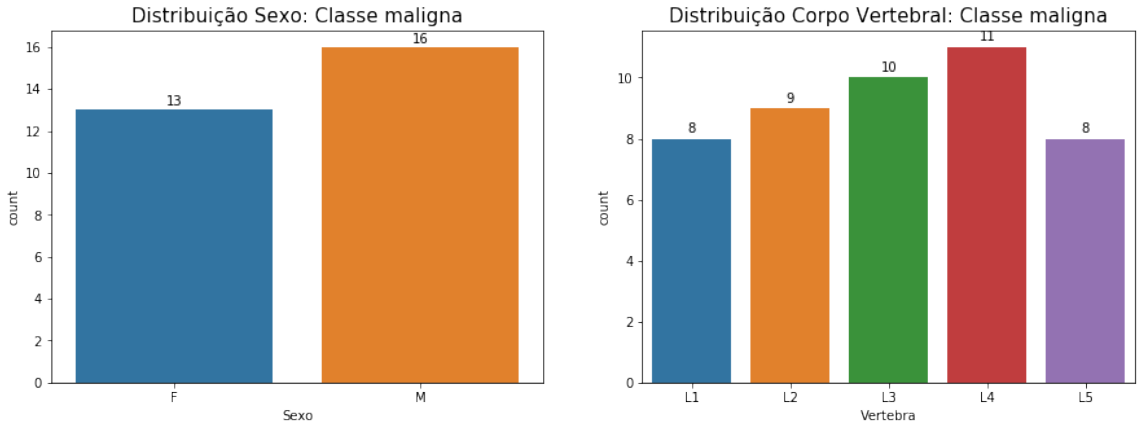


Figura 18 – Distribuição de todos os pacientes presentes na base de dados diagnosticados com fratura vertebral maligna. a) Distribuição por sexo. b) Distribuição por corpo vertebral.

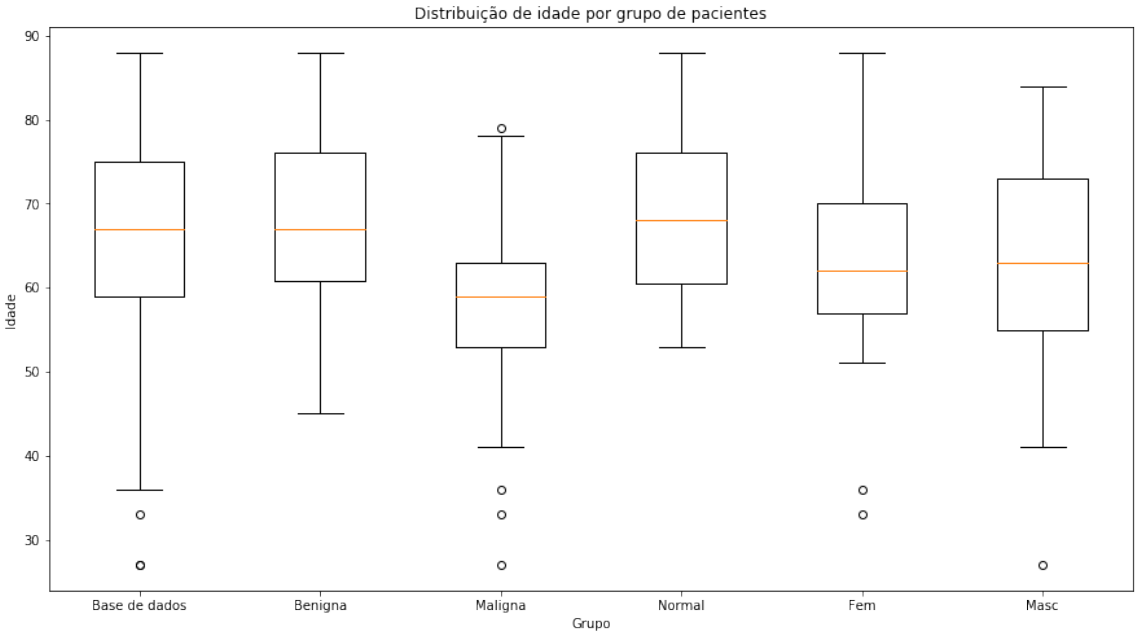


Figura 19 – Boxplot da distribuição de idade por grupo de pacientes.

Em relação aos corpos vertebrais, ocorre uma maior incidência de fratura benigna no corpo vertebral L1 (Figura 17), enquanto lesões malignas ocorre com maior incidência no corpo vertebral L2, L3, L4 (Figura 18). Já o corpo vertebral L5 foi o corpo vertebral que apresentou a menor incidência de lesão (Figura 16).

O corpo vertebral L1 é o primeiro corpo vertebral superior da região lombar, enquanto o corpo vertebral L5 é último corpo vertebral, ou seja, o corpo mais inferior.

De acordo com a base de dados utilizada neste trabalho, é possível concluir que normalmente a fratura benigna ocorre no primeiro corpo vertebral da região lombar, enquanto uma fratura maligna ocorre nos corpos vertebrais centrais, e o corpo vertebral inferior normalmente é corpo que sobre menos lesões.

Também é possível concluir que a fratura maligna ocorreu com maior frequência em pacientes mais novos em relação a fratura benignas. As mulheres sofrem, com maior incidência, fraturas benignas principalmente no corpo vertebral L1 enquanto os homens sofrem, com maior incidência, fraturas malignas principalmente no corpo vertebral L3.

B

Tempo de execução dos algoritmos

A Tabela 31 apresenta o tempo médio gasto em cada execução para cada um dos modelos estudados. [1] Os experimentos foram realizados em 2 hardwares: i) computador com processador Intel i5-8400 (com 9 MB Cache e 4.00 GHz), 16 GB de memória RAM e placa de video GTX 1060 (6 GB de memória dedicada); ii) um servidor com 2 processadores Intel Xeon E5-2620 v2 (com 15 MB Cache e 2.10 GHz) e 32 GB de memória RAM. [2]

Tabela 31 – Tempo médio por execução para cada um dos modelos treinados neste trabalho.

Modelo	**Cross-validation**	**Teste**	**AG**	**Hardware utilizado**
MLP	1 min	6 min	-	Hardware ii
CNN otimizada manualmente	6 min	8 min	-	Hardware i
CNN otimizada manualmente utilizando aumento de dados	13 min	13 min	-	Hardware i
CNN Pré-treinada	50 min	106 min	-	Hardware ii
Modelo Híbrido utilizando CNN pré-treinada	53 min	275 min	-	Hardware ii
Modelo Híbrido otimizado pelo AG	3 min	4 min	85 h	Hardware ii

[1] Durante a fase de teste foram geradas a Curva ROC, Matriz de Confusão, Tabela de resultados do conjunto de testes e foram calculadas as outras métricas utilizadas para avaliação dos modelos. Por esse motivo, o tempo de execução do teste foi maior do que o tempo de treinamento em alguns casos.

[2] O servidor utilizado é do Departamento de Computação e Matemática - USP e não foi utilizado exclusivamente para o treinamento dos modelos, sendo utilizado por outros usuários em paralelo.

C

Arquitetura do melhor modelo

Esta seção apresenta a arquitetura e os parâmetros de cada camada do melhor modelo encontrado em cada abordagem estudada. Todos os modelos foram treinados utilizando a função de métrica *categorical_accuracy* e a função de loss *categorical_crossentropy*. As CNNs pré-treinadas foram treinada com imagens RGB, por esse motivo os modelos que utilizaram a *VGG16* foram treinados com imagens de 3 canais.

As Tabelas 32 - 36 apresentam as camadas e seus respectivos parâmetros para todos os modelos estudados.

Tabela 32 – Parâmetros da arquitetura do modelo de MLP utilizado. O modelo foi treinado por 100 épocas, utilizando o otimizador Adam com lr=0.001 e os dados foram padronizados utilizando o Standard scale. Consulte a Figura 5(b).

Camada	**Parâmetros**
Dense	units=113, input_dim=113
Dense	units=58
Dropout	rate=0.2
Dense	units=29
Dropout	rate=0.2
Dense	units=3

Tabela 33 – Parâmetros da arquitetura do melhor modelo encontrado para a CNN otimizada manualmente. O modelo foi treinado por 50 épocas, utilizando o otimizador Adam com lr=0.001. Consulte a Figura 5(a).

Camada	**Parâmetros**
Conv2D	filters=128, kernel_size=3x3, input_shape=(125, 91, 1)
BatchNormalization	-
MaxPooling2D	pool_size=2x2
Conv2D	filters=64, kernel_size=3x3
BatchNormalization	-
MaxPooling2D	pool_size=2x2
Conv2D	filters=64, kernel_size=3x3
BatchNormalization	-
MaxPooling2D	pool_size=2x2
Flatten	-
Dense	units=256
Dropout	rate=0.4
Dense	units=64
Dropout	rate=0.4
Dense	units=3

Tabela 34 – Parâmetros da arquitetura do modelo de CNN utilizando *VGG16*. O modelo foi treinado por 100 épocas, utilizando o otimizador Adam com lr=0.0001.

Camada	**Parâmetros**
VGG16	input_shape=(100, 100, 3)
Flatten	-
Dense	units=512
Dropout	rate=0.2
Dense	units=256
Dropout	rate=0.2
Dense	units=128
Dropout	rate=0.2
Dense	units=3

Tabela 35 – Parâmetros da arquitetura do modelo híbrido utilizando *VGG16*. O modelo foi treinado por 300 épocas, utilizando o otimizador Adam com lr=0.0001 e os dados foram padronizados utilizando o Standard scale. Consulte a Figura 5(c).

Camada	**Parâmetros**
VGG16	input_shape=(100, 100, 3)
Flatten	-
Concatenate Flatten + 113 Features adicionais	-
Dense	units=628
Dropout	rate=0.2
Dense	units=314
Dropout	rate=0.2
Dense	units=157
Dropout	rate=0.2
Dense	units=3

Tabela 36 – Parâmetros da arquitetura do melhor modelo híbrido encontrado pelo AG. O modelo foi treinado por 150 épocas, utilizando o otimizador SGD com lr=0.05 e os dados foram normalizados utilizando o Standard scale. Consulte a Figura 5(d).

Camada	**Parâmetros**
Conv2D	filters=128, kernel_size=4x4, padding='valid', input_shape=(150, 150, 1), activation='tanh'
BatchNormalization	-
MaxPooling2D	pool_size=5x5
Dropout	rate=0.3
Conv2D	filters=512, kernel_size=3x3, padding='same', activation='tanh'
Dropout	rate=0.1
Conv2D	filters=64, kernel_size=6x6, padding='valid', activation='tanh'
BatchNormalization	-
MaxPooling2D	pool_size=3x3
Dropout	rate=0.1
Flatten	-
Concatenate Flatten + 53 Features adicionais	-
Dense	units=200
Dropout	rate=0.2
Dense	units=200
Dropout	rate=0.3
Dense	units=100
Dropout	rate=0.1
Dense	units=3

D

Atributos Selecionados pelo Algoritmo Genético

A Tabela 37 apresenta quais foram os atributos selecionados pela melhor solução encontrada em cada uma das 5 execuções do Algoritmo Genético para otimização dos parâmetros do modelo híbrido proposto.

Tabela 37 – Atributos selecionados pela melhor solução encontrada em cada uma das 5 execuções do Modelo Híbrido otimizado pelo Algoritmo Genético. Eram passíveis de seleção 113 atributos.

Feature	**Exec 1**	**Exec 2**	**Exec 3**	**Exec 4**	**Exec 5**
idade	X	X	X	X	X
sexo			X		
Max	X	X	X	X	
Mean	X		X	X	
Min		X		X	X
STD			X	X	X
Vertebra	X		X	X	
firstorder_10Percentile	X		X		X
firstorder_90Percentile	X		X		X
firstorder_Energy	X	X	X	X	
firstorder_Entropy		X	X		X
firstorder_InterquartileRange	X	X	X	X	
firstorder_Kurtosis			X	X	
firstorder_Maximum			X	X	
firstorder_Mean		X	X	X	
firstorder_Median				X	X
firstorder_MeanAbsoluteDeviation	X		X		
firstorder_Minimum	X	X	X		
firstorder_Range	X		X	X	
firstorder_RobustMeanAbsoluteDeviation	X	X	X	X	X
firstorder_RootMeanSquared			X		
firstorder_StandardDeviation		X	X		X

Continuação na próxima página

Tabela 37 – *Continuação da tabela da pagina anterior*

Feature	Exec 1	Exec 2	Exec 3	Exec 4	Exec 5
firstorder_Skewness	X	X	X		X
firstorder_TotalEnergy	X			X	X
firstorder_Uniformity	X	X			
firstorder_Variance	X		X		
glcm_Autocorrelation		X	X	X	
glcm_ClusterProminence				X	
glcm_ClusterTendency					X
glcm_ClusterShade	X	X		X	
glcm_Contrast	X	X			X
glcm_Correlation	X	X		X	X
glcm_DifferenceAverage			X	X	
glcm_DifferenceEntropy	X	X		X	
glcm_DifferenceVariance	X	X		X	
glcm_Id			X		
glcm_Idm		X	X		X
glcm_Idn	X	X			
glcm_Idmn			X		
glcm_Imc1		X	X		X
glcm_Imc2	X				
glcm_InverseVariance		X	X		X
glcm_JointAverage					X
glcm_JointEnergy	X	X	X	X	
glcm_JointEntropy					X
glcm_MaximumProbability	X		X		
glcm_SumAverage	X			X	
glcm_SumEntropy		X			X
glcm_SumSquares		X	X		X
gldm_DependenceEntropy		X	X		
gldm_DependenceNonUniformity		X			X
gldm_DependenceNonUniformityNormalized				X	X
gldm_DependenceVariance	X	X		X	X
gldm_GrayLevelNonUniformity	X		X		X
gldm_GrayLevelVariance		X	X		X
gldm_HighGrayLevelEmphasis	X	X			
gldm_LargeDependenceEmphasis		X		X	
gldm_LargeDependenceHighGrayLevelEmphasis	X	X			
gldm_LargeDependenceLowGrayLevelEmphasis	X				
gldm_LowGrayLevelEmphasis	X				X
gldm_SmallDependenceEmphasis		X		X	
gldm_SmallDependenceHighGrayLevelEmphasis	X			X	X
glrlm_GrayLevelNonUniformity			X	X	
glrlm_GrayLevelNonUniformityNormalized				X	X
glrlm_GrayLevelVariance	X				X
glrlm_HighGrayLevelRunEmphasis	X	X		X	

Continuação na próxima página

Tabela 37 – *Continuação da tabela da pagina anterior*

Feature	Exec 1	Exec 2	Exec 3	Exec 4	Exec 5
glrlm_LongRunEmphasis	X	X		X	
glrlm_LongRunHighGrayLevelEmphasis	X	X		X	
glrlm_LowGrayLevelRunEmphasis	X				
glrlm_LongRunLowGrayLevelEmphasis		X		X	
glrlm_RunEntropy	X		X	X	X
glrlm_RunLengthNonUniformity	X				X
glrlm_RunLengthNonUniformityNormalized	X	X		X	
glrlm_RunPercentage X			X		
glrlm_RunVariance	X				
glrlm_ShortRunEmphasis	X	X		X	X
glrlm_ShortRunHighGrayLevelEmphasis	X	X	X		
glrlm_ShortRunLowGrayLevelEmphasis		X	X		
glszm_GrayLevelNonUniformity				X	X
glszm_GrayLevelNonUniformityNormalized					X
glszm_GrayLevelVariance				X	X
glszm_HighGrayLevelZoneEmphasis	X	X	X		X
glszm_LargeAreaEmphasis	X	X		X	X
glszm_LargeAreaHighGrayLevelEmphasis		X	X		
glszm_LargeAreaLowGrayLevelEmphasis	X	X	X		X
glszm_LowGrayLevelZoneEmphasis		X			X
glszm_SizeZoneNonUniformity	X	X			X
glszm_SizeZoneNonUniformityNormalized	X			X	
glszm_SmallAreaHighGrayLevelEmphasis	X			X	X
glszm_SmallAreaLowGrayLevelEmphasis	X			X	
glszm_ZoneEntropy	X		X	X	X
glszm_ZonePercentage				X	X
glszm_ZoneVariance	X			X	X
ngtdm_Busyness			X		X
ngtdm_Coarseness	X		X	X	
ngtdm_Contrast				X	
ngtdm_Strength	X				X
shape_ElongationX					
shape_Flatness X			X		X
shape_LeastAxis	X			X	

Continuação na próxima página

Tabela 37 – *Continuação da tabela da pagina anterior*

Feature	**Exec 1**	**Exec 2**	**Exec 3**	**Exec 4**	**Exec 5**
shape_MajorAxis		X	X		X
shape_Maximum2DDiameterColumn	X	X		X	
shape_Maximum2DDiameterRow	X	X		X	X
shape_Maximum2DDiameterSlice		X			X
shape_Maximum3DDiameter	X		X		X
shape_MinorAxis	X		X		
shape_Compactness1		X	X	X	X
shape_Compactness2			X		X
shape_Sphericity	X			X	X
shape_SphericalDisproportion	X				X
shape_SurfaceArea		X			
shape_SurfaceVolumeRatio	X				
shape_Volume		X		X	
Total atributos selecionadas	65	56	50	55	56

E

NNGA

A biblioteca produzida durante esse mestrado recebeu o nome de **Neural Network optimized by Genetic Algorithms - (NNGA)**. Com ela é possível treinar modelos de classificação de forma fácil, utilizando arquitetura de uma MLP, CNN pré-treinada e arquitetura Híbrida proposta neste projeto. É possível ativar/desativar seleção de atributos e a otimização da arquitetura do modelo por meio de Algoritmo Genético. Um modelo de segmentação também está disponível. A biblioteca estará disponível atráves do link `https://github.com/rafaelsdellama/nnga` após o termino do projeto e a publicação dos resultados.

F

COVID-19

Neste momento estamos passando pela pandemia de COVID-19, uma doença respiratória aguda causada pelo *coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2)* [Bastos, 2020]. A doença foi identificada pela primeira vez na China em 1 de dezembro de 2019, mas o primeiro caso foi reportado em 31 de dezembro 2020. Em 11 de março de 2020, a Organização Mundial da Saúde declarou pandemia [Organization, 2020]. A gravidade dos sintomas varia, desde sintomas ligeiros semelhantes à constipação até pneumonia viral grave com insuficiência respiratória potencialmente fatal [Practice, 2020], e sendo a maioria dos pacientes submetidos a radiografia de tórax na avaliação inicial [Barreto, 2020].

A metodologia proposta neste trabalho foi aplicada a uma base de dados contendo raio-x de pacientes diagnosticados com COVID-19, com o objetivo de auxiliar no diagnóstico de pneumonia e na distinção entre pneumonia causada pela COVID-19 e pneumonia não resultante de COVID-19.

A base de COVID-19 foi obtida através de um conjunto de dados público de imagens de raios-X de pacientes positivos ou suspeitos de COVID-19 ou outras pneumonias virais e bacterianas (MERS , SARS e ARDS). Novos dados são adicionados ao repositório constantemente sendo coletados de fontes públicas, bem como por meio de coleta indireta de hospitais e médicos. Todas as imagens e dados são divulgados publicamente em `https://github.com/ieee8023/covid-chestxray-dataset`. Este projeto foi aprovado pelo Comitê de Ética da Universidade de Montreal. No momento da aquisição das imagens para este estudo, o repositório contava com 183 imagens de raio-X de pacientes diagnosticados com COVID-19.

As imagens de raio-x das classes Normal e Pneumonia (não provocada por COVID-19) foram retiradas do *Desafio de detecção de pneumonia da RSNA* [1].

O dataset de treinamento foi balanceado baseado na classe minoritária (COVID-19), resultando no conjunto da Tabela 38. Posteriormente os atributos radiômicos foram

[1] `https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/`

extraidos utilizando a biblioteca *pyRadiomics* e padronizados utilizando o método Standard scale. As máscaras dos pulmões foram geradas utilizando-se um modelo de segmentação treinado utilizando o dataset *Chest Xray Masks and Labels* [2]. Nenhum preprocessamento foi aplicado.

Tabela 38 – Distribuição images de raio-x.

Conjunto	**Normal**	**Pneumonia**	**COVID-19**	**Total**
Treino	152	152	152	456
Teste	100	100	31	231

F.1 MLP

A acurácia balanceada média obtida no conjunto de treinamento (usando a validação cruzada) foi de 69.98%, enquanto o melhor resultado para o conjunto de testes foi de 71.47%. A Tabela 39 mostra a matriz de confusão obtida pelo modelo, enquanto a Tabela 40 apresenta outras métricas usadas para avaliar o classificador. A Figura 20 mostra as curvas ROC obtidas pelo modelo. Todos esses resultados foram calculados para o conjunto de testes.

Tabela 39 – Matriz de confusão obtida pela MLP.

		Classe obtida		
	Classificação	**COVID-19**	**Normal**	**Pneumonia**
Classe real	COVID-19	24	3	4
	Normal	19	65	16
	Pneumonia	13	15	72

Tabela 40 – Resultados de métricas utilizadas para avaliação da MLP.

	Precisão	**Recall**	**F1-score**	**Support**	**Especificidade**	**Sensibilidade**
COVID-19	0.43	0.77	0.55	31	0.84	0.77
Normal	0.78	0.65	0.71	100	0.86	0.65
Pneumonia	0.78	0.72	0.75	100	0.85	0.72

[2] `https://www.kaggle.com/nikhilpandey360/chest-xray-masks-and-labels`

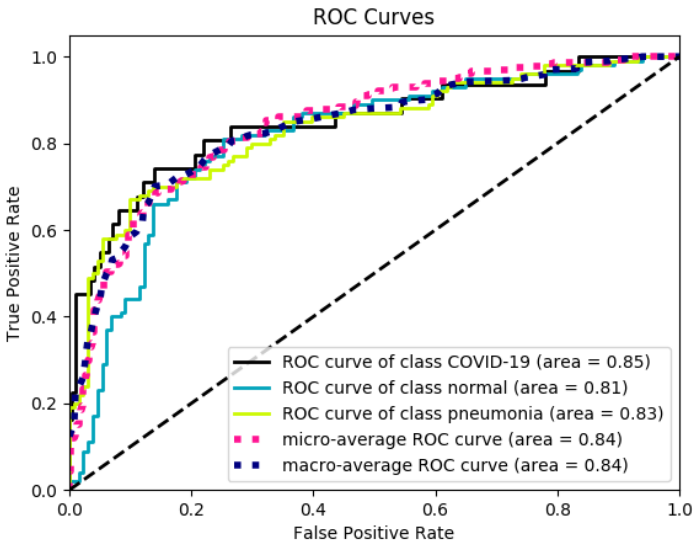


Figura 20 – Curva ROC obtida pela MLP.

F.2 CNN Pré-treinada

A acurácia balanceada média obtida no treinamento (usando a validação cruzada) foi de 87.3%, e no conjunto de imagens de teste foi de 90.1%.

A Tabela 41 mostra a matriz de confusão obtida pelo modelo no conjunto de teste, enquanto a Tabela 42 apresenta outras métricas utilizadas para avaliar o classificador, e finalmente a Figura 21 apresenta a curva ROC obtida pelo modelo.

Tabela 41 – Matriz de confusão obtida pela *VGG16*.

		Classe obtida		
	Classificação	**COVID-19**	**Normal**	**Pneumonia**
Classe real	COVID-19	28	2	1
	Normal	0	88	12
	Pneumonia	3	5	92

Tabela 42 – Resultados de métricas utilizadas para avaliação da *VGG16*.

	Precisão	**Recall**	**F1-score**	**Support**	**Especificidade**	**Sensibilidade**
COVID-19	0.90	0.90	0.90	31	0.98	0.90
Normal	0.93	0.88	0.90	100	0.94	0.88
Pneumonia	0.88	0.92	0.90	100	0.90	0.92

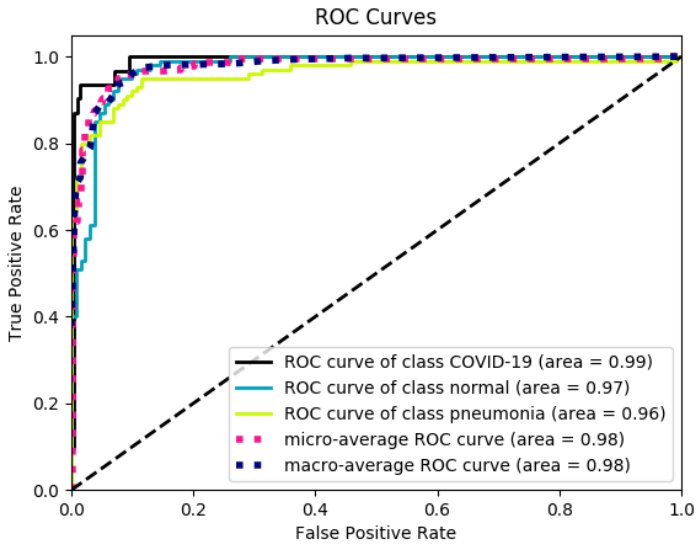


Figura 21 – Curva ROC obtida pela *VGG16*.

F.3 Modelo Híbrido Proposto utilizando CNN pré-treinada

O resultado obtido pelo modelo híbrido proposto utilizando *VGG16* como modelo de CNN pré-treinada são apresentados aqui. O modelo utiliza como entrada as imagens (brutas) e o vetor de características radiômicas, dados clínicos adicionais e atributos extraídos do histograma da imagem são inseridos na primeira camada densa. A acurácia balanceada média obtida no conjunto de treinamento (usando a validação cruzada) foi de 89.0% (resultados não mostrados na tabela), enquanto a acurácia do conjunto de testes foi de 92.18%.

A Tabela 43 apresenta a matriz de confusão, enquanto a Tabela 44 apresenta os resultados de métricas adicionais. A Figura 22 mostra as curvas ROC obtidas pelo modelo.

Tabela 43 – Matriz de confusão obtida pelo Modelo Híbrido.

		Classe obtida		
	Classificação	**COVID-19**	**Normal**	**Pneumonia**
Classe real	COVID-19	29	1	1
	Normal	0	88	12
	Pneumonia	1	4	95

Tabela 44 – Resultados de métricas utilizadas para avaliação do Modelo Híbrido.

	Precisão	**Recall**	**F1-score**	**Support**	**Especificidade**	**Sensibilidade**
COVID-19	0.97	0.94	0.95	31	0.99	0.93
Normal	0.95	0.88	0.91	100	0.96	0.88
Pneumonia	0.88	0.95	0.91	100	0.90	0.95

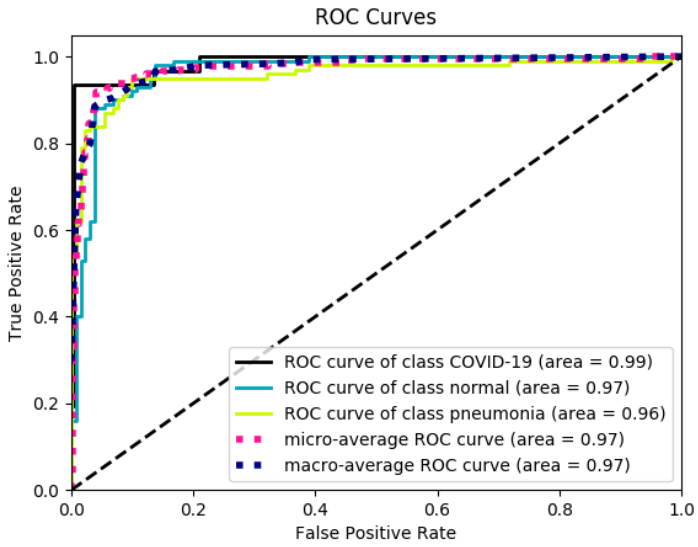


Figura 22 – Curva ROC obtida pelo Modelo Híbrido.

F.4 Modelo Híbrido Proposto otimizado pelo AG

Devido à necessidade de um longo tempo para otimização, o Algoritmo Genético foi executado uma única vez utilizando uma população de tamanho 30, durante 50 gerações. A taxa de crossover utilizada foi de 60% enquanto a taxa de mutação foi de 2%. Foi utilizado o método de elitismo, onde os 2 melhores indivíduos de cada geração foram passados automaticamente para a próxima geração. O método de seleção utilizado foi o Torneio de tamanho 3.

A acurácia balanceada média obtida no treinamento (usando a validação cruzada) foi de 70.56%, e no conjunto de imagens de teste foi de 86.10%.

A Tabela 45 ilustra a matriz de confusão obtida pelo conjunto de teste, na sequência, a Tabela 46 apresenta os resultados de métricas utilizadas para avaliar o classificador, e finalmente a Figura 23 apresenta a curva ROC obtida pelo modelo.

Tabela 45 – Matriz de confusão obtida pelo Modelo Híbrido otimizado pelo AG.

		Classe obtida		
	Classificação	**COVID-19**	**Normal**	**Pneumonia**
Classe real	COVID-19	28	1	2
	Normal	3	85	12
	Pneumonia	4	13	83

Tabela 46 – Resultados de métricas utilizadas para avaliação do Modelo Híbrido otimizado pelo AG.

	Precisão	**Recall**	**F1-score**	**Support**	**Especificidade**	**Sensibilidade**
COVID-19	0.80	0.90	0.85	31	0.96	0.90
Normal	0.86	0.85	0.85	100	0.89	0.85
Pneumonia	0.86	0.83	0.84	100	0.89	0.83

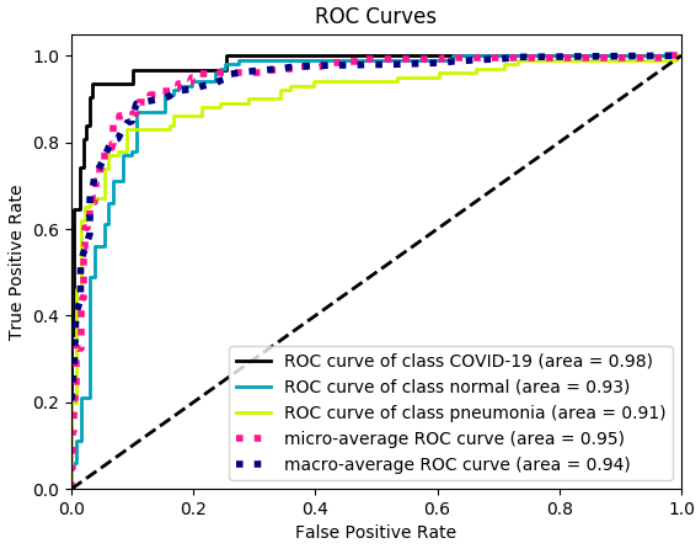


Figura 23 – Curva ROC obtida pelo Modelo Híbrido otimizado pelo AG.

F.5 Comparação dos Resultados

Tabela 47 – Acurácia obtida em cada uma das execuções com 10-fold stratified cross validation para separação das classes Benigna x Maligna x Normal. A letra S indica que o resultado é estatisticamente significante, considerando-se o *Teste t*, se comparado com o modelo referência. Os símbolos + e - significam respectivamente que médias obtidas pelo modelo foi maior ou menor que a média do modelo referência.

Exec	**MLP**	**CNN Pré-treinada**	**Modelo Híbrido utilizando CNN pré-treinada**	**Modelo Híbrido otimizado pelo AG**
1	0.68	0.93	0.93	0.35
2	0.66	0.83	0.89	0.79
3	0.64	0.84	0.93	0.53
4	0.73	0.88	0.88	0.86
5	0.68	0.82	0.95	0.84
6	0.80	0.84	0.88	0.48
7	0.46	0.95	0.84	0.88
8	0.75	0.93	0.88	0.82
9	0.75	0.86	0.88	0.75
10	0.80	0.80	0.77	0.71
Média	0.70 (S-)	0.87	0.89 (+)	0.70 (S-)
std	0.09	0.05	0.05	0.18

A Tabela 48 resume os resultados obtidos pelos modelos que utilizaram diferentes fontes de informação como entrada. A Tabela 49 mostra o(s) modelo(s) com maior(es) sensibilidade e especificidade para cada classe. A Tabela 52 apresenta o tempo de execução para os algoritmos estudados. Os experimentos foram realizados em um servidor com 2 processadores Intel Xeon E5-2620 v2 (com 15 MB Cache e 2.10 GHz) e 32 GB de memória RAM.

Tabela 48 – Resumo das métricas para cada um dos modelos estudados. Em negrito está o maior valor encontrado para cada uma das métricas.

Métrica	**Conjunto**	**MLP**	**CNN Pré-treinada**	**Modelo Híbrido utilizando CNN pré-treinada**	**Modelo Híbrido otimizado pelo AG**
Acurácia balanceada (cross-validation)	Treinamento	69.98%	87.3%	**89%**	70.56%
Acurácia balanceada	Teste	71.47%	90.01%	**92.18%**	86.1%
F1-Score	Teste	67%	90%	**92%**	85%
Especificidade	Teste	85%	94%	**95%**	91%
Sensibilidade	Teste	71%	90%	**92%**	86%
Macro-average ROC	Teste	84%	**98%**	97%	86%

Tabela 49 – Modelo com maior sensibilidade e especificidade para cada classe de FVC.

	Especificidade	**Sensibilidade**
COVID-19	Modelo Híbrido CNN Pré-treinada (0.99)	Modelo Híbrido CNN Pré-treinada (0.93)
Normal	Modelo Híbrido CNN Pré-treinada (0.96)	Modelo Híbrido CNN Pré-treinada (0.88)
Pneumonia	Modelo Híbrido CNN Pré-treinada (0.9)	Modelo Híbrido CNN Pré-treinada (0.95)

A Tabela 51 compara os resultados obtidos pelo Modelo Hibrído com a COVID-Net [Linda Wang and Wong, 2020]. O dataset utilizado para treinar o Modelo Hibrído é um subset balanceado do dataset utilizado para treinar a COVID-Net [3], e o dataset de teste é o mesmo (COVIDx2). A distribuição do dataset utilizado para treinar a COVID-Net é apresentado na Tabela 50. O Modelo Híbrido conseguiu alcançar resultados muito próximos da COVID-Net utilizando 3.36% do dataset utilizado para treinar a COVID-Net.

Tabela 50 – Distribuição imagens de raio-x utilizadas para treinar a COVID-Net.

Conjunto	**Normal**	**Pneumonia**	**COVID-19**	**Total**
Treino	7966	5451	152	13569
Teste	100	100	31	231

[3] Iniciativa de código aberto, em fase de pesquisa, que está sendo muito utilizado como modelos de referência. A COVID-Net está disponível em: `https://github.com/lindawangg/COVID-Net`

Tabela 51 – Comparação modelo híbrido proposto x COVID-Net.

	Sensibilidade			**Precisão**		
Modelo	**Normal**	**Pneumonia**	**COVID-19**	**Normal**	**Pneumonia**	**COVID-19**
COVIDNet-CXR Small	97.0%	**90.0%**	87.1%	89.8%	94.7%	**96.4%**
COVIDNet-CXR Large	**99.0%**	89.0%	**96.8%**	91.7%	**98.9%**	90.9%
Modelo Híbrido utilizando CNN Pré-treinada	93.0%	88.0%	95.0%	**97.0%**	95.0%	88.0%

Tabela 52 – Tempo médio por execução para cada um dos modelos treinados neste trabalho.

Modelo	**Cross-validation**	**Teste**	**AG**
MLP	3 min	18 min	-
CNN Pré-treinada	5 h	38 h	-
Modelo Híbrido utilizando CNN pré-treinada	3 h	53 h	-
Modelo Híbrido otimizado pelo AG	48 min	72 min	124 h