Convolutional Neural Networks in Breast Cancer Diagnosis - A Comparative Study with CBIS-DDSM Data

Vitor Negromonte
Departamento de Estatística
Universidade Federal de Pernambuco
Recife, Brazil
vnco@cin.ufpe.br

Eduardo Guimarães Medeiros
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
egm3@cin.ufpe.br

I. Introdução

Com a crescente popularização de modelos de inteligência artificial, urge a necessidade de integrar tais tecnologias em problemas reais, para que a sociedade possa usufruir de avanços que muitas vezes ficam restritos a problemas acadêmicos. Isto posto, nosso estudo propõe uma aplicação direta para modelos de redes neurais convolucionais quanto à aplicações médicas. O volume de imagens analisadas por um médico em um plantão pode ser absurdo, o que, portanto, com o passar do tempo, geraria um cansaço, que pode vir a impactar diretamente a qualidade das análises realizadas pelo profissional e, consequentemente, pode prejudicar um correto diagnóstico do problema enfrentado pelo paciente. Com isso em mente, decidimos por desenvolver uma pesquisa comparativo quanto ao desempenho de modelos CNN no diagnóstico de câncer de mama, de maneira a contribuir para o avanço da análise médica e melhorando a qualidade de vida do paciente, com análises mais estáveis, e do profissional, simplificando o trabalho que ele virá a realizar no seu dia-a-dia.

II. Justificativa

No atual cenário de avanço tecnológico, a integração de inteligência artificial em problemas do mundo real é crucial para garantir benefícios tangíveis à sociedade, especialmente na área da saúde.

A crescente popularização de modelos de inteligência artificial, como as redes neurais convolucionais (CNNs), oferece oportunidades significativas para melhorar a eficiência e a precisão das análises médicas.

A sobrecarga de trabalho enfrentada pelos profissionais da saúde, especialmente em ambientes de plantão, é uma preocupação relevante. O volume de imagens a serem analisadas, como no caso do diagnóstico de câncer de mama, pode ser imenso, levando a fadiga e potencialmente afetando a qualidade das avaliações médicas. Esta fadiga pode, por sua vez, comprometer a precisão do diagnóstico, prejudicando o tratamento adequado do paciente.

Diante desse contexto, a presente pesquisa se propõe a preencher uma lacuna significativa na aplicação de modelos de CNNs em problemas médicos reais. Mais especificamente, nosso estudo visa investigar a eficácia desses modelos no diagnóstico de câncer de mama, utilizando o CBIS-DDSM como fonte de dados. A escolha deste tema é motivada pela necessidade de encontrar soluções que otimizem o processo de diagnóstico, promovendo uma análise mais rápida, precisa e, consequentemente, um tratamento mais eficaz para os pacientes.

III. Metodologia

A. Conjunto de Dados

O conjunto de dados utilizado para as análises é o *CBIS-DDSM: Breast Cancer Image Dataset*, disponibilizado pelo usuário "AWSAF" no website Kaggle. O autor concede permissão para utilizar os dados para fins de estudo. Este conjunto compreende 10.239 imagens de mamografias, incluindo casos normais, benignos e malignos. O dataset é uma derivação do conjunto *Digital Database for Screening Mammography (DDSM)*, criado em 1997 por pesquisadores da Universidade do Sul da Flórida.

A versão utilizada do conjunto de dados foi modificada por pesquisadores do grupo de Ciência de Dados Biomédicas da Escola de Medicina da Universidade de Stanford.

1) Descrição do Conjunto: O conjunto de dados consiste em mamogramas e mapas de Regiões de Interesse (ROI), categorizados em um conjunto maior de calcificações ou nódulos. As seguintes características estão presentes:

- *Categoria de densidade* - a categoria de densidade da calcificação ou nódulo encontrado, variando de 1 a 4.
- *Número de anormalidades na imagem* - alguns casos podem conter mais de uma anomalia.
- *Forma da massa*
- *Margem da massa*
- *Tipo de calcificação*
- *Distribuição de calcificação*
- *Patologia* - benigno, maligno e benigno inconclusivo (casos onde uma anomalia é detectada mas não pode

ser conclusivamente caracterizada como não cancerosa ou cancerosa).
- *Classificação de sutileza*

B. Tratamento dos dados

Para preparamos o conjunto de dados para o treinamento, necessitávamos realizar algumas limpezas no dataset:

- Remoção dos valores nulos ou não respondidos (todas as observações que encontramos estavam entre o tipo de anormalidade e/ou a distribuição), para tal, utilizamos o método de preenchimento retroativo (BFILL) disponível no Sci-Kit Learn, que propaga o próximo valor observado para trás até encontrar a última observação válida no conjunto.
- Redimensionamento das imagens, mesmo considerando que os modelos tenham sido treinados com imagens de 224x224, decidimos por diminuir a imagem para 50x50, visando melhorar o desempenho do treinamento dos modelos.
- Normalização das imagens.
- Unificamos as duas classes benignas (benignas e benignas inconclusivas) para uma única classe, simplificando o treinamento e mantendo o problema binário. Cogitamos treinar os modelos apenas com os dados de benign e maligno, sem a unificação da outra classe, no entanto, percebemos que teríamos poucas imagens disponíveis.

1) Data augmentation: Para melhoramos o desempenho dos modelos, notamos a necessidade de aumentar a variabilidade das imagens, por tanto, utilizamos o método ImageDataGenerator disponível no TensorFlow. Realizamos as seguintes modificações:

- *rotation_range = 20*: Permite rotacionar aleatoriamente a imagem no intervalo de -20 a 20 graus.
- *width_shift_range = 0.2*: Permite deslocamento horizontal aleatório da imagem, variando em até 20% da largura total da imagem.
- *height_shift_range = 0.2*: Permite deslocamento vertical aleatório da imagem, variando em até 20% da altura total da imagem.
- *shear_range = 0.2*: Aplica um deslocamento de tesoura aleatório na imagem, variando em até 20%.
- *zoom_range = 0.2*: Permite aplicar um zoom aleatório na imagem, variando em até 20%.
- *horizontal_flip = True*: Realiza espelhamento horizontal aleatório da imagem.

C. Análise Exploratória de Dados

Inicialmente retiramos algumas medidas descritivas, sobretudo, em relação ao tamanho dos nódulos e das calcificações, para compreendermos melhor o que seria avaliado. Em seguida analisamos disposição geral do dados: distribuição de patologia, dos formatos das nódulos em relação às patologias, a densidade da mama em relação as patologias.

Primeiramente, checamos a distribuição dos casos de câncer ou não-câncer, para avaliarmos a distribuição dos dados. Com isso, notamos um certo desbalanceamento do conjunto, com aproximadamente 60% dos casos sendo de não-câncer, o que poderia vir a impactar o desempenho dos modelos (ver Figura 1).

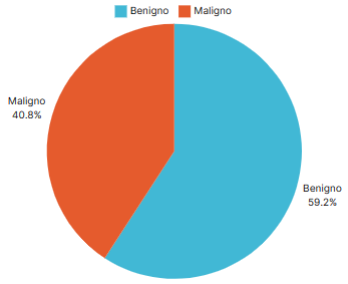

Fig. 1. Distribuição dos casos

Com isso, partimos para a avaliação dos tipo de anormalidades encontradas (ver Figura2, importante ressaltar que a anormalidade não necessariamente indica a presença de um tumor maligno. É importante ressaltar que as calcificações possuem diferentes categorias quanto a sua distribuição e estes dados estavam disponíveis no dataset, no entanto, não vimos pertinência de utilizarmos.

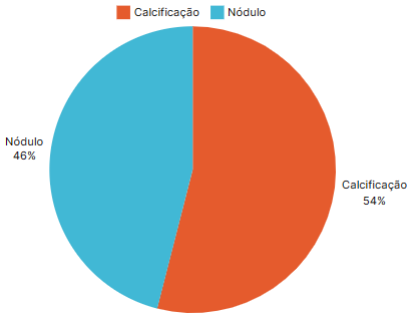

Fig. 2. Distribuição das anormalidades

No conjunto haviam 3 tipos de imagens disponíveis: mamograma completo, imagens cortadas e máscaras de ROI (Region of Interest). Precisamos avaliar a distribuição das imagens para garantirmos a escolha ideal para que os modelos possam performar bem. Portanto, ao avaliarmos a distribuição (ver Figura 3) das imagens disponíveis, decidimos por utilizarmos as imagens cortadas.

D. Modelos

Utilizamos os modelos pré-treinados na ImageNet [1] disponíveis na biblioteca padrão do Keras e do TensorFlow, fazendo apenas o Transfer Learning dos modelos selecionados para os nosso conjunto de dados. Essa estratégia simplifica o processo de treinamento, sobretudo quando avaliada a quantidade de características extraídas das imagens disponíveis na ImageNet em relação aos nossos dados.

1) VGG16: Projetada por pesquisadores do Visual Geometry Group da Universidade de Oxford, a VGG [2] é uma arquitetura de rede que consiste em 16 camadas de convolução

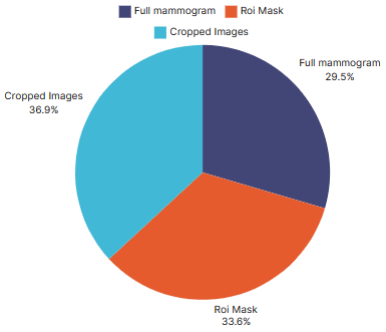

Fig. 3. Distribuição dos tipos de imagem

e *pooling* seguidas por 3 camadas totalmente conectadas e a função de ativação na saída é Softmax. Ela usa filtros de convolução 3x3 com padding e stride 1, seguidas por camadas de Max-pooling 2x2. A grande inovação promovida pela arquitetura VGG é a estrutura de blocos definidas: convolução com padding, função de ativação não linear (ReLU - rectified linear unit) e max-pooling.

2) ResNet50: A ResNet50 [3] é uma rede convolucional da família das redes residuais (ResNet - Residual Networks), desenvolvida por pesquisadores da Microsoft Research. Esta arquitetura inaugura o conceito de redes residuais, que introduzem conexões residuais que recuperam a informação original da imagem a cada bloco residual da rede. Essas conexões permitem que a rede refine as representações, reduzindo a chance de degradação da informação, como normalmente acontece em redes muito profundas.

3) DenseNet121: A DenseNet [4] utiliza blocos de conexões densas, onde cada camada recebe entradas diretamente de todas as camadas anteriores no bloco. Isso promove um fluxo de informações mais eficiente e ajuda a aliviar o problema de degradação de desempenho em redes profundas. A DenseNet121 possui 121 camadas, consistindo em blocos de convolução, operações de pooling e camadas totalmente conectadas para classificação. A configuração da rede a faz ser muito utilizada em classificação de imagens médicas, pois, a mesma dificulta que a rede sofra com Distribution Shift.

Modelo	Parâmetros (milhões)	Profundidade
VGG16	138.4	16
ResNet50	25.6	107
DenseNet121	8.1	242

TABLE I
COMPARAÇÃO DE MODELOS DE REDES NEURAIS CONVOLUCIONAIS COM BASE NO NÚMERO DE PARÂMETROS E PROFUNDIDADE.

E. Metodologia de treinamento

Para um efetivo estudo comparativo, todos os modelos foram treinados com o mesmo número de épocas (5). Na saída de todos os modelos pré treinados foi adicionada um MLP comum a todos com 512 neurônios e função de ativação ReLU, camada de Dropout(0.5) e uma saída com 1 neurônio e função de ativação Softmax. Ademais, padronizamos os otimizadores (Adam), função de perda (entropia cruzada binária) e o mesmo valor para taxa de aprendizado (0.001).

F. Métricas de avaliação

- Matriz de confusão É uma tabela utilizada para avaliar os resultados da saída de um modelo de classificação. A matriz mostra a frequência com que as classificações feitas pelo modelo correspondem ou não à verdadeira classe dos dados (Ver Figura 4).
  - TP: verdadeiros positivos.
  - TN: verdadeiros negativos.
  - FP: falsos positivos.
  - FN: falsos negativos.
- Acurácia É a métrica que nos diz quantos de nossos exemplos foram de fato classificados corretamente, independente da classe avaliada. Definida pela seguinte equação:
  $\frac{TP+TN}{TP+TN+FP+FN}$
  Onde, TP e TN são as obervações corretas (verdadeiros positivos e verdadeios negativos), FN e FP são os faltos negativos e faltos positivos, respectivamente.
- Precisão A métrica é definida pela razão entre a quantidade de exemplos classificados corretamente como positivos e o total de exemplos classificados como positivos:
  $\frac{TP}{TP+FP}$
- Recall É a métrica que lida com a taxa de verdadeiros positivos. É dada pela razão entre os exemplos corretamente positivos pelos exemplos positivos somados aos exemplos falsos negativos:
  $\frac{TP}{TP+FN}$
- F1 Score Dada pela média harmônica entre Recall e Precisão. Um modelo que apresenta um bom F1-score é um modelo capaz tanto de acertar suas predições (precisão alta) quanto de recuperar os exemplos da classe de interesse (recall alto). Portanto, esta métrica tende a ser um resumo melhor da qualidade do modelo. A métrica é definida pela seguinte fórmula:
  $F1 = 2 * \frac{precisao*recall}{precisao+recall}$

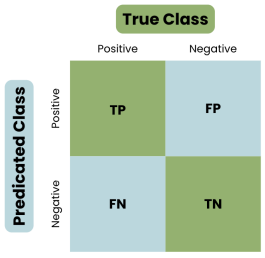

Fig. 4. Exemplo de matriz de confusão

IV. RESULTADOS

Nesta seção, apresentamos os resultados da aplicação de diferentes modelos de redes neurais convolucionais (CNN) no conjunto de dados CBIS-DDSM. Os experimentos foram conduzidos para avaliar o desempenho dos modelos em termos

de métricas de avaliação comuns, como acurácia, precisão, recall, F1-score.

1) Quanto ao tempo de treinamento: As métricas que levamos em consideração para avaliação foram a acurácia, precisão, recall, F1-score, a matriz de confusão e o tempo de treinamento (em segundos).

Modelo	**Duração (min)**	**Duração (segs)**
VGG16	22	1333
ResNet50	23	1390
DenseNet121	25	1480

TABLE II
COMPARAÇÃO DOS MODELOS EM RELAÇÃO AO TEMPO DE TREINAMENTO.

Assim, portanto, nota-se que não houve uma grande discrepância nas duração do treinamento dos modelo, o que pode ser explicado pelos modelos advirem da transferência de aprendizado os dados da ImageNet.

2) Relatório de classificação: Para avaliação das outras métricas, utilizamos o relatório de classificação disponível no pacote Sci-Kit learn, que já retorna as métricas que precisamos para correta avaliação dos modelos.

- VGG16
  - Acurácia no treinamento: 0.9568
  - Acurácia no teste: 0.9563
  - Tempo de treinamento: 1318 segundos

Classe	Precisão	Recall	F1-Score
Sem câncer	0.97	0.98	0.98
Câncer	0.47	0.29	0.36

- ResNet50
  - Acurácia no treinamento: 0.9589
  - Acurácia no teste: 0.96
  - Tempo de treinamento: 1348 segundos

Classe	Precisão	Recall	F1-Score
Sem câncer	0.97	0.99	0.98
Câncer	0.59	0.31	0.41

- DenseNet121
  - Acurácia no treinamento: 0.9573
  - Acurácia no teste: 0.9569
  - Tempo de treinamento: 1436 segundos

Classe	Precisão	Recall	F1-Score
Sem câncer	0.97	0.98	0.97
Câncer	0.41	0.34	0.37

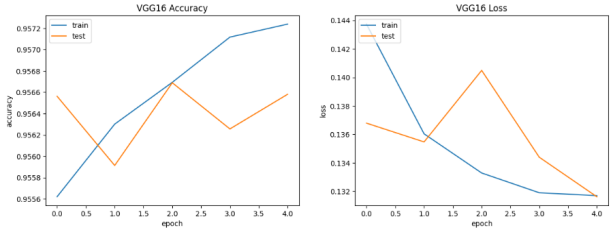

Fig. 5. Gráficos da acurácia e perda da VGG16

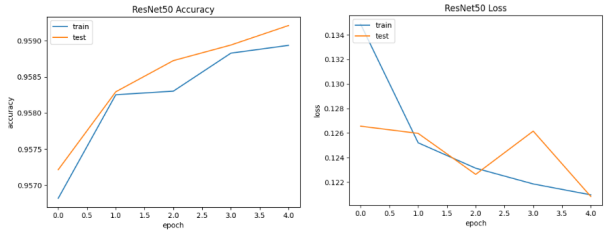

Fig. 6. Gráfico da acurácia e perda da ResNet50

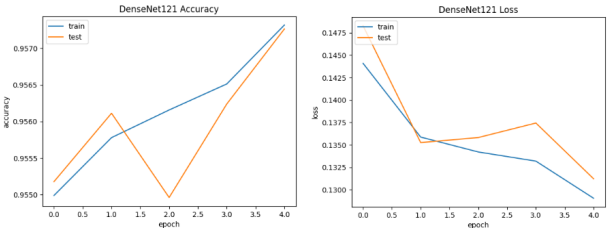

Fig. 7. Gráfico da acurácia e perda da DenseNet121

V. CONCLUSÕES

Avaliando os resultados dos modelos, podemos concluir que houve pouca variação nos resultados dos mesmos, todos obtiveram cerca de 95% de acurácia, no entanto, quanto às demais métricas a ResNet50 demonstrou melhor desempenho que os outros modelos. Ademais, o modelo também apresentou a melhor acurácia, no entanto, decidimos por avaliar os modelos baseados na F1-Score, pois é a métrica mais significativa, que nos permite entender melhor o desempenho real do modelo.

VI. LIMITAÇÕES

Dada a natureza dos dados, percebe-se um desequilíbrio significativo, o que é comum em conjuntos de dados médicos. Esse desequilíbrio representa um desafio adicional para o treinamento eficaz dos modelos. Além disso, seria benéfico reavaliar a abordagem dos estudos, corrigindo o desequilíbrio nos conjuntos de dados. Uma possível solução seria explorar o uso de modelos generativos para gerar dados sintéticos, os quais poderiam ser integrados ao conjunto de dados existente para alcançar um balanceamento mais adequado. Esta abordagem pode ajudar a mitigar os efeitos prejudiciais do desequilíbrio nos dados e melhorar a capacidade dos modelos de aprender com eficácia a partir dos dados disponíveis. Ademais, durante nossa exploração do conjunto de dados, encontramos algumas inconsistências quanto a veracidade de algumas informações contidas no mesmo. Dados de treinamento faltantes, algumas inconsistências quanto as etiquetas dos dados, as máscaras de segmentação não correspondiam as imagens que deveriam apontar, além disso, encontramos informações diferente quanto ao número de pacientes no dataset, na descrição do mesmo, era dito que haviam aproximadamente 6 mil pacientes e nas nossas análises foram encontrados aproximadamente 1.600 pacientes distintos.

REFERENCES

[1] Deng, J. and Dong, W. and Socher, R. and Li, L.-J. and Li, K. and Fei-Fei, L., ImageNet: A Large-Scale Hierarchical Image Database, CVPR09, 2009.

[2] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2015, arXiv: 1409.1556.

[3] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, arXiv: cs.CV, 2015, 1512.03385.

[4] G. Huang and Z. Liu and L. van der Maaten and K. Q. Weinberger, Densely Connected Convolutional Networks, arXiv: cs.CV, 2018, 1608.06993.