Diagnóstico de Câncer de Mama: Melhoria na Classificação entre Lesões Benignas e Malignas com Radiômica

Gean Rocha da Silva Junior[1], **Thaís Gaudencio do Rêgo**[2]

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brasil

gean.junior@academico.ufpb.br, gaudenciothais@gmail.com

Abstract. *Breast cancer is the most frequently diagnosed cancer among women and is the leading cause of cancer-related deaths in this population. Early diagnosis is crucial for reducing mortality associated with this disease. However, there are challenges in detecting breast cancer at early stages, particularly in patients with dense breast tissue. Ultrasound is an effective tool that, in conjunction with mammography, aids in the identification and diagnosis of breast anomalies. Additionally, radiomics aims to extract, process, and classify image characteristics, contributing to the differentiation between malignant and benign lesions. This study proposes the development of a Convolutional Neural Network (CNN) that utilizes not only mammographic images but also their radiomic características , aiming to improve the accuracy in classifying malignant and benign lesions.*

Resumo. *O câncer de mama é o tipo mais frequentemente diagnosticado entre as mulheres e representa a principal causa de morte por câncer nessa população. O diagnóstico precoce é crucial para a redução da mortalidade associada a essa doença. No entanto, a detecção do câncer de mama em estágios iniciais apresenta desafios, especialmente em pacientes com tecido mamário denso. A ultrassonografia é uma ferramenta eficaz que, em conjunto com a mamografia, auxilia na identificação e diagnóstico de anomalias mamárias. Além disso, a radiômica visa extrair, processar e classificar características de imagem, contribuindo para a distinção entre lesões malignas e benignas. Este estudo propõe o desenvolvimento de uma Rede Neural Convolucional (CNN) que utiliza não apenas imagens mamográficas, mas também suas características radiômicas, visando melhorar a precisão na classificação entre lesões malignas e benignas.*

1. Introdução

O câncer de mama se destaca como uma das principais preocupações de saúde global, sendo o câncer mais diagnosticado entre as mulheres em todo o mundo [Bray et al. 2018]. De acordo com as estimativas do GLOBOCAN 2018, este tipo de câncer representa aproximadamente 11,6% de todos os casos novos de câncer e é a principal causa de morte por câncer entre as mulheres, refletindo não apenas sua alta incidência, mas também a gravidade das suas implicações na saúde pública [Bray et al. 2018].

O diagnóstico precoce desempenha um papel fundamental na redução da mortalidade causada pelo câncer de mama. Dentre os métodos existentes, a mamografia é amplamente utilizada na detecção de lesões na mama. No entanto, o desafio de detectar câncer de mama em estágios iniciais permanece significativo, especialmente em pacientes com tecido mamário denso [Khalid et al. 2023].

A ultrassonografia é uma ferramenta eficaz utilizada em conjunto com a mamografia para detectar e diagnosticar anomalias na mama, permitindo diferenciar com alta precisão massas benignas de malignas, o que reduz o número de biópsias desnecessárias. Além disso, a radiômica visa extrair, processar e classificar características de imagem para determinar as características fenotípicas de lesões, ajudando a distinguir lesões malignas de benignas, e pode ser aplicada em qualquer método de imagem, incluindo a ultrassonografia [Fleury and Marcomini 2019].

Dado que o câncer de mama é mais difícil de diagnosticar em tecidos densos, torna-se essencial desenvolver métodos que ofereçam um alto grau de confiabilidade na detecção. Embora modelos existentes possam apresentar um desempenho relativamente bom, sempre existe espaço para melhorias, especialmente em aplicações médicas onde a precisão é crucial. Este estudo propõe a criação de um modelo baseado em rede neural convolucional (*Convolutional neural network* - CNN, em inglês) que integra não apenas as imagens mamográficas, mas também suas características radiômicas (*Radiomics features*, em inglês) correspondentes. A expectativa é que, ao incorporar essas características radiômicas específicas de cada imagem, o modelo possa alcançar uma precisão superior em relação a abordagens que utilizam apenas as imagens. Assim, este trabalho visa contribuir para a melhoria dos diagnósticos de câncer de mama, possibilitando uma classificação mais eficaz entre lesões benignas e malignas.

Para avaliar a eficácia do método proposto, serão utilizadas métricas de desempenho como acurácia, precisão, revocação e *F1-score*. Além disso, a matriz de confusão será empregada como uma ferramenta para visualizar e interpretar detalhadamente o desempenho do modelo, permitindo identificar o número de acertos e erros em cada classe. Essa combinação de métricas fornecerá uma análise abrangente da capacidade do modelo em classificar corretamente as lesões, oferecendo uma avaliação robusta de seu desempenho em um cenário clínico.

2. Trabalhos relacionados

Neste estudo, revisamos uma série de artigos que exploram abordagens metodológicas semelhantes às que adotamos. Essa revisão nos permite extrair *insights* práticos e orientar nosso trabalho com base em estratégias comprovadas. O objetivo é consolidar as melhores práticas identificadas, visando uma aplicação mais eficaz e fundamentada das técnicas selecionadas.

O trabalho de [Fleury and Marcomini 2019] avaliou características radiômicas computáveis do sistema de relatórios e dados de imagem da mama (*Breast Imaging Reporting and Data System* - BI-RADS, em inglês) para classificar massas mamárias, utilizando um banco de dados de 206 lesões (144 benignas e 62 malignas) confirmadas por biópsia. As lesões foram manualmente delineadas em imagens em escala de cinza, e dez características radiômicas foram extraídas. Para a classificação, foram aplicados cinco métodos de aprendizado de máquina (*Machine learning* - ML, em inglês): perceptron multicamada (*Multi-layer perceptron* - MLP, em inglês), árvore de decisão (*Decision tree* - DT, em inglês), análise discriminante linear (Linear discriminant analysis - LDA, em inglês), floresta aleatória (Random forest - RF, em inglês) e máquina de vetores de suporte (Support vector machine - SVM, em inglês), utilizando validação cruzada de 10 vezes e análise ROC para calcular a área sob a curva (AUC). O SVM obteve a melhor performance, com AUC de 0,840, 71,4% de sensibilidade e 76,9% de especificidade, destacando as características de margem e orientação da lesão como as características mais relevantes para a classificação. Esses resultados indicam que o aprendizado de máquina pode ser uma ferramenta útil na distinção entre lesões benignas e malignas em ultrassonografias mamárias, utilizando características radiômicas do BI-RADS.

Apesar das métricas utilizadas serem diferentes da abordagem que adotaremos aqui, esse estudo é importante, pois nos fornece a base para iniciar o trabalho, mostrando que as características radiômicas podem ser úteis em modelos de classificação.

No estudo de [Dong et al. 2024], foram analisados 102 pacientes com 419 imagens de tomossíntese digital, resultando em 100 pacientes e 403 imagens após a exclusão de dados inadequados. As imagens foram processadas para obter mapas de densidade volumétrica, segmentando tecidos densos e adiposos, e características radiômicas foram extraídas, totalizando 1.385 por região de interesse, utilizando a biblioteca `PyRadiomics`. Para garantir a qualidade e eficiência do conjunto de características , foi implementado um processo de seleção em várias etapas: inicialmente, removeram-se características radiômicas com variância zero e alta correlação, seguido pela aplicação da eliminação recursiva de características (*Recursive feature elimination* - RFE, em inglês) com regressão logística (*Logistic regression* - LR, em inglês), que iterativamente testou e excluiu características de menor desempenho preditivo, resultando em um conjunto final de vinte características radiômicas.

Esse estudo fornece uma base metodológica sólida para o nosso trabalho, uma vez que adotaremos técnicas semelhantes de extração e seleção de características , utilizando a mesma biblioteca, a `PyRadiomics`. Inicialmente, extrairemos todas as características radiômicas disponíveis na biblioteca e, posteriormente, aplicaremos o método RFE para selecionar as mais significativas para o nosso modelo. Esse processo nos permitirá reduzir o conjunto de dados às características radiômicas com maior impacto preditivo.

No trabalho de [Khalid et al. 2023], são analisados seis modelos de aprendizado de máquina para a classificação de câncer de mama, utilizando o conjunto de dados *Breast Cancer Wisconsin*. Os modelos avaliados incluem: LR, RF, DT, k vizinhos mais próximos (*K-nearest neighbors* - KNN em inglês), classificador de vetores de suporte (*Support vector classifier* - SVC, em inglês) e linear SVC. As métricas empregadas para avaliar os resultados são as mesmas que pretendemos utilizar neste trabalho: acurácia, revocação, precisão e F1-score. Além disso, será realizada uma análise da matriz de confusão, que também foi utilizada no estudo mencionado. Dessa forma, seguiremos os passos descritos no artigo e aplicaremos esses modelos de classificação e métricas em nossa pesquisa.

3. Metodologia

Nesta seção, apresenta-se a metodologia adotada ao longo do estudo, incluindo a descrição da base de dados, as técnicas de pré-processamento aplicadas, o ambiente de desenvolvimento utilizado, as métricas de avaliação empregadas e os modelos implementados. A metodologia utilizada neste trabalho é dividida em várias etapas sequenciais, cada uma sendo uma peça fundamental para alcançar o modelo final.

3.1. Ambiente de desenvolvimento

Para o desenvolvimento deste estudo, utilizou-se a linguagem de programação Python na versão 3.10.12, juntamente com os seguintes módulos:

- `re`: Este módulo fornece operações de correspondência de expressões regulares, semelhantes às encontradas em Perl.
- `time`: Este módulo disponibiliza várias funções relacionadas à manipulação do tempo.
- `os`: Este módulo implementa funções úteis para o trabalho com nomes de caminho.
- `ast`: Este módulo ajuda a processar árvores de sintaxe abstrata em Python.

Além disso, foram utilizadas as seguintes bibliotecas:

- `lightgbm` - versão 4.5.0: Fornece um classificador baseado em aumento de gradiente, utilizando algoritmos de aprendizado baseados em árvore. Projetada para ser eficiente e escalável, destaca-se por sua velocidade de treinamento, menor consumo de memória, maior precisão e suporte ao aprendizado paralelo e distribuído, além de sua capacidade de lidar com grandes volumes de dados.
- `matplotlib` - versão 3.7.1: Para criar visualizações estáticas, animadas e interativas.
- `numpy` - versão 1.26.4: Fornece um objeto array multidimensional, vários objetos derivados (como arrays mascarados e matrizes) e uma variedade de rotinas para operações rápidas em arrays, incluindo matemática, lógica, manipulação de formas, classificação, seleção, transformadas discretas de Fourier, álgebra linear básica, operações estatísticas básicas, simulação aleatória e muito mais.
- `pandas` - versão 2.2.2: Para análise e manipulação de dados, caracterizada por sua rapidez, flexibilidade e facilidade de uso.
- `PIL` - versão 10.4.0: Oferece amplo suporte a formatos de arquivo, uma representação interna eficiente e recursos de processamento de imagem avançados.

- `scikit-learn` - versão 1.5.2: Biblioteca de ML, que suporta tanto aprendizado supervisionado quanto não supervisionado. Fornece ferramentas para ajuste de modelo, pré-processamento de dados, seleção e avaliação de modelos, entre outros utilitários.
- `seaborn` - versão 0.13.2: Biblioteca de visualização de dados, que fornece uma interface de alto nível para criar gráficos estatísticos atrativos e informativos.
- `pyradiomics` - versão 3.0.1: Para extração de características radiômicas de imagens médicas.
- `PyTorch` - versão 2.5.0+cu121: Para operações com tensores, otimizada para aprendizado profundo (*Deep learning*, em inglês), utilizando GPUs e CPUs.

O hardware utilizado para o treinamento dos modelos e extração das características radiômicas foi a versão paga do Google Colab, especificamente no ambiente de execução selecionado: L4 GPU. Esse ambiente consome um total de 3 unidades computacionais por hora e fornece os seguintes recursos:

- RAM do sitema: 53 GB
- RAM da GPU: 22.5 GB
- Armazenamento: 235.7 GB

3.2. Métricas de avaliação

As métricas de desempenho são essenciais para avaliar a eficácia do método proposto, pois fornecem uma análise objetiva da capacidade do modelo em classificar corretamente. Elas permitem identificar acertos e erros, oferecendo uma compreensão aprofundada do desempenho em um contexto clínico. Essa avaliação é crucial para garantir que o modelo seja confiável e útil na prática, especialmente em situações em que decisões baseadas nas classificações podem impactar diretamente na saúde dos pacientes.

3.2.1. Acurácia

Em geral, a acurácia mede a proporção de previsões corretas sobre o número total de instâncias avaliadas [Hossin and Sulaiman 2015]. E pode ser dada pela seguinte formula:

$$\frac{\text{VP + VN}}{\text{VP +FP + VN + FN}} \tag{1}$$

Onde:

- VP (Verdadeiros Positivos): número de previsões corretas da classe positiva.
- VN (Verdadeiros Negativos): número de previsões corretas da classe negativa.
- FP (Falsos Positivos): número de previsões erradas da classe positiva.
- FN (Falsos Negativos): número de previsões erradas da classe negativa.

3.2.2. Precisão

A precisão é utilizada para medir a proporção de instâncias positivas que são corretamente previstas em relação ao total de previsões na classe positiva [Hossin and Sulaiman 2015]. Pode ser dada pela seguinte formula:

$$\frac{\text{VP}}{\text{VP} + \text{FP}} \tag{2}$$

3.2.3. Revocação

A revocação é utilizado para medir a proporção de instâncias positivas que são corretamente classificadas [Hossin and Sulaiman 2015]. Pode ser representado pela seguinte fórmula:

$$\frac{\text{VP}}{\text{VP} + \text{VN}} \tag{3}$$

3.2.4. F1-score

Essa métrica representa a média harmônica entre os valores de revocação e precisão [Hossin and Sulaiman 2015]. É dada pela seguinte fórmula:

$$\frac{2 * \text{revocação} * \text{precisão}}{\text{revocação} + \text{precisão}} \tag{4}$$

3.2.5. Taxa de Falsos Positivos

A taxa de FP é calculada pelo número de FP divididos pelo total de negativos 5. A melhor taxa de FP é 0, enquanto a melhor taxa de FP é 1. Também pode ser calculado fazendo 1 - revocação [Vujović et al. 2021].

$$\frac{\text{FP}}{\text{FP} + \text{VN}} \tag{5}$$

3.2.6. Matriz de confusão

Matriz de confusão para um classificador binário (Tabela 1), os valores reais são marcados como Verdadeiro (1) e Falso (0), e são preditos como Positivo (1) e Negativo (0). As estimativas das possibilidades do modelo de classificação são derivadas das expressões VP, VN, FP e FN, que existem na matriz de confusão [Vujović et al. 2021].

	Verdadeiro (1)	**Falso (0)**
Positivo (1)	VP	FN
Negativo (0)	FP	VN

Tabela 1. Matriz de confusão para classificação binária. Fonte: Elaborado pelo autor.

3.2.7. Normalização

Normalização é o processo de converter os dados para o intervalo específico, como entre 0 e 1 ou entre -1 e +1. A normalização é necessária quando há grandes diferenças nos intervalos de diferentes recursos. Este método de dimensionamento é útil quando o conjunto de dados não contém *outliers* [Ali and Faraj 2014].

$$\text{X'} = \frac{X - \text{mínimo}}{\text{máximo -mínimo}} \tag{6}$$

3.2.8. Padronização

Criando um conjunto de dados com média = 0 e desvio padrão = 1. Este método de escala é útil quando os dados seguem uma distribuição normal (distribuição gaussiana), se os dados não seguem uma distribuição normal, isso causará problemas. [Ali and Faraj 2014].

$$\text{X''} = \frac{X - \text{média}}{\text{desvio padrão}} \tag{7}$$

3.3. Base de dados

As imagens utilizadas neste trabalho foram obtidas do *Curated Breast Imaging Subset of Digital Database for Screening Mammography* (CBIS-DDSM) [Sawyer-Lee et al. 2016], uma versão atualizada e padronizada do *Digital Database for Screening Mammography* (DDSM). Esse banco de dados reúne 2.620 estudos de mamografia digitalizados, incluindo a segmentação de regiões de interesse (*Region of Interest* - ROI, em inglês). As imagens estão classificadas em três categorias: *benign* (benigno), *benign without callback* (benigno sem acompanhamento) e *malignant* (maligno). Na Figura 1, alguns exemplos dessas regiões de interesse são apresentados, onde as linhas correspondem às classes: a primeira linha exibe amostras da classe *benign*, a segunda linha apresenta exemplos da classe *benign without callback* e a terceira linha mostra amostras da classe *malignant*.

3.3.1. Filtragem das imagens e redimensionamento

Primeiro, foi realizado um processo de filtragem, pois na pasta original temos as regiões de interesse já cortadas (*crops*), as máscaras e a mama completa. As regiões de interesse já cortadas foram separadas em uma pasta e organizadas entre as três classes: *benign*, *benign without callback* e *malignant*. Elas foram filtradas de modo a preservar aquelas com dimensões menores ou iguais a 512 por 512, que eram a maior parte delas. Para as que eram menores que isso, foi aplicada a técnica que adiciona pixels de valor zero em torno da imagem para ajustá-la a um tamanho específico (*zero padding*) [Muquet et al. 2002].

3.4. Primeiro experimento

O primeiro experimento tem como finalidade a extração das características radiômicas do conjunto de imagens selecionadas e depois determinar as 10 mais impactantes na classificação do nosso conjunto de dados.

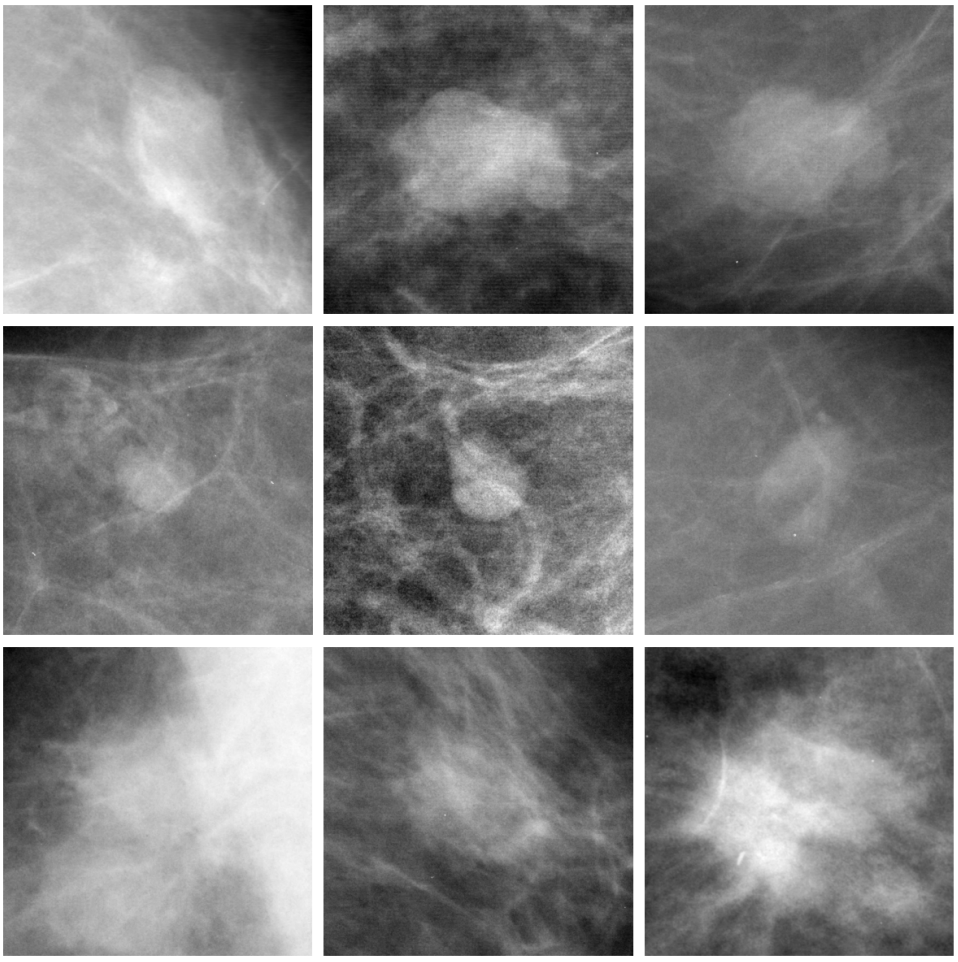

Figura 1. Exemplos de imagens de regiões de interesse (ROIs) que compõem o CBIS-DDSM. As linhas correspondem às diferentes classes de amostras: a primeira linha exibe amostras da classe *benign*, a segunda linha apresenta exemplos da classe *benign without callback*, e a terceira linha mostra amostras da classe *malignant*. Fonte: Elaborado pelo autor.

3.4.1. Extração das características radiômicas

As características radiômicas foram extraídas utilizando a biblioteca `PyRadiomics`. O método `extractor.execute` foi empregado para calcular as características a partir das imagens filtradas anteriormente. Após a extração, os resultados foram organizados em um *DataFrame* e salvos em um arquivo CSV para análise posterior.

Com as características obtidas, partimos para a etapa de filtragem delas, de modo que agora pretendemos ficar apenas com as mais relevantes para a classificação dos dados. Para isso, iremos utilizar a mesma abordagem descrita por [Dong et al. 2024] em seu artigo, ou seja, vamos usar um RFE. No caso dele, um total de 20 características foram selecionadas, neste trabalho, iremos manter apenas 10, como [Fleury and Marcomini 2019] fez em sua pesquisa.

3.4.2. Pré-processamento

Utilizando uma instância do `Label Encoder` da biblioteca `Scikit-learn` na coluna correspondente às classes, pudemos transformar esses campos que antes eram *Strings* em representações numéricas dessas classes, sendo a classe *benign* agora representada pelo número 0 a classe *benign without callback* representada pelo número 1 e a classe *malignant* agora representada pelo número 2.

Mais de 100 características radiômicas foram extraídas para cada ROI e, a partir daí, iniciou-se o pré-processamento para usar os dados nos modelos de classificação. O pré-processamento consistiu inicialmente em analisar a distribuição de classes. Foi constatado que havia 996 amostras pertencentes à classe *benign* e 814 amostras pertencentes à classe *malignant*. As demais amostras pertencentes à classe *benign without callback* foram descartadas e não serão mais usadas no decorrer do estudo. Isso porque a quantidade de amostras da classe *malignant* já é bem menor que a quantidade de amostras pertencentes à classe *benign*, então, para evitar o desbalanceamento dos dados, optamos por desconsiderar essa classe. Além disso, a natureza da classe *benign without callback* introduz uma incerteza adicional, uma vez que ela está relacionada a amostras que não receberam um acompanhamento posterior, deixando uma margem de dúvida quanto à sua classificação precisa.

Foram removidas as características que eram altamente correlacionadas, conforme mencionado no artigo de [Dong et al. 2024], que também excluiu características altamente correlacionadas. Foi desenvolvido um método interativo que mantém apenas uma das duas colunas correlacionadas, usando um limiar de 99%. Após essa etapa, restaram apenas 77 características . Analisando as tabelas, também era possível observar que algumas características geradas eram constantes, independentes das amostras, e essas colunas foram removidas do *DataFrame* também, resultando agora em um total de 66 características radiômicas.

3.4.3. Divisão dos dados

Para a divisão dos dados entre treino e teste foi utilizada a função `train_test_split` da biblioteca `scikit-learn` com os seguintes parâmetros:

- `test_size = 0.2`: Este parâmetro determina a proporção do conjunto de dados que será utilizada para testes. Neste caso, 20% dos dados serão reservados para validação do modelo, enquanto 80% serão utilizados para treino.
- `stratify = y`: Este parâmetro garante que a divisão dos dados mantenha a mesma proporção das classes presentes no conjunto de dados de treinamento. Isso é importante porque, como mencionado anteriormente, nossas classes já possuem um leve desbalanceamento. Assim, garantimos que tanto o conjunto de treino quanto o de teste reflitam a distribuição das classes.
- `random_state = 42`: Aqui asseguramos que as amostras escolhidas tanto para treino quanto para teste sejam selecionadas aleatoriamente. Usar o valor 42 como semente (seed, em inglês) do gerador garante a reprodução da divisão dos dados em execuções futuras.

3.4.4. Normalização

A normalização dos dados foi feita usando a biblioteca `scikit-learn` e o método `MinMaxScaler`, que faz a normalização dos dados exatamente como foi descrito anteriormente no artigo. A função `fit_transform` foi aplicada ao conjunto de treino para calcular os valores mínimos e máximos das características , transformando os dados para uma faixa entre 0 e 1. Em seguida, a função `transform` foi aplicada ao conjunto de teste para deixa-los na mesma escala que os dados de treino.

3.4.5. Múltiplos modelos

Após a normalização, os dados foram aplicados em oito modelos de ML distintos, sendo que sete deles eram da biblioteca `scikit-learn` que eram: RF, KNN, LR, SVC, DT, *gradiente boosting* e *adaboost*. O oitavo modelo foi o *LightGBM* da biblioteca de mesmo nome. O *LightGBM* foi incluído na lista pois é um algorítimo muito comumente utilizado em competições do kaggle devido sua alta performasse em tarefas de previsão [Li et al. 2022].

3.4.6. RFE

A biblioteca `scikit-learn` já nos fornece o método `RFE`, que permite realizar a eliminação recursiva de características . O método `RFE` do `scikit-learn` possui um parâmetro chamado `estimator`, que recebe uma instância de um estimador de aprendizado supervisionado com o método `fit` implementado, ou seja, pode receber qualquer um dos modelos utilizados anteriormente. Como a classificação anterior foi feita utilizando 66 características e nosso objetivo final é reduzir esse número para 10, nossa preocupação agora é garantir que o modelo passado tenha a melhor performance possível nesse conjunto de dados, de modo que, mesmo com a diminuição do número de características , seja possível atingir uma acurácia semelhante à obtida com mais informações. Adicionalmente, como anteriormente apenas utilizamos os parâmetros padrão, nesta segunda classificação aplicaremos um *grid search* para encontrar os melhores hiperparâmetros e garantir que a perda das demais características não impacte tanto no modelo, que já não possui uma acurácia global extremamente alta.

3.5. Segundo experimento

O segundo experimento tem como finalidade a criação de uma rede neural convolucional que receba as imagens e as classes, realizando a classificação entre elas. Para essa parte do trabalho, foi escolhida a biblioteca `PyTorch`, que nos oferece uma ampla gama de operações com tensores e suporta computação em GPU, acelerando o treinamento de modelos complexos. Assim, a principal finalidade desse experimento, além de nos dar uma ideia geral sobre o desempenho do nosso classificador de imagens, é também gerar uma base sólida para o experimento seguinte, onde adicionaremos as características radiômicas que coletamos no primeiro experimento às camadas convolucionais do segundo experimento.

3.5.1. Pré-processamento

Como queremos fazer três experimentos com técnicas de classificação diferentes, mas compará-los entre si, é prudente que também eliminemos as imagens referentes a classe *benign without callback*, já que no primeiro experimento optamos por excluir as características radiômicas referentes a essa classe, e é o que foi feito, nos deixando com um conjunto final de 1.810 imagens, sendo 996 da classe *benign* e 814 imagens pertencentes à classe *malignant*.

A divisão dos dados foi feita como no primeiro experimento, de modo que os mesmos dados nos conjuntos de treino e teste foram garantidos pelo uso da mesma semente e pela mesma proporção dos dados: 80% para treinamento e 20% para teste.

A padronização foi aplicada tanto nos dados de treino quanto nos dados de teste, mas usando apenas os dados de treino para obter a média e o desvio padrão, evitando assim o vazamento de dados.

Os dados foram passados para a rede convolucional em lotes de 64, de modo que os lotes provenientes do conjunto de treino são sempre embaralhados para garantir que o modelo aprenda em vez de memorizar os dados, enquanto os lotes provenientes do conjunto de teste foram mantidos sempre na mesma ordem, garantindo a consistência dos testes.

3.5.2. Criação da Rede convolucional

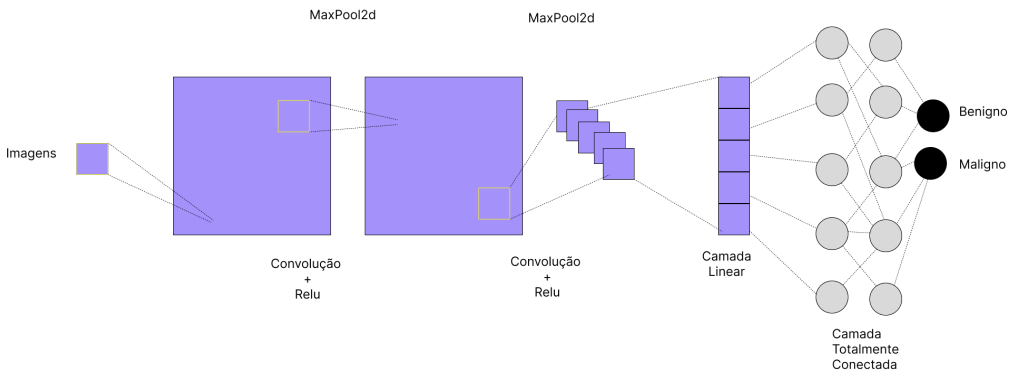

Figura 2. Representação da rede convolucional (imagens). Fonte: Elaborado pelo autor

Arbitrariamente, foi escolhido usar duas camadas convolucionais. A primeira, recebe uma entrada com 1 canal (nossa imagem em escala de cinza) e aplica 16 filtros de convolução, cada um com um `kernel` de 3 por 3, resultando em uma saída com 16 canais. Em seguida, `nn.ReLU()` transforma cada valor negativo em zero e mantém os valores positivos inalterados, ajudando a introduzir não-linearidade na rede. A camada `nn.MaxPool2d(2, 2)` aplica uma operação de pooling 2D com uma janela de 2 por 2, reduzindo a resolução da entrada pela metade. A segunda camada recebe uma entrada com 16 canais e aplica 32 filtros de convolução com um `kernel` de 3 por 3, seguida de outra `nn.ReLU()` e outro `nn.MaxPool2d(2, 2)`.

Em seguida, é definida uma camada linear que recebe como entrada a saída das camadas convolucionais e transforma essa entrada em um vetor de 256 unidades, `nn.Linear(32 * 126 * 126, 256)`. Depois, é aplicada uma `nn.ReLU()` e outra camada linear, `nn.Linear(256, 2)`, de modo que a saída corresponde às nossas classes.

3.5.3. Treinamento

Para essa etapa, foi utilizada `nn.CrossEntropyLoss()` como função de perda, `torch.optim.SGD(...)` como otimizador e uma taxa de aprendizado definida como `lr=0.001`. Adicionalmente, um *Early stopping* [Prechelt 2002] foi implementado, de modo que o número de épocas será definido pela perda acumulada do modelo e o critério de parada do treinamento como sendo 10 épocas. Ou seja, à medida que o treinamento é executado, verifica-se a perda. Se a perda não melhorar em 10 épocas, o treinamento é interrompido e retornado o modelo no estado em que teve a melhor performance.

3.6. Terceiro experimento

O terceiro experimento tem como finalidade a criação de uma rede neural convolucional que receba as imagens, as classes e as 10 características radiômicas selecionadas no primeiro experimento, realizando a classificação entre elas. Para este terceiro experimento, foi aproveitado basicamente tudo o que foi feito no segundo, pois a estrutura é essencialmente a mesma, com a diferença de que este experimento inclui as características radiômicas.

3.6.1. Pré-processamento

Conforme foi feito no segundo experimento, as imagens que serão utilizadas nesta etapa são as 996 da classe *benign* e 814 imagens pertencentes à classe *malignant*. Quando as características, elas foram extraídas e foram passadas para um CSV, que estava organizado da seguinte forma: a primeira coluna continha o nome da imagem, as colunas do meio continham as características correspondentes e a última coluna indicava a classe daquela imagem. Portanto, agora iremos criar um novo `DataFrame` de modo que as colunas centrais sejam apenas as 10 características selecionadas no primeiro experimento. Para fazer isso, simplesmente usamos os nomes que estão na tabela 3 para selecionar as colunas a serem mantidas, dessa forma, foram mantidos os valores originais sem qualquer tipo de normalização ou padronização.

Os dados foram divididos da mesma forma que nos experimentos anteriores, garantindo os mesmos dados nos conjuntos de treinamento e teste e na mesma proporção: 80% dos dados para treino e 20% para teste. Depois disso, a padronização foi aplicada aos dados e é importante ressaltar que, até este ponto, os dados das imagens e das características radiômicas ainda não foram combinados. Cada operação foi realizada separadamente. Tanto para as imagens quanto para as características radiômicas, os valores de média e desvio padrão foram calculados com base apenas nos conjuntos de treino e a padronização foi aplicada aos conjuntos de teste com base nesses valores, evitando assim o vazamento de dados.

3.6.2. Criação da Rede convolucional

A rede convolucional foi criada exatamente como descrita no segundo experimento, com as mesmas camadas convolucionais e lineares. A única diferença agora é que uma nova camada linear foi criada para as características radiômicas transforma-as em um vetor com 256 unidades, no final as informações são concatenadas em um único tensor, que será passado pela camada linear final, que agora conta com uma entrada maior para acomodar a adição das características radiômicas.

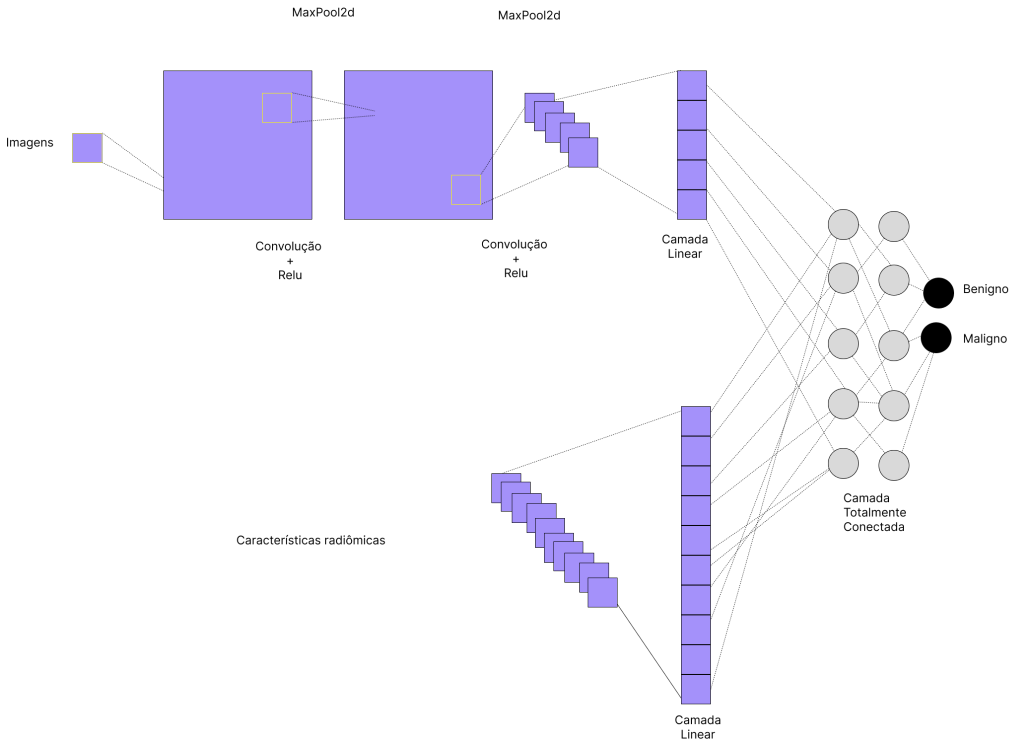

Figura 3. Representação da rede convolucional (imagens + características radiômicas). Fonte: Elaborado pelo autor

3.6.3. Treinamento

Nessa etapa, foram usados os mesmos parâmetros que no segundo experimento: a mesma função de perda `nn.CrossEntropyLoss()`, o mesmo otimizador `torch.optim.SGD(...)` e a mesma taxa de aprendizado `lr=0.001`, além do *Early stopping* para garantir que a quantidade de épocas no treinamento seja determinada pela perda acumulada do treinamento, com uma paciência definida em 10 épocas.

4. Resultados e discussões

Agora que serão apresentados os resultados, é importante contextualizar as métricas escolhidas ao tipo de problema que estamos estudando, para que possamos fazer uma análise correta dos resultados obtidos.

Em nosso estudo, os casos positivos referem-se à classe *benign*, enquanto os negativos pertencem à classe *malignant*. Nesse contexto, é extremamente importante considerarmos a gravidade das consequências associadas a falsos positivos e falsos negativos. Um falso positivo ocorre quando a paciente está doente, mas o modelo a classifica como saudável, enquanto um falso negativo ocorre quando a paciente não está doente, mas o modelo a classifica erroneamente como doente.

Podemos imaginar as consequências: um falso positivo é altamente problemático em termos de risco imediato à saúde da paciente, pois pode literalmente levá-la à morte por falta de tratamento. Portanto, esse é o caso que devemos dar prioridade em evitar no nosso modelo, e ele pode ser numericamente representado pela taxa de falsos positivos. Na matriz de confusão (tabela 1), é o canto inferior esquerdo da tabela. Quanto menor for esse número e, consequentemente, maior a revocação (mais próximo de 1) referente a classe *malignant*, melhor será o nosso modelo.

Um falso negativo, então, não seria tão problemático, porque se a paciente fosse falsamente diagnosticada com a doença, à medida que outros exames e tratamentos fossem feitos, logo se perceberia que ela não estava doente. Obviamente, também queremos aumentar os verdadeiros positivos e os verdadeiros negativos, que são representados pela revocação do modelo.

A acurácia é uma métrica importante que devemos observar. No entanto, no nosso contexto, ela pode passar uma falsa sensação de confiança. Em cenários com desbalanceamento de classes, um modelo pode apresentar alta acurácia apenas por classificar corretamente a maioria dos casos positivos, enquanto falha em identificar os casos negativos. Por exemplo, se 90% da população for saudável, um modelo que classifica todos os indivíduos como saudáveis terá 90% de acurácia. Embora nosso conjunto de dados não apresente uma discrepância tão grande, é importante lembrar que também temos um desbalanceamento de classes no nosso estudo, e esse foi um dos fatores que motivaram a desconsideração das amostras da classe *benign without callback*.

Dessa forma, enquanto a acurácia nos fornece uma visão geral do desempenho do modelo, ela precisa ser analisada em conjunto com métricas como a revocação e a taxa de falsos positivos para que nossa avaliação seja consistente e adequada ao problema em questão. Em situações onde os falsos positivos e os falsos negativos têm impactos distintos e graves, como no nosso caso, é fundamental priorizar métricas que reflitam a capacidade do modelo em minimizar esses erros específicos. Assim, devemos utilizar essas métricas de maneira complementar.

4.1. Primeiro experimento

Todos os classificadores foram instanciados com os parâmetros padrão das respectivas bibliotecas, e os resultados da classificação estão listados na Tabela 2.

Ao final da segunda classificação, utilizando o *LightGBM* como estimador do RFE e aplicando o *grid search* para encontrar os melhores hiperparâmetros, obtivemos

Modelo	Acurácia
LightGBM	0,640884
RF	0,627072
AdaBoost	0,616022
LR	0,610497
Gradient Boosting	0,610497
KNN	0,574586
SVC	0,569061
DT	0,513812

Tabela 2. Acurácia dos Modelos. Fonte: Elaborado pelo autor.

uma acurácia de 60%. Na Tabela 3, podemos ver as características que restaram após a aplicação do RFE.

Características Radiômicas Selecionadas
original_firstorder_90Percentile
original_firstorder_Energy
original_firstorder_Entropy
original_firstorder_Kurtosis
original_firstorder_Skewness
original_glcm_Idmn
original_gldm_DependenceNonUniformityNormalized
original_ngtdm_Coarseness
original_ngtdm_Contrast
original_ngtdm_Strength

Tabela 3. Características radiômicas selecionadas pelo RFE. Fonte: Elaborado pelo autor.

Para mais informações sobre a modelagem matemática e a fundamentação teórica por trás dessas características , o trabalho [Zwanenburg et al. 2016] foi a referência utilizada na construção da biblioteca `PyRadiomics`.

Conforme vimos na Tabela 2, a maior acurácia para esse experimento foi de 64%, mas será que isso reflete um bom desempenho? Primeiro, uma acurácia de 64% à princípio não parece ser muita coisa, mas, se levarmos em conta o resultado do segundo experimento, em que, usando as mesmas imagens de onde as características para este experimento foram extraídas, obteve-se uma acurácia global de 60%, esse resultado passa a ser bem satisfatório. Isso mostra que as características radiômicas, apesar de serem características limitadas da imagem, nos dão uma representação levemente melhor que as das imagens completas.

Analisando o relatório de classificação do modelo (Tabela 4) e a matriz de confusão correspondente (Figura 4), podemos ver que o revocação referente à classe *malignant* (56%) está bem abaixo da revocação da classe *benign* (71%), isso indica que o modelo teve mais facilidade em identificar corretamente os casos refentes a classe *benign*.

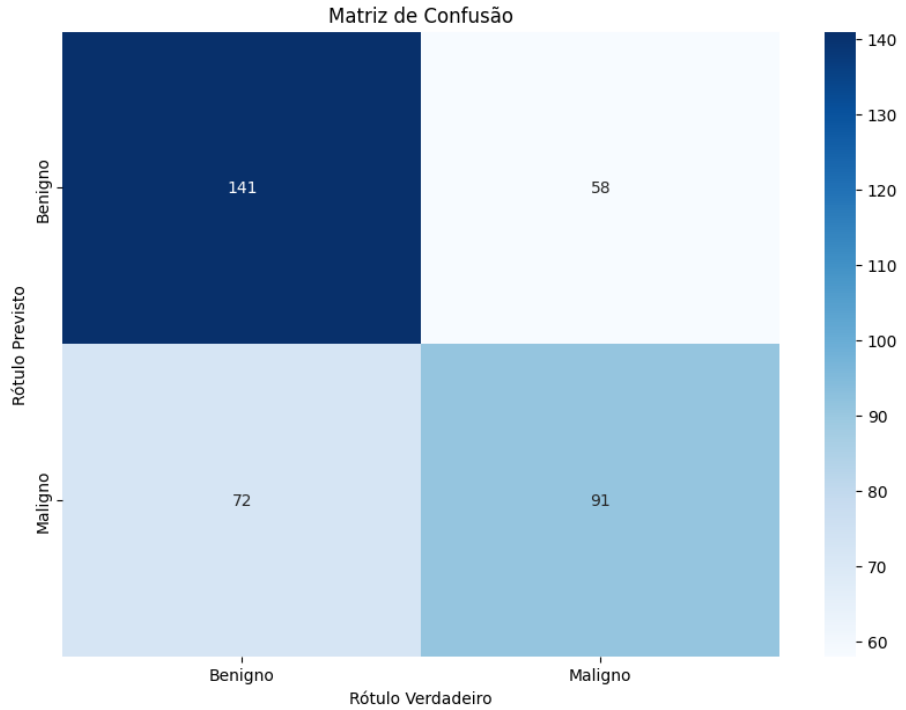

Figura 4. Matriz de confusão: LightGBM com 66 características . Fonte: Elaborado pelo autor

	Precisão	Revocação	F1-score	Quantidade de Amostras
Benigno	0,66	0,71	0,68	199
Maligno	0,61	0,56	0,58	163
Acurácia			0,64	362
Média geral	0,64	0,63	0,63	362
Média ponderada	0,64	0,64	0,64	362

Tabela 4. Relatório de classificação: LightGBM com 66 características radiômicas. Fonte: Elaborado pelo autor

Utilizando o mesmo classificador, mas com o número de características reduzido a 10, a queda na acurácia não foi tão grande: foi de apenas 4%. No entanto, ao analisarmos a matriz de confusão correspondente (Figura 5) e seu relatório de classificação associado (Tabela 5), percebemos que ocorreu o pior cenário possível. A perda na acurácia global foi toda ao custo da revocação referente à classe *malignant* (caiu de 56% para 46%), ou seja, tivemos o resultado diametralmente oposto ao que buscávamos, já que aumentaram os falsos positivos. Isso sugere que usar um número tão reduzido de características não é uma boa escolha, pelo menos em uma classificação que leva apenas as características radiômicas em consideração.

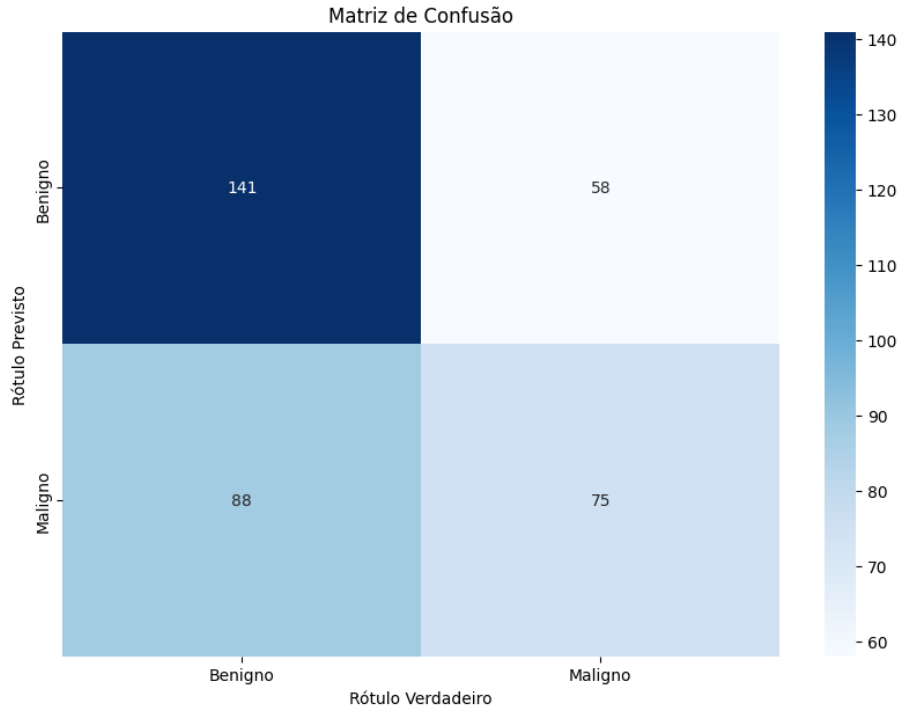

Figura 5. Matriz de confusão: LightGBM com 10 características . Fonte: Elaborado pelo autor

	Precisão	Revocação	F1-score	Quantidade de Amostras
Benigno	0,62	0,71	0,66	199
Maligno	0,56	0,46	0,51	163
Acurácia			0,60	362
Média geral	0,59	0,58	0,58	362
Média ponderada	0,59	0,60	0,59	362

Tabela 5. Relatório de classificação: LightGBM com 10 características radiômicas. Fonte: Elaborado pelo autor

4.2. Segundo experimento

Com base na matriz de confusão (Figura 6) e no relatório de classificação (Tabela 6), podemos ver os falsos positivos e analisar a taxa de falsos positivos, que é de 48%. Esse valor é menor ao obtido no experimento um usando 10 características radiômicas (54%), mas é maior ao mesmo experimento usando 66 características radiômicas (44%), indicando que esse modelo ficou no meio-termo.

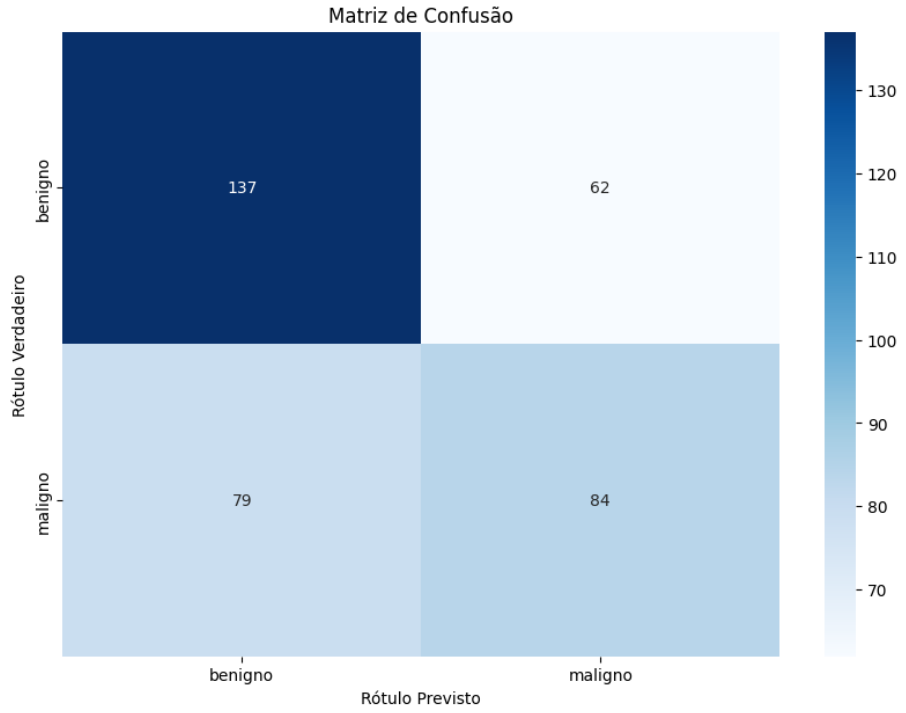

Figura 6. Matriz de confusão Imagens: rede convolucional com imagens. Fonte: Elaborado pelo autor

	Precisão	Revocação	F1-score	Quantidade de Amostras
Benigno	0,63	0,69	0,66	199
Maligno	0,58	0,52	0,54	163
Acurácia			0,61	362
Média geral	0,60	0,60	0,60	362
Média ponderada	0,61	0,61	0,61	362

Tabela 6. Relatório de classificação: Rede convolucional com imagens. Fonte: Elaborado pelo autor

4.3. Terceiro experimento

Com base na matriz de confusão (Figura 7) e no relatório de classificação (Tabela 7), podemos observar os falsos positivos e analisar a taxa de falsos positivos, que é de 35%. Esse valor é inferior a qualquer outro obtido até o momento, seja no primeiro ou no segundo experimento (44%, 48% e 54%, respectivamente), e a acurácia de 78% é a maior até agora, indicando que este foi o experimento com os melhores resultados obtidos até o momento.

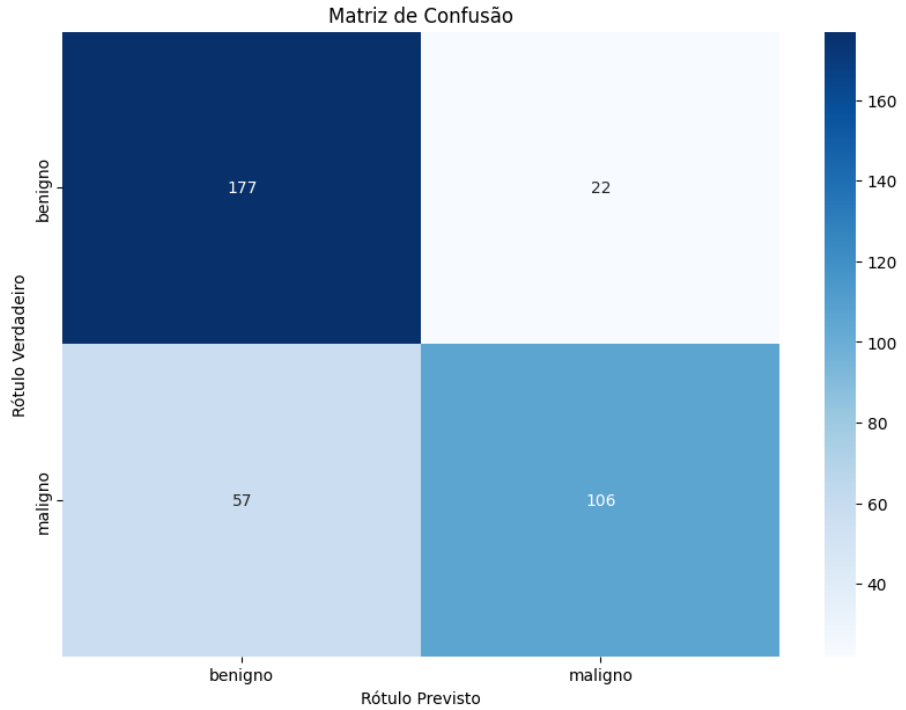

Figura 7. Matriz de confusão Imagens: rede convolucional com imagens e as principais características radiômicas. Fonte: Elaborado pelo autor

	Precisão	Revocação	F1-score	Quantidade de Amostras
Benigno	0,76	0,89	0,82	199
Maligno	0,83	0,65	0,73	163
Acurácia			0,78	362
Média geral	0,79	0,77	0,77	362
Média ponderada	0,79	0,78	0,78	362

Tabela 7. Relatório de classificação: Rede convolucional com imagens e as principais características radiômicas. Fonte: Elaborado pelo autor

5. Conclusão

Este trabalho propôs a criação de uma rede convolucional para aumentar a acurácia na classificação de imagens mamográficas, possibilitando uma classificação mais eficaz entre lesões benignas e malignas incorporando, além das imagens, suas respectivas características radiômicas. Para isso, foram conduzidos três experimentos distintos. No primeiro, buscou-se identificar as características radiômicas mais relevantes associadas ao conjunto de imagens, reduzindo-as às 10 melhores entre mais de 100 disponíveis. O segundo experimento consistiu em desenvolver uma rede convolucional capaz de classificar as imagens em duas classes: maligno ou benigno. Por fim, o terceiro experimento integrou as características radiômicas ao modelo convolucional inspirado no segundo experimento. Os resultados demonstraram que o objetivo foi alcançado com sucesso: o modelo final obteve um aumento de 17% na acurácia e uma redução de 13% na taxa de falsos positivos em relação ao segundo modelo (que utilizava apenas as imagens). Em comparação ao primeiro experimento, que usava 66 características radiômicas, houve um aumento de 14% na acurácia e uma redução de 9% na taxa de falsos positivos.

5.1. Trabalhos Futuros

O modelo final foi construído com base em uma análise multi-classes. No entanto, devido às particularidades deste trabalho, as classes foram limitadas a duas: maligno e benigno. Dito isso, uma abordagem alternativa poderia ter sido adotada, utilizando uma classificação binária. Com a flexibilidade do modelo construído, diversas modificações poderiam ser realizadas e testadas nas camadas convolucionais, como o aumento da dimensionalidade ou da quantidade de camadas. Também poderiam ser testadas outras combinações e quantidades de características radiômicas.

Referências

Ali, P. J. M. and Faraj, R. H. (2014). Data normalization and standardization: A technical report. *Machine Learning Technical Reports*, 1(1):1–6.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.

Dong, V., Barufaldi, B., Mankowski, W., Silva Filho, T. M., McCarthy, A. M., Kontos, D., and Maidment, A. D. (2024). Explainable radiomics to characterize breast density and tissue complexity: preliminary findings. In *17th International Workshop on Breast Imaging (IWBI 2024)*, volume 13174, pages 393–406. SPIE.

Fleury, E. and Marcomini, K. (2019). Performance of machine learning software to classify breast lesions using bi-rads radiomic features on ultrasound images. *European Radiology Experimental*, 3(1):34.

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):11.

Khalid, A., Mehmood, A., Alabrah, A., Alkhamees, B. F., Amin, F., AlSalman, H., and Choi, G. S. (2023). Breast cancer detection and prevention using machine learning. *Diagnostics*, 13(19):3113.

Li, X., Bai, Y., and Kang, Y. (2022). Exploring the social influence of the kaggle virtual community on the m5 competition. *International Journal of Forecasting*, 38(4):1507–1518.

Muquet, B., Wang, Z., Giannakis, G. B., De Courville, M., and Duhamel, P. (2002). Cyclic prefixing or zero padding for wireless multicarrier transmissions? *IEEE Transactions on communications*, 50(12):2136–2148.

Prechelt, L. (2002). Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

Sawyer-Lee, R., Gimenez, F., Hoogi, A., and Rubin, D. (2016). Curated breast imaging subset of digital database for screening mammography (cbis-ddsm). Data set. `https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY`.

Vujović, Ž. et al. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 12(6):599–606.

Zwanenburg, A., Leger, S., Vallières, M., and Löck, S. (2016). Image biomarker standardisation initiative. *arXiv preprint arXiv:1612.07003*.