Carla Estefany Caetano Silva

Detecção de câncer de mama por meio de análise de imagem com os descritores de Haralick e aprendizado de máquina.

Niterói - RJ, Brasil

18 de dezembro de 2023

Universidade Federal Fluminense

Carla Estefany Caetano Silva

Detecção de câncer de mama por meio de análise de imagem com os descritores de Haralick e aprendizado de máquina.

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Prof. Dra. Karina Yuriko Yaginuma

Niterói - RJ, Brasil

18 de dezembro de 2023

Universidade Federal Fluminense

Carla Estefany Caetano Silva

Detecção de câncer de mama por meio de análise de imagem com os descritores de Haralick e aprendizado de máquina.

Monografia de Projeto Final de Graduação sob o título *"Detecção de câncer de mama por meio de análise de imagem com os descritores de Haralick e aprendizado de máquina."*, defendida por Carla Estefany Caetano Silva e aprovada em 18 de dezembro de 2023, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profa. Dra. Nome do Orientador
Departamento de Estatística – UFF

Prof. Me. Nome do 1o membro da banca
Instituicao do 1º membro da banca

Profa. Ma. Nome do 2o membro da banca
Instituicao do 2º membro da banca

Niterói, 18 de dezembro de 2023

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

C127d Caetano Silva, Carla Estefany
Detecção de câncer de mama por meio de análise de imagem com os descritores de Haralick e aprendizado de máquina. / Carla Estefany Caetano Silva. - 2023.
57 f.: il.

Orientador: Karina Yuriko Yaginuma.
Trabalho de Conclusão de Curso (graduação)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2023.

1. Descritores de Haralick. 2. Segmentação binária. 3. Aprendizado de máquinas. 4. Câncer de mama. 5. Produção intelectual. I. Yaginuma, Karina Yuriko, orientadora. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

Este trabalho de conclusão de curso aborda a detecção de câncer de mama por meio de técnicas de processamento de imagem e aprendizado de máquina. O objetivo é desenvolver um modelo de classificação capaz de analisar as mamografias e auxiliar radiologistas no diagnóstico precoce de possíveis casos de câncer de mama. A metodologia proposta envolve etapas de segmentação binária para identificação de regiões de interesse, extração de descritores de Haralick para caracterização das características das regiões, e utilização de modelos de aprendizado de máquina para a classificação dos casos. A implementação e avaliação da metodologia será realizada em um conjunto de dados clínicos relevantes, com o intuito de validar a eficácia do sistema proposto no auxílio ao diagnóstico de câncer de mama. Os resultados obtidos para imagens de compressão médio lateral, o melhor modelo foi o *Adaboost* que se destacou com 85,71% de acurácia, 87,50% de sensibilidade e 83,93% de especificidade. Já para as imagens do tipo crânio caudal o melhor modelo foi o KNN destacando-se com uma acurácia de 88,39%, sensibilidade de 91,07% e especificidade de 85,71%. Os modelos criados, alcançaram resultados desejáveis, contudo, embora os resultados não tenham atingindo patamares ideias, os objetivos deste estudo foram alcançados. E é evidente que existe um potencial significativo para melhorar futuramente.

Palavras-chave: Descritores de Haralick. Segmentação binária. Aprendizado de máquinas. Câncer de mama.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Neste trabalho visamos analisar imagens de mamografia em pacientes do sexo feminino com o intuito de entender e detectar o câncer de mama que é uma patologia que ocorre devido a um crescimento e multiplicação desordenado de células anormais na mama.

1.1 Motivação

O câncer de mama é uma das doenças mais comuns entre as mulheres e, se não for detectado precocemente, pode ser fatal. Os métodos computacionais para detecção de câncer de mama são uma área de pesquisa em rápido crescimento que visa melhorar a precisão e a eficiência do diagnóstico. Esses métodos envolvem o uso de algoritmos de aprendizado de máquina e outras técnicas de inteligência artificial para analisar imagens médicas e identificar anomalias que indiquem a presença de câncer de mama. A detecção precoce do câncer de mama é fundamental para aumentar as chances de sobrevivência e melhorar a qualidade de vida das pacientes. Além disso, o estudo desses métodos pode levar a avanços significativos na área de inteligência artificial e aprendizado de máquina, que podem ter aplicações em outras áreas da medicina e além.

Segundo o Instituto Nacional do Câncer (INCA)[1], o câncer de mama é uma doença causada pela multiplicação desordenada de células anormais que formam um tumor na mama com potencial de invadir outros órgãos. No Brasil, em 2020, 108.318 mulheres morreram por algum tipo de neoplasia [2], cerca de 16,5% foram de câncer de mama, no ano de 2022 houve 362.730 novos casos de mulheres com algum tipo de neoplasia, 30,1% desses casos foram de câncer de mama.

Os sintomas mais comuns são nódulos (caroços) onde 90% dos casos é percebido pela mulher, pele da mama avermelhada, alterações no bico do peito, pequenos nódulos nas

[1] https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/c/cancer-de-mama

[2] Massa de tecido anormal que crescer em qualquer lugar do corpo, onde esse crescimento exagerado pode provocar riscos diferentes ao ser humano devido as suas características.

axilas ou no pescoço e saída espontânea de líquido anormal pelos mamilos. O diagnóstico é realizado mediante a exames de imagens como mamografias, ultrassonografia ou ressonância magnética [3].

Para analisar uma mamografia de maneira eficiente é preciso uma boa interpretação e atenção nos mínimos detalhes. O exame pode identificar lesões benignas e cânceres, que frequentemente são encontrados como nódulos, ou calcificações. A detecção precoce de câncer de mama no seu estágio inicial é fundamental para a recuperação do paciente, aumentando a taxa de sucesso do tratamento e a redução da mortalidade. Com um diagnóstico tardio, pode-se levar a um tratamento agressivo, diminuindo a taxa de sobrevivência das pacientes.

Conforme o Portal Hospitais Brasil [4], os estágios (ou estadiamentos) do câncer são classificados segundo o grau de gravidade da doença e os custos aumentam segundo a magnitude do problema observado, exigindo maior intensidade no tratamento e custos elevados.

Segundo o INCA[5], no Brasil, ao longo das últimas duas décadas pode ser observado um aumento da proporção de cânceres *in situ* (estádio 0, é um carcinoma restrito a área que aparece) e em estádio I, com redução da apresentação em estádio II. Estes estádios são os tumores iniciais onde o estádio I não apresenta comprometimento linfático, e o estádio II é o início do espalhamento no tecido inicial ou em mais de um tecido com comprometimento linfático. Cerca de 40% dos casos são diagnosticados em fase avançada no estádio III, os tumores são localmente avançados, espalhados por mais de um tecido e causando comprometimento linfático e o estádio IV é considerado metástase a distância, ou seja, quando o tumor está se espalhando para outros órgãos ou todo o corpo.

A maioria dos casos de câncer de mama recebeu o primeiro tratamento oncológico após 60 dias da data do diagnóstico. Entre os anos de 2019 a 2021, a proporção de casos de câncer de mama tratados após 60 dias diminuiu. Contudo, a proporção de casos tratados em até 30 dias praticamente não se alterou (INCA,2022) [6].

Este trabalho visa contribuir para melhorias para o estudo e diagnostico de doenças, ajudando e auxiliando o profissional da saúde com diagnósticos mais preciso e rápidos,

[3]https://www.nucleodoconhecimento.com.br/saude/mamografia-aspectos-gerais

[4]https://portalhospitaisbrasil.com.br/artigo-custos-do-tratamento-do-cancer-no-brasil-como-melhorar-o-foco/

[5]https://www.gov.br/inca/pt-br/assuntos/gestor-e-profissional-de-saude/controle-do-cancer-de-mama/dados-e-numeros/estadiamento

[6]https://www.gov.br/inca/pt-br/assuntos/gestor-e-profissional-de-saude/controle-do-cancer-de-mama/dados-e-numeros/tempo-para-o-tratamento

designando o paciente para um melhor tratamento. Tendo em vista também a diminuição dos custos do tratamento realizado.

1.2 Revisão Bibliográfica

Nesta seção abordaremos brevemente os artigos relacionados ao uso de aprendizado de máquinas e redes neurais para a detecção de câncer de mama e suas medidas de qualidade de ajuste.

O artigo de Roy et al. (2021) apresenta um estudo sobre a detecção e classificação de câncer de mama em imagens de mamografia. Os resultados mostraram que o método CatBoost foi o melhor modelo, com uma acurácia de 92,5% e uma precisão média de 93,4% na detecção de câncer de mama em imagens de mamografia, superando outras técnicas de classificação utilizadas em estudos anteriores.

Já o artigo de Gonçalves (2021) mostra a importância do aprendizado de máquinas na análise dos dados e seu potencial uso no campo da medicina, em especial no diagnóstico de câncer de mama. O objetivo do estudo foi construir um modelo de predição para diagnóstico de câncer de mama usando aprendizado de máquina. O modelo de predição utilizado para o treinamento é o algoritmo máquina vetorial de suporte (SVM) que apresentou melhores resultados. O algoritmo sendo utilizado com os métodos de reamostragem atingiu métricas 99,11% (acurácia), 98,61% (precisão), 100% (sensibilidade), 97,62% (Especificidade).

O artigo de Marques et al. (2017) propõe uma metodologia que utiliza características de textura para reconhecer regiões de massa em mamografias. Segundo o artigo, a extração de características de textura foi feita usando Vetores de Descritores Localmente Agregados (VLAD) para obter informações de textura em regiões de massa nas mamografias. O VLAD é um método de descrição de imagem que utiliza a técnica de clusterização para gerar um dicionário de descritores locais, os quais são então utilizados para criar um vetor de características localmente agregadas. O algoritmo *Random Forest* foi utilizado para classificação dessas regiões como benignas ou malignas. Os testes alcançaram resultados promissores, foi utilizado o Banco de Dados Digital para Mamografia de Rastreamento (DDSM) e o modelo desenvolvido obteve 91,83% (acurácia), 92,28% (sensibilidade) e 94.37% (especificidade).

De acordo com Haralick, Shanmugam e Dinstein (1973), a utilização de características texturais são medidas benéficas e podem ser aplicadas para a classificação de imagens. Os

descritores são calculados utilizando a matriz de co-ocorrência de níveis de cinza, sendo uma medida objetiva e direta da distribuição de intensidade da imagem. Essas matrizes são a base para os descritores propostos por Robert Haralick e colaboradores. Esses descritores de textura são utilizados para extrair recursos de imagens médicas, incluindo mamografias, que podem auxiliar na detecção precoce do câncer de mama. Os descritores de Haralick conseguem capturar informações de textura de alta ordem, que podem ser difíceis de descrever com outros descritores.

A revisão bibliográfica revela um panorama animador na detecção de câncer de mama, com abordagens de aprendizado de máquinas ressaltada em Roy et al. (2021) e técnicas texturais propostas em Haralick, Shanmugam e Dinstein (1973) e a junção do aprendizado de máquinas com características de textura mencionada em Marques et al. (2017). Essa combinação pode fornecer informações e padrões importantes e complementares. Ao explorar essas duas vertentes, busca-se uma abordagem mais abrangente e precisa para o diagnóstico auxiliado por computador do câncer de mama. Este estudo busca não apenas contribuir para a pesquisa, mas também para a prática clínica, promovendo avanços significativos na detecção precoce e precisa desta doença desafiadora.

1.3 Objetivos

Este trabalho pretende:

- Utilizar morfologia para analisar e segmentar imagens de mamografia;
- Analisar as partes segmentadas separadamente e utilizar técnicas para quantificar cada segmento (objeto) em medidas descritivas como área, perímetros, medidas de textura entre outras;
- Usando as medidas extraídas dos objetos segmentados, deseja-se desenvolver um modelo de classificação baseado em aprendizado de máquinas capaz de identificar quais objetos possuem câncer;
- Comparar o desempenho de diferentes algoritmos de classificação para detectar o câncer de mama em imagens de mamografia;
- Analisar o desempenho dos descritores de textura na detecção de câncer de mama em imagens de mamografia.

1.4 Organização

O presente trabalho está organizado de maneira a proporcionar uma abordagem clara e estruturada para a análise da problemática proposta. O Capítulo 2 apresenta a metodologia adotada, detalhando os procedimentos experimentais, a coleta de dados e as ferramentas utilizadas para a análise. O Capítulo 3 apresenta os resultados e sua interpretação para fornecer percepções significativos para a resolução da problemática. No Capítulo 4, as conclusões do estudo são apresentadas, destacando contribuições específicas, limitações identificadas e possíveis direções para pesquisas futuras. Além disso, são fornecidas recomendações práticas derivadas dos resultados obtidos.

2 Materiais e Métodos

2.1 Base de dados projeto DDSM

Conforme o site da Universidade do Sul da Flórida (USF) (CAI et al., 2023), o Banco de Dados Digital para Mamografia de Rastreamento (DDSM) (HEATH et al., 1998) é um recurso para uso pela comunidade de pesquisa em análise de imagens mamográficas, com o objetivo principal de facilitar a pesquisa sólida no desenvolvimento de algoritmos computacionais para auxiliar na triagem. O banco de dados contém aproximadamente 2.500 sujeitos. Cada estudo inclui duas imagens de cada mama, com informações sobre a paciente como idade no momento do estudo, classificação de densidade mamária (ACR) e classificação de sutileza para anormalidades, entre outros. As imagens contendo áreas suspeitas foram segmentadas em vermelho por um profissional. Na Figura 1 apresentada as informações referente ao paciente e seu respectivo diagnóstico. Objetivo deste trabalho e classificar qual área da imagem possui uma anormalidade. Mas detalhes de como esta base será utilizada será discutida na sessão de pré-processamento. Ou seja, nesta sessão discutiremos a utilização desta base de dados para construir a base dados que será utilizada neste trabalho.

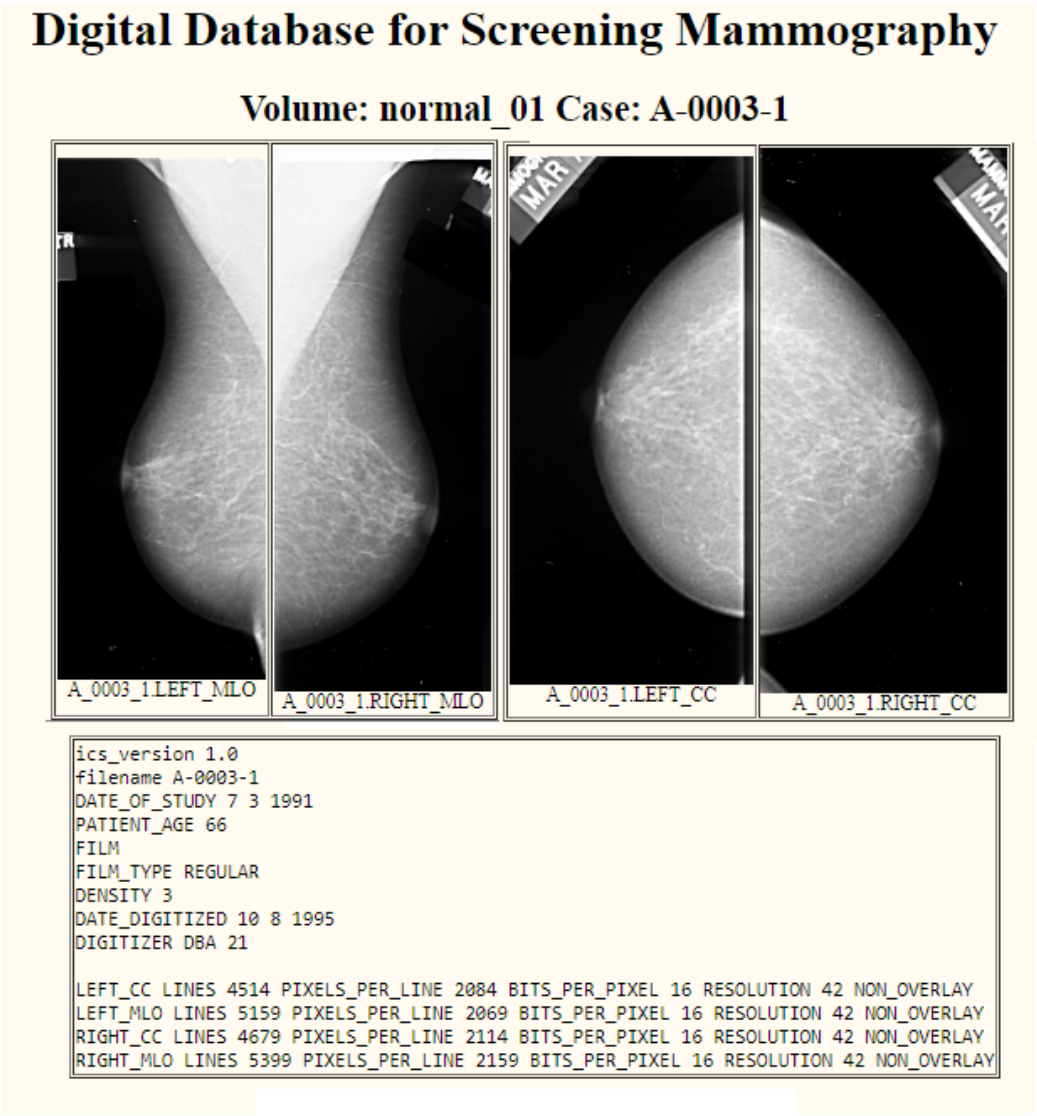


Figura 1: A página da web mostrando a versão em miniatura das imagens para case 0003 no DDSM.

2.2 Matriz de co-ocorrência

Nesta sessão será explicado o conceito da matriz de co-ocorrência que será utilizada para calcular os descritores de textura que serão discutidos na próxima sessão. As matrizes de co-ocorrência são a base para podermos calcular os descritores de textura de Haralick (HARALICK; SHANMUGAM; DINSTEIN, 1973). A utilização dos descritores de textura será explicado na sessão de pré-processamento.

A matriz de co-ocorrência é uma ferramenta muito utilizada na análise de imagens para extrair informações que descrevem a distribuição espacial da intensidade de cinza em uma imagem. Segundo Conci (2017) uma matriz de co-ocorrência (*GLCM, do inglês,*

Gray-Level Co-occurrence Matrix) é uma tabulação de quantas combinações diferentes de valores de intensidade dos píxeis (níveis de cinza) ocorrem em uma imagem. Assim é possível calcular várias medidas de textura como, por exemplo, a entropia, homogeneidade, energia, contraste entre outras que serão definidas mais adiante. No entanto, à medida que o número de tons de cinza aumenta na imagem, o seu custo computacional aumenta muito.

A construção da matriz de co-ocorrência é obrigatoriamente uma matriz quadrada, porque conterá a mesma quantidade de linhas e colunas que será o número total de níveis de cinza. Ou seja, o tamanho da matriz vai depender da quantidade de tons de cinza (se uma imagem tiver 256 tons de cinza, a matriz de co-ocorrência será uma matriz 256×256 porque inclui o nível de cinza 0).

Segundo Conci (2017) a matriz de co-ocorrência pode ser definida como uma matriz de frequências relativas $P(i, j, d, \theta)$, onde dois vizinhos i e j são separados por uma distância d e um ângulo, θ um com um nível de cinza i e outro com um nível de cinza j. Na matriz de co-ocorrência, a referência dos níveis de cinza i e j representaram as posições da matriz. Onde o nível de cinza i representa a linha i, e o nível de cinza j representará a coluna j da matriz de co-ocorrência. Por isso, cada posição da matriz se refere a comparação de dois níveis de cinza. Por exemplo, para o tom de cinza i = 0 e j = 1, na matriz de co-ocorrência a célula da posição i e j terá a frequência de ocorrência do nível de cinza 0 sendo tendo como vizinho o nível de cinza 1, de acordo com uma distância d e um ângulo θ estipulados.

Considerando um d (distância) fixo, para cada um dos ângulos 0^{o}, 45^{o}, 90^{o} e 135^{o} a direção dos pares de níveis de cinza serão comparados na matriz de co-ocorrência. Será considerado uma distância d entre as posições de um nível de cinza central e seu vizinho adjacente dispostos nessa orientação específica na Figura 2.

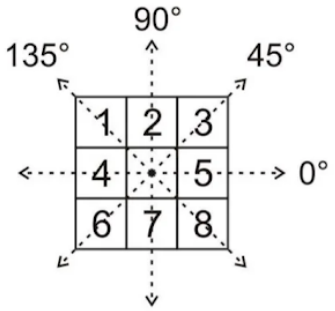


Figura 2: Ângulos definidos para calcular a matriz de co-ocorrência

Segundo Schwartz (2005) a especificação desses quatro ângulos 0º, 45º, 90º e 135º expressaram a associação espacial, para calcular essas transições deve-se associar o píxel central com seus vizinhos conforme o ângulo escolhido.

Por fim, normaliza-se a matriz de co-ocorrência dividindo cada elemento (i,j) pelo soma de todos os elementos (i,j) da própria matriz quadrada de co-ocorrência. A matriz de co-ocorrência no final é uma matriz de probabilidades P(i,j) onde cada elemento (i,j) representa a probabilidade de um certo valor aparecer na matriz, sendo calculado pela equação P(i,j) = $\frac{M_{i,j}}{\sum_{i,j=1}^{N}(M_{i,j})}$, onde i é o número da linha e j é o número de coluna, P(i,j) é a probabilidade da célula (i,j), $M_{i,j}$ é o conteúdo da célula (i,j), N é o número de linhas ou colunas de uma matriz quadrada, considerando uma distância d e um ângulo θ, note que a equação poderia ser escrita também como P(i,j,d,θ).

Tabela 1: Matriz de co-ocorrência com três níveis de tons de cinza 0,1 e 2.

	0	1	2
0	P(0,0,d,θ)	P(0,1,d,θ)	P(0,2,d,θ)
1	P(1,0,d,θ)	P(1,1,d,θ)	P(1,2,d,θ)
2	P(2,0,d,θ)	P(2,1,d,θ)	P(2,2,d,θ)

Suponha uma imagem 5x5 em tons de cinza, assuma que os valores de píxeis em escala de cinza variam de 0 a 2 na Tabela 2 Esta imagem possui apenas 3 níveis de cinza, $Z_0 = 0, Z_1 = 1$ e $Z_2 = 2$.

Tabela 2: Imagem com níveis de cinza 0, 1 e 2, retirada de Gonzalez e Woods, 2002, p. 363.

0	0	0	1	2
1	1	0	1	1
2	2	1	0	0
1	1	0	2	0
0	0	1	0	1

Para esse exemplo considere a distância d = 1, para cada ângulo θ terá uma matriz de co-ocorrência. Veja a ilustração abaixo do resultado da matriz de co-ocorrência para cada ângulo. Consideraremos a metodologia na qual a diagonal principal da matriz é dobrada, ou seja, multiplicada por 2.

A matriz de co-ocorrência é calculada da seguinte maneira, para o ângulo $\theta = 0º$ e d = 1 (distância), cada célula (i,j) da imagem que será vista como uma matriz, será

comparada com a próxima célula adjacente (i+1,j) para cada tom de cinza Z_i (0, 1 ou 2) e será contabilizada toda vez que essa co-ocorrência ocorrer. Isso será feito para os pares de tons de cinza que podem ser formados, isso é ilustrado na Tabela 1. Veja que, são comparados os níveis de cinza (0,0), (0,1), (0,2), (1,0) e assim respectivamente até o último parte de comparações (2,2).

Considere a notação $I_{i,j}$ sendo a posição do píxel (i,j) da imagem. Observe a ilustração na Tabela 2, se queremos saber a ocorrência do tom zero com um vizinho adjacente que possui o tom zero, ou seja, queremos saber a frequência dessas ocorrências para preencher o ponto (0,0) da matriz de co-ocorrência ilustrada na Tabela 1. Essas ocorrências acontecem nas posições $I_{0,0} \longrightarrow I_{0,1}$, $I_{0,1} \longrightarrow I_{0,2}$, $I_{2,3} \longrightarrow I_{2,4}$, $I_{4,0} \longrightarrow I_{4,1}$, veja que tivemos 4 ocorrências, como estamos usando a metodologia onde a diagonal principal da matriz é multiplicada por 2, logo a célula da matriz de co-ocorrência no ponto (0,0) será 8.

De forma geral, para cada ângulo θ considerando uma distância d, onde $d \geq 1$, as comparações das células da imagem serão da forma:

- Para o ângulo 0^{o}: cada célula (i,j) da imagem será comparada com a próxima célula adjacente (i+d,j).
- Para o ângulo 45^{o}: cada célula (i,j) da imagem será comparada com a próxima célula adjacente (i-d,j+d).
- Para o ângulo 90: cada célula (i,j) da imagem será comparada com a próxima célula adjacente (i-d,j).
- Para o ângulo 135^{o}: cada célula (i,j) da imagem será comparada com a próxima célula adjacente (i-d,j-d).

Por fim, deve-se exemplificar que para cada valor de distância escolhido haverá uma matriz de co-ocorrência distinta, para cada ângulo. Ou seja, para cada d (distância) fixo teremos 4 matrizes de co-ocorrência diferentes, uma para cada ângulo, 0^{o}, 45^{o}, 90^{o} e 135^{o}. Veja a seguir as Figuras 3, 4, 5 e 6 que ilustram a construção completa da matriz de co-ocorrência para todas as combinações de níveis de cinza possíveis e seus respectivos ângulos, e a Figura 7 com as matrizes de co-ocorrência normalizadas. Observe nas Figuras 3, 4, 5 e 6 a direção do ângulo utilizado para o cálculo de cada matriz, note que o resultado de cada uma delas são diferentes mesmo utilizando a mesma distância.

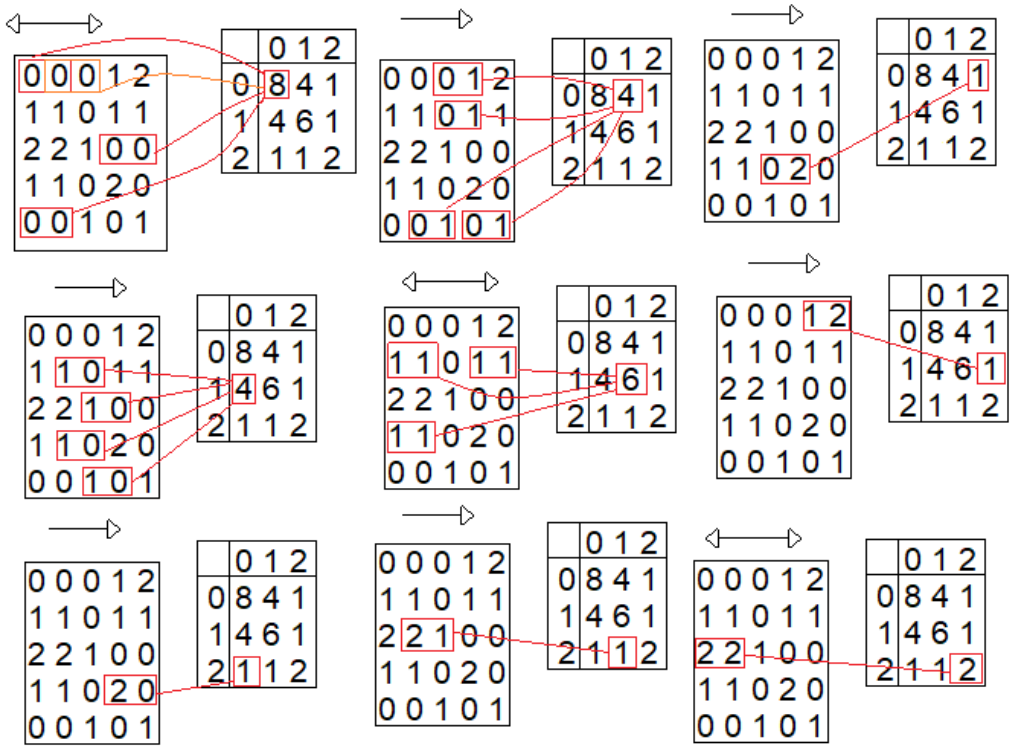


Figura 3: Construção da matriz de co-ocorrência para o ângulo 0^{o} e distância 1.

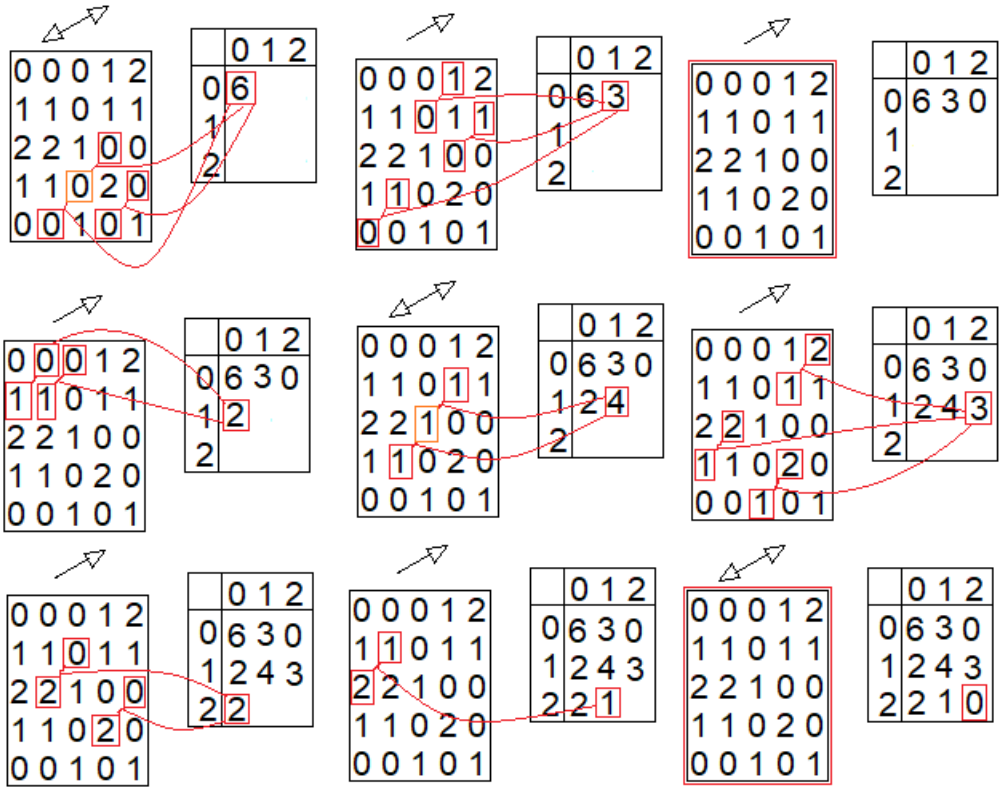


Figura 4: Construção da matriz de co-ocorrência para o ângulo 45^{o} e distância 1.

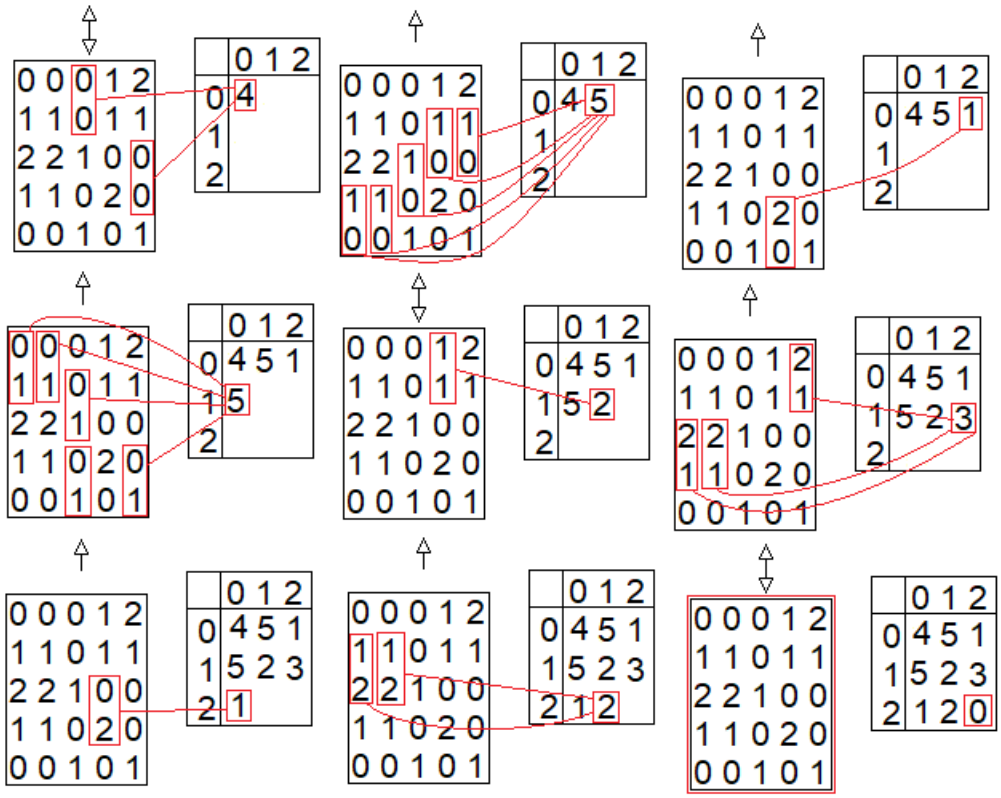


Figura 5: Construção da matriz de co-ocorrência para o ângulo 90º e distância 1.

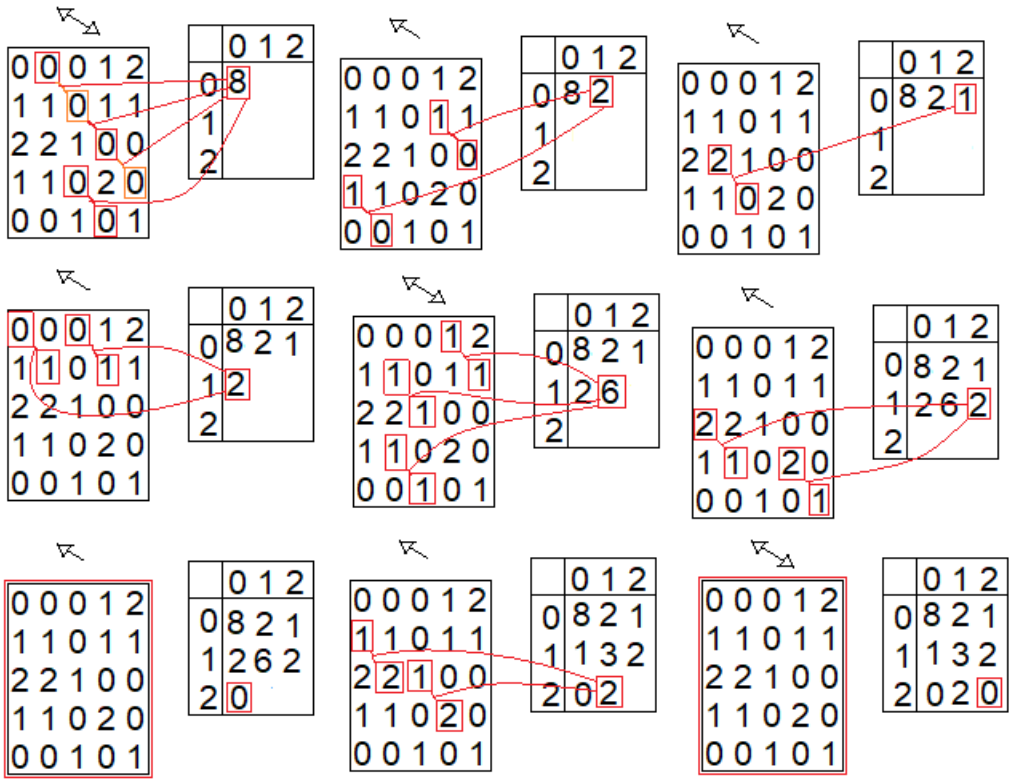


Figura 6: Construção da matriz de co-ocorrência para o ângulo 135º e distância 1.

Assim, finalizamos as matrizes na Figura 7 dividindo cada elemento da posição (i,j) pela soma de todos os elementos da matriz, ou seja, seria calcular a probabilidade P(i,j) $= \frac{M_{i,j}}{\sum_{i,j=1}^{N}(M_{i,j})}$ descrito anteriormente, onde i é o número da linha e j é o número de coluna, P(i,j) é a probabilidade da célula (i,j), $M_{i,j}$ é o conteúdo da célula (i,j), N é o número de linhas ou colunas de uma matriz quadrada. Considerando neste exemplo, uma distância 1 e cada um dos ângulos 0º, 45º, 90º e 135º.

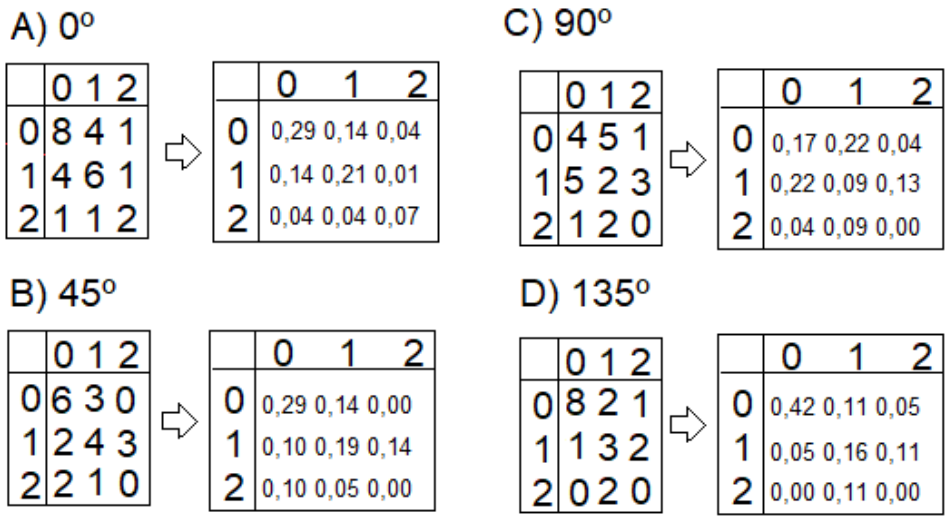


Figura 7: Resultado da construção da matriz de co-ocorrência para cada ângulo com as matrizes normalizadas.

2.3 Descritores de Haralick

A utilização de características texturais pode ser aplicada para a classificação de imagens. A matriz de co-ocorrência de níveis de cinza é a base dos descritores de Haralick, que foram propostos por Robert Haralick e colaboradores. Analisando seu artigo, Haralick, Shanmugam e Dinstein (1973) as características texturais podem ser extraídas de regiões suspeitas, esses descritores são utilizados para extrair recursos de imagens médicas, incluindo mamografias, que podem auxiliar na detecção precoce do câncer de mama.

Os descritores de Haralick têm algumas vantagens, como a facilidade de cálculo a partir da matriz de co-ocorrência de níveis de cinza, sendo uma medida objetiva e direta da distribuição de intensidade da imagem. Além disso, eles são robustos a variações de iluminação e contraste nas imagens, tornando-os úteis em imagens médicas, onde a iluminação e a qualidade da imagem podem variar. Os descritores de Haralick também conseguem capturar informações de textura de alta ordem, que podem ser difíceis de

descrever com outros descritores.

Entretanto, as desvantagens dos descritores de Haralick incluem a sensibilidade ao tamanho (distância estipulada) e orientação (ângulo) da janela de textura utilizada para calcular a matriz de co-ocorrência de níveis de cinza. Isso significa que diferentes tamanhos e orientações de janelas podem resultar em diferentes descritores de textura. Outro ponto é que eles podem ser afetados pelo nível de ruído na imagem, o que pode levar a descritores de textura menos precisos. Ou seja, é mais vantajoso utilizar esses descritores para partes específicas da imagem e não para a imagem inteira porque os descritores deixaram de ser representativos.

Além disso, eles podem não capturar todas as informações de textura relevantes na imagem, principalmente em casos de texturas complexas. Portanto, a aplicação dos descritores de Haralick na detecção de câncer de mama em mamografias pode ser útil, mas devemos ter em mente suas vantagens e desvantagens.

A seguir teremos as fórmulas dos descritores de textura de Haralick. Veja abaixo o significado de algumas expressões que serão simplificadas nas equações dos descritores. Para melhor visualização e entendimento.

Notações:

$\boldsymbol{P(i,j)}$ - matriz de frequências relativas (matriz de co-ocorrência normalizada) com as quais duas células (níveis de cinza naquelas posições) de resolução vizinhas separadas pela distância d ocorrem na imagem.

$\boldsymbol{R = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i,j)}$ - constante de normalização, ou seja, número de pares de células de resolução vizinhos usados na computação de uma matriz de dependência espacial de tom de cinza específico e a soma de todos os elementos da matriz de frequência.

$\boldsymbol{p(i,j) = \frac{P(i,j)}{R}}$ - (i,j)-ésima entrada em uma matriz de dependência espacial normalizada em tons de cinza.

$\boldsymbol{p_x(i) = \sum_{j=1}^{N_g} P(i,j)}$ - i-ésima entrada na matriz de probabilidade marginal obtida pela soma das linhas de p(i,j).

$\boldsymbol{N_g}$ - Número de níveis de cinza distintos na imagem.

$\boldsymbol{p_y(j) = \sum_{i=1}^{N_g} P(i,j)}$ - j-ésima entrada na matriz de probabilidade marginal obtida pela soma das colunas de p(i,j).

As medidas $p_{x+y}(k)$ e $p_{x-y}(k)$ são usadas para avaliar a dependência entre os níveis de cinza dos píxeis vizinhos. $p_{x+y}(k)$ é a soma das probabilidades conjuntas que têm a mesma soma de índices de linha e coluna, enquanto $p_{x-y}(k)$ é a soma das probabilidades conjuntas que têm a mesma diferença de índices de linha e coluna.

$$\boldsymbol{p_{x+y}(k)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N_g}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{N_g}} \boldsymbol{\delta_{i+j,k} P(i,j)} \text{ , k = 2,3,...,}2N_g.$$

$$\boldsymbol{p_{x-y}(k)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N_g}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{N_g}} \boldsymbol{\delta_{i-j,k} P(i,j)} \text{ , k = 0,1,...,}N_g - 1.$$

Onde a função delta de Kronecher é definada como $\delta_{m,n} = \begin{cases} 1, & \text{quando} \quad \text{m = n.} \\ 0, & \text{quando} \quad \text{m} \neq \text{n.} \end{cases}$

As medidas **H(X)**, **H(Y)**, **H(X,Y)**, **H1(X,Y)** e **H2(X,Y)** são utilizadas para quantificar a informação mútua entre duas variáveis aleatórias, **X** e **Y**. **H(X)** e **H(Y)** medem a entropia marginal de **X** e **Y**, respectivamente, enquanto **H(X,Y)** mede a entropia conjunta de **X** e **Y**. **H1(X,Y)** e **H2(X,Y)** são medidas de informação mútua que podem ser usadas para avaliar a dependência entre **X** e **Y**.

$$\mathbf{H(X)} = -\sum_{\boldsymbol{i}} \boldsymbol{p_x(i) log(p_x(i))}$$

$$\mathbf{H(Y)} = -\sum_{\boldsymbol{j}} \boldsymbol{p_y(j) log(p_y(j))}$$

$$\mathbf{H(X,Y)} = -\sum_{\boldsymbol{i}} \sum_{\boldsymbol{j}} \boldsymbol{p(i,j) log(p(i,j))}$$

$$\mathbf{H1(X,Y)} = -\sum_{\boldsymbol{i}} \sum_{\boldsymbol{j}} \boldsymbol{p(i,j) log(p_x(i) p_y(j))}$$

$$\mathbf{H2(X,Y)} = -\sum_{\boldsymbol{i}} \sum_{\boldsymbol{j}} \boldsymbol{p_x(i) p_y(j) log(p_x(i) p_y(j))}$$

Haralick, Shanmugam e Dinstein (1973) sugere um conjunto de 28 características texturais que podem ser extraídas das matrizes de co-ocorrência (matrizes com dependência espacial) de tons de cinza. O último descritor de correlação máxima possui uma instabilidade computacional muito grande, não sendo calculado pelo pacote utilizado, ele não será apresentado. A seguir vejam as equações que definem esses recursos de textura descrito por Haralick em seu artigo Haralick, Shanmugam e Dinstein (1973).

- **Segundo Momento Angular [asm]:** é uma medida de homogeneidade para analisar imagens. O Segundo momento angular é alto quando a imagem tem uma homogeneidade ótima ou quando os píxeis são muito semelhantes Kumar e Sreekumar

(2014).

$$f_1 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(\frac{P(i,j)}{R}\right)^2 = \sum_i\sum_j P(i,j)^2. \tag{2.1}$$

- **Contraste [con]:** O recurso de contrate é um momento de diferença da matriz P e é uma medida do contraste ou da quantidade de variações locais presentes na imagem.

$$f_2 = \sum_{i=0}^{N_g-1} k^2 \left\{\sum_{j=1}^{N_g}\sum_{j=1}^{N_g}\delta_{|i-j|}P(i,j)\right\} = \sum_{k=0}^{N_g-1} k^2 P_{x-y}(k). \tag{2.2}$$

- **Correlação [cor]:** A correlação mede a dependência linear dos níveis de cinza dos píxeis vizinhos Kumar e Sreekumar (2014).

$$f_3 = \frac{\sum_{i=0}^{N_g}\sum_{j=1}^{N_g}(i,j)P(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y} \tag{2.3}$$

Sendo μ_x, μ_y, σ_x e σ_y são medias e os desvios padrões de p_x e p_y.

- **Soma dos Quadrados: Variância [var]**: Mede a variância dos níveis de cinza dos píxeis.

$$f_4 = \sum_{i=0}^{N_g}\sum_{j=1}^{N_g}(i-\mu)^2 P(i,j) \tag{2.4}$$

- **Momento de Diferença Inversa [idm]:** O Momento de Diferença Inversa (IDM) é a homogeneidade local. É alto quando o nível de cinza local é uniforme e o GLCM inverso é alto Kumar e Sreekumar (2014).

$$f_5 = \sum_{i=0}^{N_g}\sum_{j=1}^{N_g}\frac{1}{1+(i-j)^2}P(i,j) \tag{2.5}$$

- **Média de soma [sav]:** Mede a média de soma.

$$f_6 = \sum_{i=2}^{2N_g} i \cdot P_{x+y}(i) \tag{2.6}$$

- **Variância da soma [sva]:** Mede a variância da soma.

$$f_7 = \sum_{i=2}^{2N_g}(i-f_8)^2 P_{x+y}(i) \tag{2.7}$$

- **Entropia da soma [sen]:** Como algumas das probabilidades podem ser zero, e

log(0) não está definido, recomenda-se que o termo log(p + ϵ) seja usado no lugar de log(p) em cálculos de entropia, onde ϵ é uma constante positiva arbitrariamente pequena. O pacote EBImage usa um valor de 1E-7 para ϵ e calcula logaritmos de base 2.

$$f_8 = -\sum_{i=2}^{2N_g} P_{x+y}(i) log(P_{x+y}(i)) \tag{2.8}$$

- **Entropia [ent]:** Fórmulas de Miyamoto e Merryman (2005) esclarecem que $f_9 = H(X, Y)$ no artigo original de Haralick.

$$f_9 = -\sum_{i=2}^{2N_g}\sum_{i=2}^{2N_g} P(i, j) log(P(i, j)) \tag{2.9}$$

- **Diferença de variância [dva]:** Mede a diferença entre as médias dos níveis de cinza dos píxeis.

$$f_{10} = \sum_{i=0}^{N_g-1} i^2 P_{x-y}(i) \tag{2.10}$$

- **Diferença de entropia [den]**: A diferença de entropia é uma medida da variação da entropia entre duas imagens. Ela é calculada subtraindo a entropia da imagem original da entropia da imagem modificada.

$$f_{11} = -\sum_{i=0}^{N_g-1} p_{x-y}(i) log(p_{x-y}(i)) \tag{2.11}$$

- **Medidas de correlação [f12 e f13]**: A medida f12 é a soma das probabilidades conjuntas que têm a mesma soma de índices de linha e coluna, enquanto a medida f13 é a soma das probabilidades conjuntas que têm a mesma diferença de índices de linha e coluna.

$$f_{12} = \frac{f_9 - H1(X, Y)}{max(H(X), H(Y))} \tag{2.12}$$

$$f_{13} = [1 - e^{-2(H2(X,Y) - f_9)}]^2 \tag{2.13}$$

Loewke (2013) realizou uma análise usando o intervalo e a média dos 13 descritores, em média nas quatro direções, totalizando 26 medidas.

2.4 Técnicas de redução e tratamento do banco de dados

Nesta seção, serão apresentadas as técnicas de redução e tratamento de banco de dados utilizadas nos modelos de aprendizado de máquina supervisionados. Em grandes bancos de dados, é fundamental empregar técnicas eficazes de redução e tratamento de informações, uma vez que tais bancos podem conter informações desnecessárias ou redundantes, comprometendo a precisão do modelo criado. Desse modo, o uso dessas técnicas pode facilitar a tomada de decisão e contribuir para o sucesso de projetos de pesquisa, além de preparar o banco de dados de forma mais eficiente e precisa para o modelo a ser desenvolvido.

- Padronização dos dados: é um processo de transformação dos dados para terem média zero e desvio padrão igual a 1. Isso permite comparar variáveis que possuem escalas diferentes e para garantir que variáveis com maiores magnitudes não tenham mais peso na análise do que as demais.

- Variância zero ou quase zero: ocorre quando uma variável possui uma variância muito pequena ou igual a zero. Essas variáveis não fornecem informações úteis para a análise e podem ser removidas do conjunto de dados (KUHN, 2015).

- Variáveis correlacionadas: A correlação entre variáveis é uma medida estatística que busca identificar a relação entre duas ou mais variáveis. Ela indica se existe uma associação entre as variáveis e qual é a direção e a intensidade dessa relação. O coeficiente de correlação varia de -1 a 1, sendo que valores próximos de -1 indicam uma correlação negativa, valores próximos de 1 indicam uma correlação positiva e valores próximos de 0 indicam que não há correlação entre as variáveis. A correlação é importante porque nos permite entender como diferentes variáveis se relacionam entre si. Isso pode levar a problemas na análise, como a multicolinearidade, que pode afetar a precisão dos resultados. É importante identificar e tratar variáveis correlacionadas antes da análise (KUHN, 2015).

- Rebalanceamento: é uma técnica utilizada em conjuntos de dados desbalanceados, onde há uma grande diferença entre o número de observações em cada classe. Isso pode levar a uma falta de representatividade das classes minoritárias na análise e afetar a precisão dos resultados Amorim (2022). Neste trabalho foram utilizadas essas duas abordagens de separação da base de dados:

– *Downsampling*: conhecido como subamostragem ou *undersampling*, é uma técnica de processamento de dados usada para reduzir o tamanho de um conjunto de dados. Isso é feito selecionando aleatoriamente uma amostra de dados da classe majoritária para criar um novo subconjunto de dados. A importância do *downsampling* reside em sua capacidade de lidar com conjuntos de dados desbalanceados. O *downsampling* ajuda a equilibrar (balancear) o conjunto de dados conforme a classe de interesse, permitindo que o modelo aprenda de maneira mais eficaz a partir de ambas as classes. No entanto, é importante notar que o *downsampling* pode levar à perda de informações, pois algumas amostras da classe majoritária são removidas[1].

– Separação de 2 para 1: significa que, para cada duas amostras da classe majoritária, há uma amostra da classe minoritária. Para bases muito desbalanceadas, uma forma de não se perder muita informação, às vezes é selecionando aleatoriamente um pouco mais de amostras da classe majoritária. Isso pode ser utilizado quando os modelos não conseguem aprender. Mesmo com a base balanceada ou totalmente desbalanceada com o tamanho original, eles podem se tornar enviesado para a classe majoritária. Assim pode-se utilizar outras separações como 4 para 1, 3 para 1 entre outras. Neste trabalho utilizaremos a separação 2 para 1.

2.5 Etapas para a modelagem com aprendizado de máquinas.

O aprendizado de máquinas é uma área da inteligência artificial que permite que as máquinas aprendam mediante um conjunto de dados, em vez de serem explicitamente programadas para realizar uma tarefa específica. Ele é usado para analisar dados, identificar padrões e fazer previsões ou classificações com base nesses padrões. O aprendizado de máquinas pode ser dividido em três categorias, supervisionado, não supervisionado e por reforço. Neste trabalho adotaremos os modelos de aprendizado supervisionados, onde os dados de treinamento que serão rotulados, serão utilizados para a criação dos modelos que serão treinados para estimar rótulos para novos dados, dados nunca vistos. Antes de apresentar os modelos que serão utilizados, a seguir apresentamos conceitos fundamentais para utilização dos modelos.

[1]https://ichi.pro/pt/5-tecnicas-para-trabalhar-com-dados-desequilibrados-no-aprendizado-de-maquina-141301962363660

2.5.1 Separação, treino e teste.

A separação dos dados em conjuntos de treino e teste é uma técnica fundamental no processo de treinamento e avaliação de modelos de aprendizado de máquina. A ideia básica é utilizar uma parte dos dados disponíveis para treinar o modelo e outra parte para avaliá-lo. O conjunto de treino é usado para ajustar os parâmetros do modelo, enquanto o conjunto de teste é utilizado para avaliar a capacidade de generalização do modelo para novos dados.

A escolha adequada da proporção entre os conjuntos de treino e teste é importante para evitar problemas de *overfitting* ou *underfitting*. O *overfitting* ocorre quando o modelo se ajusta muito bem aos dados de treino, mas não apresenta resultados bons para novos dados, enquanto o *underfitting* ocorre quando o modelo é muito simples e não consegue capturar a complexidade dos dados. Além disso, é importante garantir que a seleção dos dados para treino e teste seja feita de forma aleatória e representativa da população a ser estudada. A seguir temos algumas maneiras de separação da base em treino e teste (MONARD; BARANAUSKAS, 2003).

- Separação da base: os dados são divididos em duas partes, geralmente 80% para treinamento e 20% para teste. Essa é uma das estratégias mais simples e mais usadas, mas pode resultar em uma variação significativa no desempenho do modelo, dependendo da forma como os dados são divididos (MONARD; BARANAUSKAS, 2003).

- *Cross-validation*: os dados são divididos em k partes iguais, sendo que uma parte é usada para teste e as outras k-1 partes são usadas para treinamento. Esse processo é repetido k vezes, alternando as partes usadas para teste e treinamento. Isso ajuda a mitigar a variação do desempenho do modelo (MONARD; BARANAUSKAS, 2003).

2.5.2 Tipos de erros

Na avaliação de modelos de aprendizado de máquina supervisionado, exitem alguns tipos de erros que se referem a duas medidas diferentes de desempenho de um modelo. O erro dentro da amostra (erro *in sample*) e o erro fora da amostra (erro *out of sample*) Hastie et al. (2009).

Existem três tipos de erros que podem ocorrer:

- Erro de viés (*Bias error*): ocorre quando o modelo não consegue capturar a relação entre as variáveis de entrada e de saída, ou seja, quando o modelo é muito simples e não consegue se ajustar aos dados de treinamento.

- Erro de variância (*Variance error*): ocorre quando o modelo é muito complexo e se ajusta muito bem aos dados de treinamento, mas não generaliza bem para dados novos, ou seja, apresenta uma baixa capacidade de generalização.

- Erro irredutível (*Irreducible error*): é um erro inerente aos dados em si, ou seja, é causado por fatores externos que não podem ser controlados. Esse erro não pode ser reduzido por nenhum modelo e é inevitável.

O erro dentro da mostra é calculado com base nos dados de treinamento utilizados para construir o modelo, também conhecido como erro de treinamento, refere-se à taxa de erro obtida ao avaliar o desempenho do modelo nos dados de treinamento utilizados para treiná-lo. Essa medida indica quão bom o modelo se ajusta aos dados de treinamento.

O erro fora da amostra é calculado usando dados que não foram usados durante o treinamento do modelo. Também conhecido como erro de teste, é calculado ao avaliar o desempenho do modelo em dados não utilizados no treinamento, ou seja, em novos exemplos que o modelo nunca viu antes. Esse erro é uma medida crítica para avaliar a capacidade de generalização do modelo, ou seja, quão bom ele pode fazer previsões precisas em dados não vistos anteriormente.

Geralmente, espera-se que o erro dentro da amostra seja menor do que o erro fora da amostra. Se o erro dentro da amostra for muito baixo, mas o erro fora da amostra for alto, pode ser um indicativo de que o modelo está sofrendo de *overfitting*, ou seja, está se ajustando excessivamente aos dados de treinamento e não está generalizando bem para novos dados. É importante considerar tanto o erro dentro da amostra quanto o erro fora da amostra ao avaliar o desempenho de um modelo de aprendizado de máquina, a fim de obter uma visão abrangente de sua capacidade de fazer previsões precisas em diferentes conjuntos de dados.

2.5.3 Modelos supervisionados

Nesta seção, serão abordados algumas técnicas de aprendizado de máquinas que utilizam dados rotulados para treinar um modelo a fim de fazer previsões em novos dados. É o tipo mais comum de aprendizado de máquina usado em uma ampla gama de aplicativos,

desde reconhecimento de fala até detecção de fraudes em transações financeiras. Durante o processo de treinamento, o modelo aprende a relacionar as entradas aos rótulos correspondentes. Após o treinamento, o modelo consegue fazer previsões precisas em dados não vistos anteriormente, avaliado por métricas como acurácia, sensibilidade e especificidade da matriz de confusão.

A seguir apresentamos um breve resumo sobre os métodos que serão utilizados neste trabalho. Esses métodos não serão detalhados visto que o foco do estudo é destacar o uso da segmentação, e todo o pre-processamento feito no banco de dados DDSM para a construção da base de dados que será discutido na sessão de pré-processamento.

- Floresta aleatória: é um algoritmo de *bagging* (BASTOS; NASCIMENTO; LAURETTO, 2013) que usa várias árvores de decisão para fazer previsões. Cada árvore é treinada em uma amostra aleatória dos dados e as previsões são feitas pela média das previsões de todas as árvores.

- Máquina de Vetores de Suporte (SVM): é um algoritmo de aprendizado supervisionado que analisa os dados e os divide em diferentes grupos, de acordo com seus padrões, para assim classificar as observações Semolini et al. (2002). O SVM visa encontrar o hiperplano que melhor separa as classes. A classificação foi feita com base no lado do hiperplano no qual a observação se encontra.

- *Adaboost*: é um método que usa várias árvores de decisão, mas ao contrário da Floresta Aleatória (veja mais em Bastos, Nascimento e Lauretto (2013)), ele treina as árvores sequencialmente e dá mais peso aos exemplos que foram bem classificados pelas árvores anteriores, ou seja, em cada iteração, um modelo fraco é criado a partir dos dados de treinamento e os erros são ponderados para se concentrar em exemplos classificados incorretamente. O modelo fraco é então adicionado ao modelo anterior com um peso, formando um modelo mais forte. A técnica de *Adaboost* é útil para dados com poucas variáveis explicativas Vezhnevets e Vezhnevets (2005). O AdaBoost combina vários modelos fracos para criar um modelo forte. A classificação foi feita com base na soma ponderada das previsões dos modelos fracos.

- Regressão logística: é um modelo linear generalizado (GLM) usado para prever a probabilidade de uma variável Bernoulli com base em uma ou mais variáveis explicativas usando a função de ligação logística para fazer a previsão. Durante a modelagem, a função logística é usada para calcular a probabilidade da resposta pertencer a uma das categorias. A regressão logística é útil para dados com

uma ou mais variáveis explicativas e é frequentemente usada para análise de dados biomédicos(MINUSSI; DAMACENA; JR, 2002). A classificação foi feita com base na probabilidade prevista pelo modelo, e essa probabilidade é binariada para os valores de 0 ou 1.

- K vizinho mais próximos (KNN): é um classificador que utiliza a ideia de aprendizado por analogia. Ele determina a classe de uma instância de entrada com base nas classes de instâncias similares na base de treinamento. É um dos métodos mais simples e eficazes de classificação em aprendizado de máquina. O KNN encontra os k vizinhos mais próximos da instância de entrada com base em uma medida de distância, como a distância euclidiana.[2] Esses vizinhos são selecionados na base de treinamento, onde cada instância possui atributos e uma classe correspondente. Após identificar os k vizinhos mais próximos, o KNN atribui a classe mais frequente entre eles à instância de entrada Hastie et al. (2009).

- XGboost: é um algoritmo de boosting que usa árvores de decisão como seus modelos base. Ele usa gradient descent para minimizar a função de perda e adicionar árvores de decisão fracas para melhorar o modelo (CHEN et al., 2015). A classificação foi feita com base na soma das previsões de todas as árvores no modelo.

2.5.4 Medidas de qualidade de ajuste.

As medidas de qualidade de ajuste são utilizadas para avaliar a performance dos modelos de aprendizado de máquina supervisionados em dados de teste. Elas são importantes para determinar a eficácia do modelo em prever novos dados e identificar possíveis problemas de *overfitting* ou *underfitting*. Existem várias medidas de qualidade de ajuste disponíveis, cada uma com sua própria interpretação e uso apropriado. A seguir apresentamos as medidas que serão utilizadas neste trabalho.

Matriz de confusão

A matriz de confusão na Tabela 3 é uma ferramenta fundamental na avaliação de modelos de aprendizado de máquina. Ela permite analisar visualmente a performance do modelo, mostrando a frequência com que os dados foram classificados correta ou incorretamente Hastie et al. (2009). A matriz é composta por 4 elementos:

[2] A distância euclidiana $D(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$, é uma medida utilizada para calcular a distância entre dois pontos em um espaço euclidiano, como, por exemplo, em um gráfico cartesiano. Essa medida é amplamente utilizada em algoritmos de aprendizado de máquina, como o KNN, para determinar a proximidade entre instâncias e tomar decisões com base nessas distâncias.

- Verdadeiros positivos (VP): são os casos em que o modelo classificou corretamente a presença da classe de interesse.
- Verdadeiros negativos (VN): correspondem aos casos em que o modelo classificou corretamente a ausência da classe de interesse.
- Falsos positivos (FP): representam os casos em que o modelo classificou incorretamente a presença da classe de interesse.
- Falsos negativos (FN): são os casos em que o modelo classificou incorretamente a ausência da classe de interesse.

Tabela 3: Matriz de confusão.

	Classe real 0	Classe real 1
Previsão 0	VN	FN
Previsão 1	FP	VP

A partir desses elementos, podemos calcular diversas medidas de qualidade de ajuste, como:

- Acurácia = $\frac{VP+VN}{VP+VN+FP+FN}$, proporção de observações corretamente classificadas em relação ao total de observações.
- Especificidade = $\frac{VN}{VN+FN}$, é a proporção de verdadeiros negativos em relação ao total de casos negativos, indicando a capacidade do modelo de identificar corretamente a ausência da classe de interesse.
- Sensibilidade = $\frac{VP}{VP+FN}$ proporção de verdadeiros positivos (VP) em relação à soma de verdadeiros positivos (VP) e falsos negativos (FN).

2.5.5 Pré-processamento da base de dados.

Como este trabalho será realizado no software R Core Team (2014), a imagem da mamografia será analisada utilizando o pacote EBImage. De acordo com o criador Pau et al. (2010) esse pacote foi criado para o processamento de imagens biológicas, ele é muito utilizado em pesquisas com biologia celular, microbiologia e neurociência. Este pacote fornece funções para leitura, escrita, visualização, segmentação de imagem e extração de descritores celulares quantitativos, útil para processamento de sinais, modelagem estatísticas, aprendizado de máquina e visualização de dados.

Os descritores de Haralick, Shanmugam e Dinstein (1973) serão calculados a partir de objetos que serão classificados como suspeitos ou não utilizando a imagem toda, dado que foi comprovado que esses descritores são sensíveis a uma janela grande de textura. A classificação destes objetos se dará pela utilização de algumas operações morfológicas preparando a imagem para ser analisada e rotulando conjuntos de píxeis conectados ou semelhantes como um objeto.

As funções morfológicas [3] posicionam o centro do elemento estruturante sobre cada píxel na imagem de entrada. Para píxeis próximos à borda de uma imagem, partes da vizinhança definidas pelo elemento estruturante podem se estender além da borda da imagem. Nesse caso, um valor é atribuído a esses píxeis indefinidos, como se a imagem fosse preenchida com linhas e colunas adicionais. O valor desses píxeis de preenchimento varia para operações de dilatação e erosão. Para dilatação, os píxeis além da borda da imagem recebem o valor mínimo fornecido pelo tipo de dados, que no caso de imagens binárias é equivalente a definir como plano de fundo. Para a erosão, os píxeis além da borda da imagem recebem o valor máximo fornecido pelo tipo de dados, que no caso de imagens binárias é equivalente a definir como primeiro plano.

Segue abaixo algumas definições importantes que serão utilizadas, para realização do trabalho proposto.

- Otsu: o método de limiar de (OTSU, 1979) que permite executar automaticamente o limite de imagem baseado em clusterização. O algoritmo assume que a distribuição das intensidades dos píxeis da imagem segue um histograma bimodal e separa esses píxeis em duas classes (por exemplo, primeiro plano e fundo). O valor limite ideal é determinado minimizando a variância intraclasse combinada. O valor limite é calculado para cada quadro de imagem separadamente, resultando em um vetor de saída de comprimento igual ao número total de quadros na imagem.

- Dilatação (*dilate*): aplica a máscara posicionando seu centro sobre cada píxel da imagem, o valor de saída do píxel é o valor máximo coberto pela máscara. No caso de imagens binárias, isso é equivalente a colocar a máscara sobre cada píxel de fundo e defini-la para primeiro plano se qualquer um dos píxeis cobertos pela máscara for de primeiro plano (URBACH; WILKINSON, 2007).

- Erosão (*erode*): aplica a máscara posicionando seu centro sobre cada píxel da ima-

[3] Operações morfológicas são implementadas utilizando o eficiente algoritmo de Urbach e Wilkinson (2007). Seu tempo de computação necessário é independente do conteúdo da imagem e do número de níveis de cinza usados.

gem, o valor de saída do píxel é o valor mínimo de coberto pela máscara. No caso de imagens binárias, isso é equivalente a colocar a máscara sobre cada píxel de primeiro plano e defini-la como plano de fundo se qualquer um dos píxeis cobertos pela máscara for do plano de fundo Urbach e Wilkinson (2007).

- Abertura e fechamento (*opening*): a abertura é uma erosão seguida de uma dilatação e o fechamento é uma dilatação seguida de uma erosão (URBACH; WILKINSON, 2007);

- Segmentação binária (bwlabel): rotula objetos conectados (conjuntos conectados) em uma imagem binária. Todos os píxeis para cada conjunto conectado de píxeis de primeiro plano (diferente de zero) em x são definidos como um inteiro crescente único, começando em 1. Portanto, Max(x) fornece o número de objetos conectados em x (PAU et al., 2010);

- Extração de medidas estatísticas: esse descritor calcular estatísticas básicas como média, desvio padrão, o valor mínimo e máximo referente a intensidade dos píxeis em uma imagem (PAU et al., 2010);

- Extração de medidas de áreas: esse descritor extrair informações presente na imagem referente a um objeto, tais como raio, perímetro e área (PAU et al., 2010);

- Extração de medidas de momentos: este descritor descreve medidas relacionadas a distribuição espacial de intensidade de uma imagem, eles são usados para descrever orientação, tamanho e forma de objetos presente em uma imagem (PAU et al., 2010);

- Extração de medidas de textura: esse descritor extrai medidas com base em uma matriz de co-ocorrência de níveis de cinza, essa matriz indica quantas vezes cada par de níveis de cinza aparece em uma determinada relação espacial na imagem. Essas medidas calculadas são os descritores de textura de Haralick, sendo medidas objetivas sobre a distribuição de intensidades na imagem (HARALICK; SHANMUGAM; DINSTEIN, 1973).

Esses descritores são úteis para quantificar diferentes aspectos de uma imagem e podem ser usados para extrair informações relevantes para várias aplicações em processamento de imagens e análise de dados de imagem.

Veja na ilustração da Figura 8 o pré-processamento da imagem. Na leitura da imagem os píxeis serão definidos valores no intervalo de 0 a 1. Após isso, a imagem será passada

pelo método de limiarização de Otsu adaptado, ou seja, a imagem será analisada em janelas de tamanho de 35×35, que distinguirá grupos que representam o primeiro plano e o plano de fundo, logo em seguida serão retirados ruídos da imagem com a operação morfológica *opening*. Assim, por último, essa imagem passará por uma segmentação binária, para encontrar conjunto de píxeis conectados (semelhantes) que serão denotados como objetos.

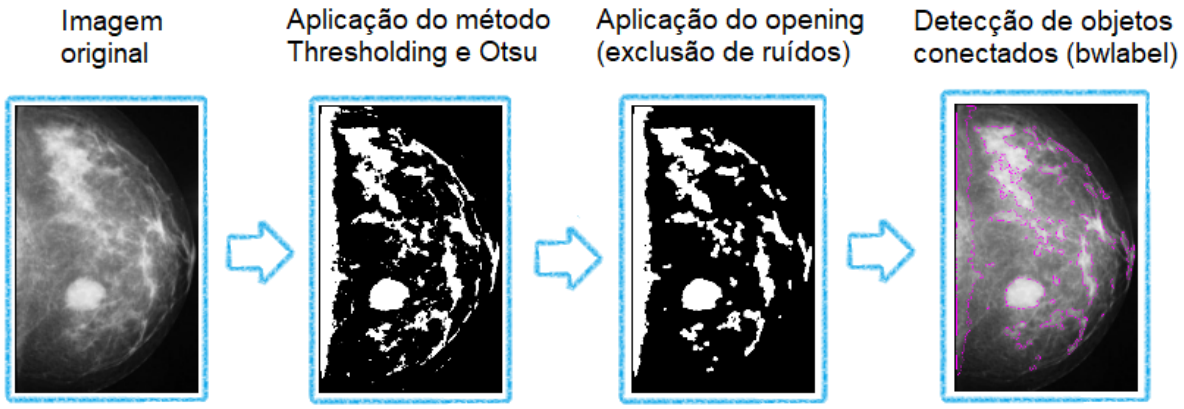


Figura 8: Pré-processamento

Destes objetos presentes na Figura 9 serão extraídos medidas estatísticas, de texturas e informações para esses objetos, veja a Tabela 4 com as medidas extraídas.

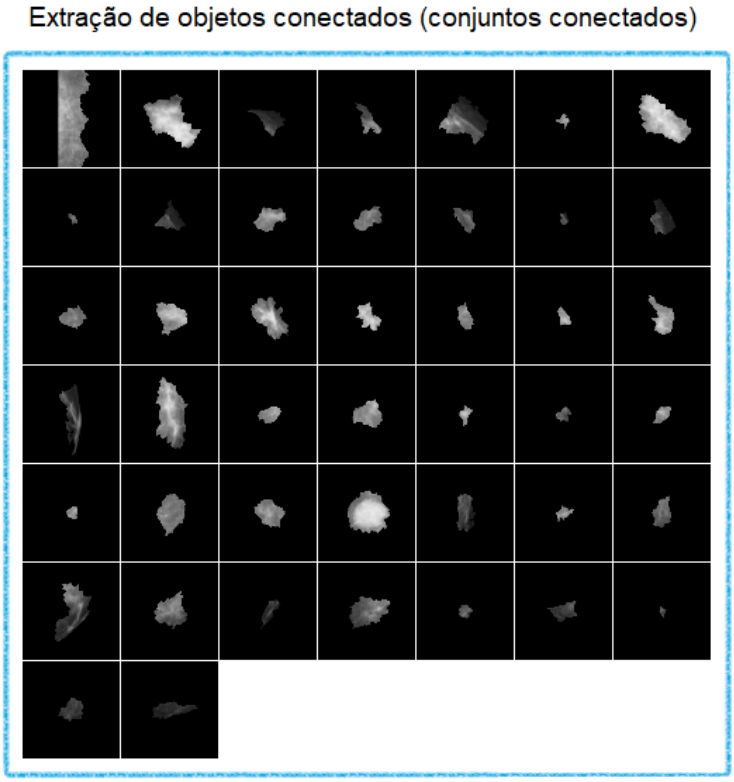


Figura 9: Exemplos dos objetos extraídos que serão quantificados.

Tabela 4: Medidas extraídas dos objetos.

Descritores de Textura	Medidas de Momento	Medidas Estatísticas	Medidas de área
Segundo Momento Angular [asm]	Eixo Maior Elíptico	Média	Raio Médio
Contraste [con]	Ângulo	Desvio Padrão	Desvio Padrão do Raio.
Soma dos Quadrados: Variância [var]	Excentricidade	Desvio Absoluto Médio	Raio Mínimo
Momento de Diferença Inversa [idm]	-	Percentil 1%	Raio Máximo
Média de Soma [sav]	-	Percentil 5%	-
Variância da soma [sva]	-	Percentil 50%	-
Entropia da Soma [sen]	-	Percentil 95%	-
Entropia [ent]	-	Percentil 99%	-
Correlação [cor]	-	-	-
Diferença de ariância [dva]	-	-	-
Diferença de entropia [den]	-	-	-
Medidas de correlação [f12]	-	-	-
Medidas de correlação [f13]	-	-	-

Após as extrações de medidas desses objetos, cada objeto será classificado com um rótulo, 1 se aquele objeto é um tumor ou 0 caso aquele objeto não seja um tumor. Depois que a base estiver completa será aplicado a metodologia de separação de treino de teste, onde 80% da base de dados serão designados para treinamento e 20% dos dados

serão usados para testar o modelo, veja a ilustração na Figura 10. Em seguida, após a separação serão aplicados um novo pré-processamento, mas na base de dados com as medidas extraídas. Na base de treino, os dados serão normalizados, serão excluídas variáveis correlacionadas, variáveis com variância zero ou quase zero, variáveis com dependência linear. Variáveis com essas características podem influenciar negativamente na previsão do modelo. Todos esses passos serão aplicados nos dados de teste, menos o rebalanceamento que será aplicado nos dados de treino somente se a base de dados for muito desbalanceada. O balanceamento dos dados e alguns detalhes da base de dados serão discutidos na sessão dos resultados.

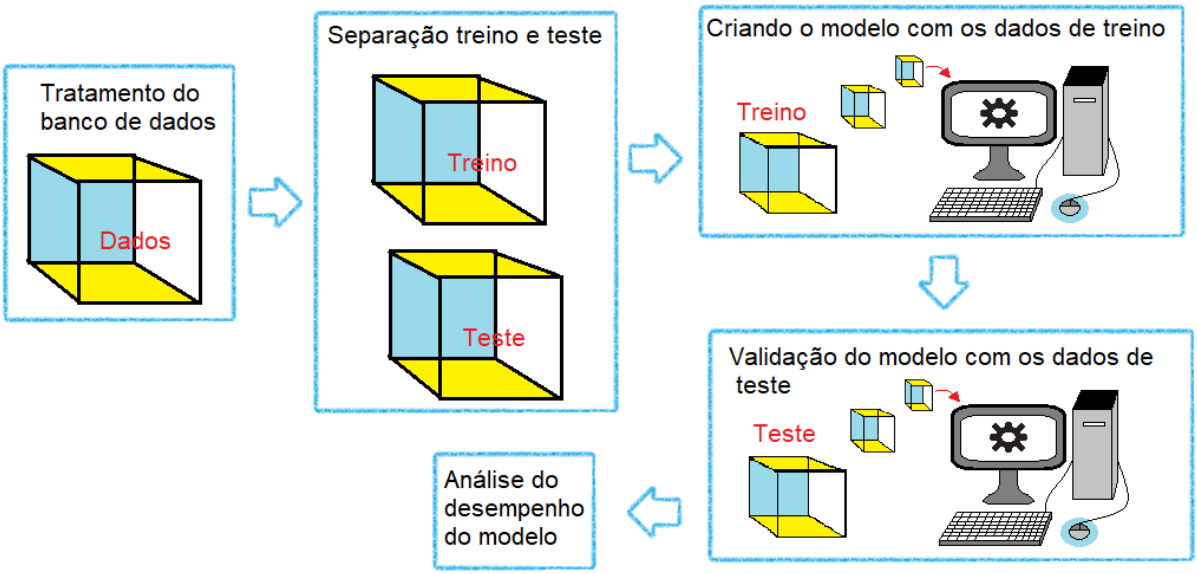


Figura 10: Divisão em treino e teste.

2.6 Base de dados construída para a analise.

2.6.1 Tipos de exames.

Analisaremos cada imagem de maneira individual e quantitativamente, com os recursos extraídos conforme a Seção 2.5.5, criando assim dois modelos de classificação, porque há imagens do tipo Obliquidade Médio-Lateral (MLO) e de Compressão Cranio-caudal (CC) da mesma paciente. Veja a Figura 11 com a representação das imagens das aquisições MLO e CC.

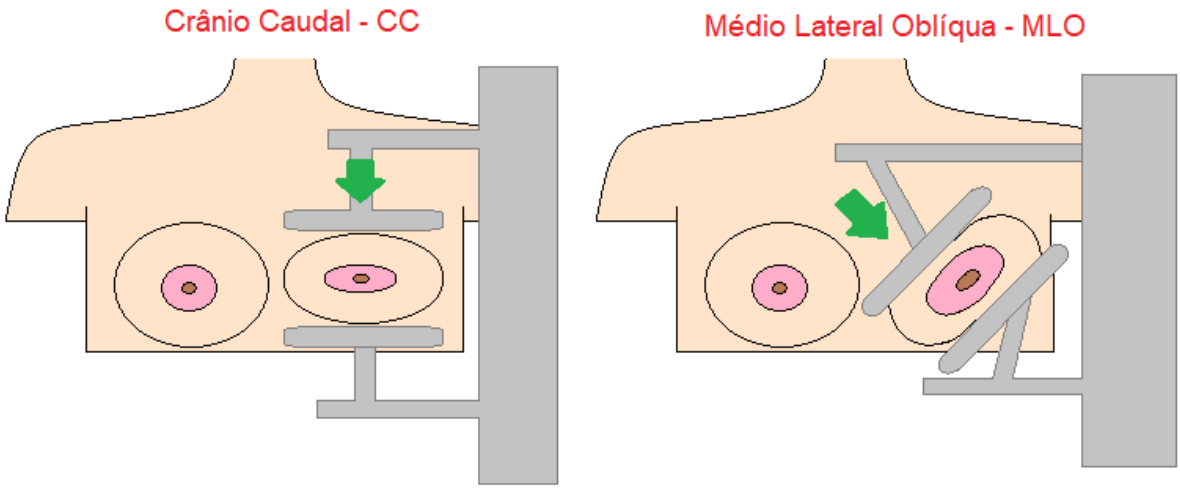


Figura 11: Representação visual de um exame de mamografia.

2.6.2 Descrição da coleta

Foram analisadas 1.100 imagens do Banco de Dados Digital para Mamografia de Rastreamento (DDSM) (HEATH et al., 1998), aplicando o processamento descrito na Seção 2.5.5 nas imagens de pacientes com diagnóstico positivo, foram coletados 564 objetos com o diagnóstico de câncer veja na Figura 12. Além disso, foram selecionadas aleatoriamente 500 imagens de pacientes saudáveis para completar a base de dados, foram gerados 9.972 objetos saudáveis. Não foram usadas as imagens benignas neste trabalho.

A análise dos resultados, focando separadamente em distinguir as imagens (MLO) de (CC), será pelo fato de não utilizar os mesmo objetos encontrados nas imagens MLO e CC para evitar *Double dipping*, ou seja, evitar contaminação e influência nos modelos de classificação por ter o mesmo paciente no treinamento e no teste. Cada imagem oferece percepções e aspectos valiosos para os modelos. A base de dados utilizada para este estudo consiste em um conjunto representativo, onde cada estudo possui 4 imagens, um MLO e uma CC para cada mama, provenientes do Banco de Dados Digital para Mamografia de Rastreamento (DDSM) (HEATH et al., 1998). Essa base possui aproximadamente 2.500 sujeitos, 4 imagens para cada paciente, resultando em aproximadamente 10.000 imagens. Neste trabalho foram usadas 1.600 imagens deste banco.

Primeiramente, veja na Figura 12 a avaliação das imagens MLO revelou 284 objetos com diagnosticados com câncer de mama e 5.089 objetos com o diagnosticados saudável. De forma análoga, a análise das imagens CC resultou em 280 objetos com diagnosticados

com câncer de mama e 4.319 objetos com o diagnosticados saudável. A abordagem individualizada dessas imagens proporcionou uma visão específica das particularidades dessa projeção mamográfica em relação ao diagnóstico de câncer.

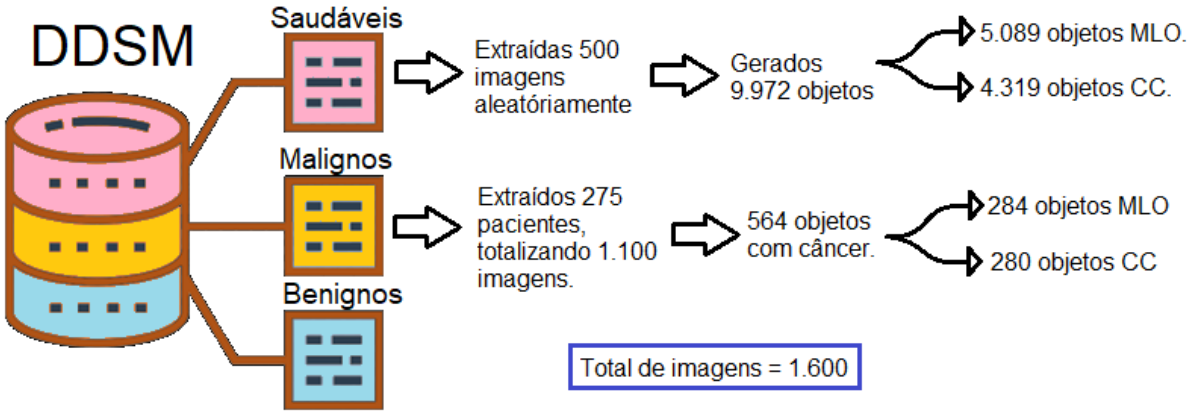


Figura 12: Divisão dos dados para a construção dos modelos.

3 Resultados

3.1 Explorando as diferenças entre os tipos de exame.

Durante a análise minuciosa dos resultados, uma comparação abrangente entre os exames de Mamografia de Obliquidade Médio-Lateral (MLO) e Compressão Cranio-caudal (CC) foi conduzida, focando em métricas essenciais, como medidas de textura de Haralick, momentos estatísticos, forma e área. Constatou-se que não há diferenças significativas nas distribuições dessas variáveis em relação ao tipo de exame, sugerindo uma uniformidade notável nos padrões texturais e estruturais capturados por ambas as projeções mamográficas, isso pode ser visualizado nas Figuras 14, 13 e 15. Entretanto, observe na Figura 14 que algumas medidas de textura dos objetos com o diagnóstico possuem uma diferença relevante em comparação com os objetos saudáveis.

É relevante observar que, apesar dessa uniformidade geral, muitas variáveis apresentaram um número substancial de *outliers*, valores discrepantes que podem potencialmente impactar negativamente a precisão dos modelos criados. Essa presença notável de *outliers* destaca a importância de estratégias robustas de pré-processamento e tratamento de dados, visando mitigar o impacto desses valores atípicos na construção e desempenho dos modelos de classificação.

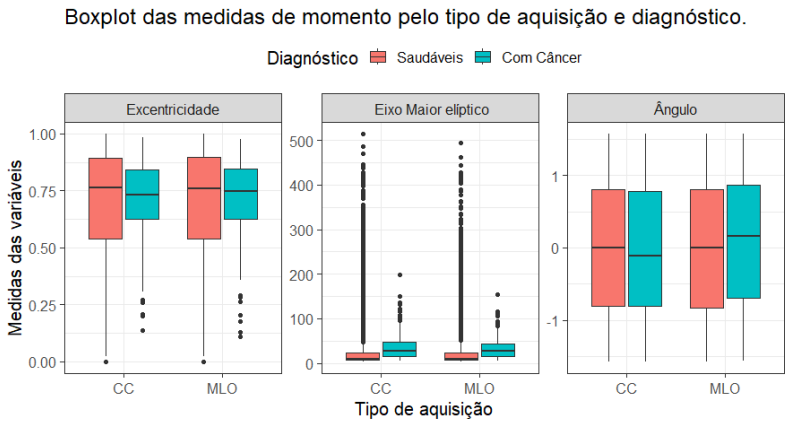


Figura 13: Variação das medidas de momento segundo o tipo de aquisição e o diagnóstico.

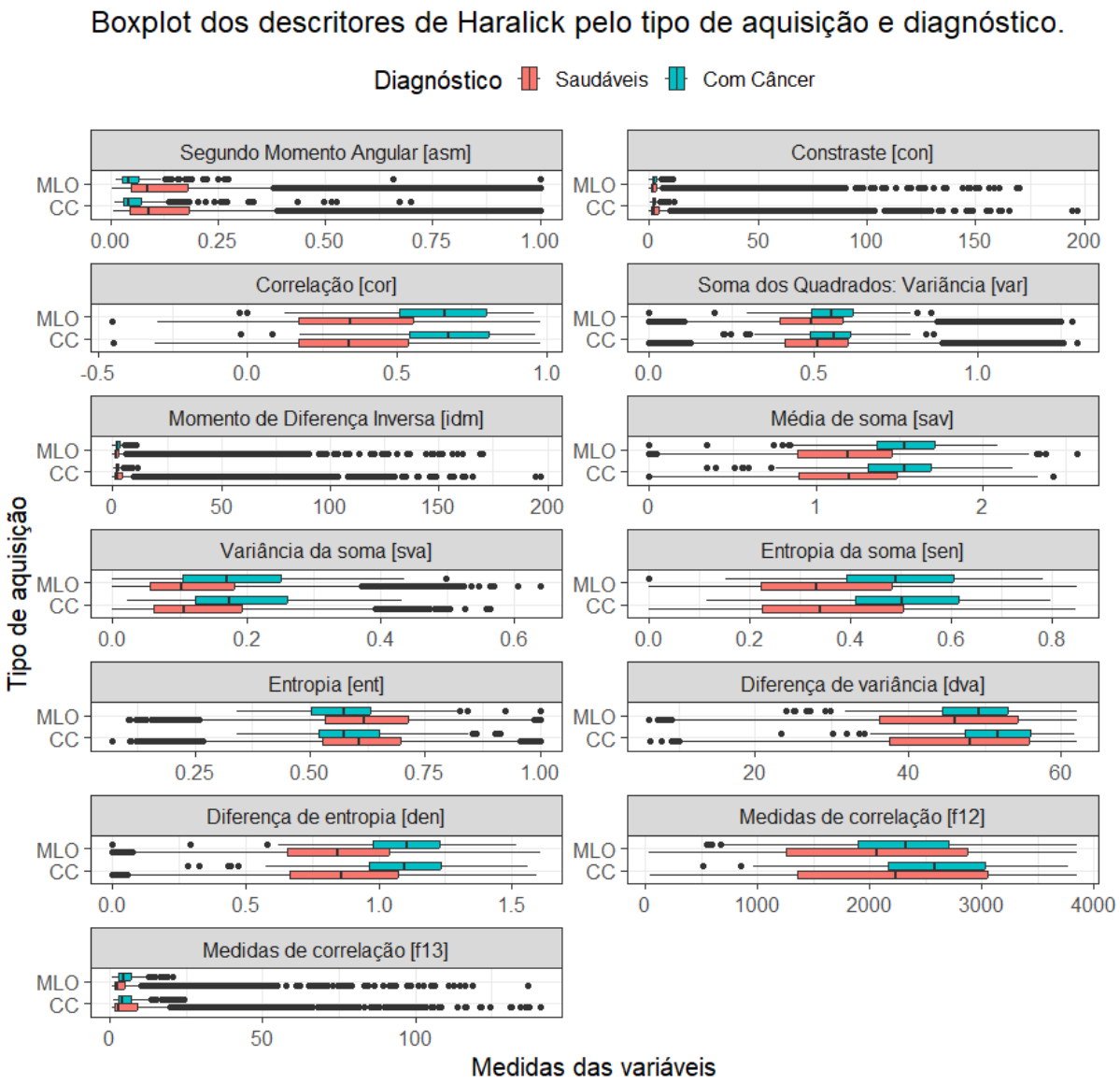


Figura 14: Variação das medidas de textura de Haralick segundo o tipo de aquisição e o diagnóstico.

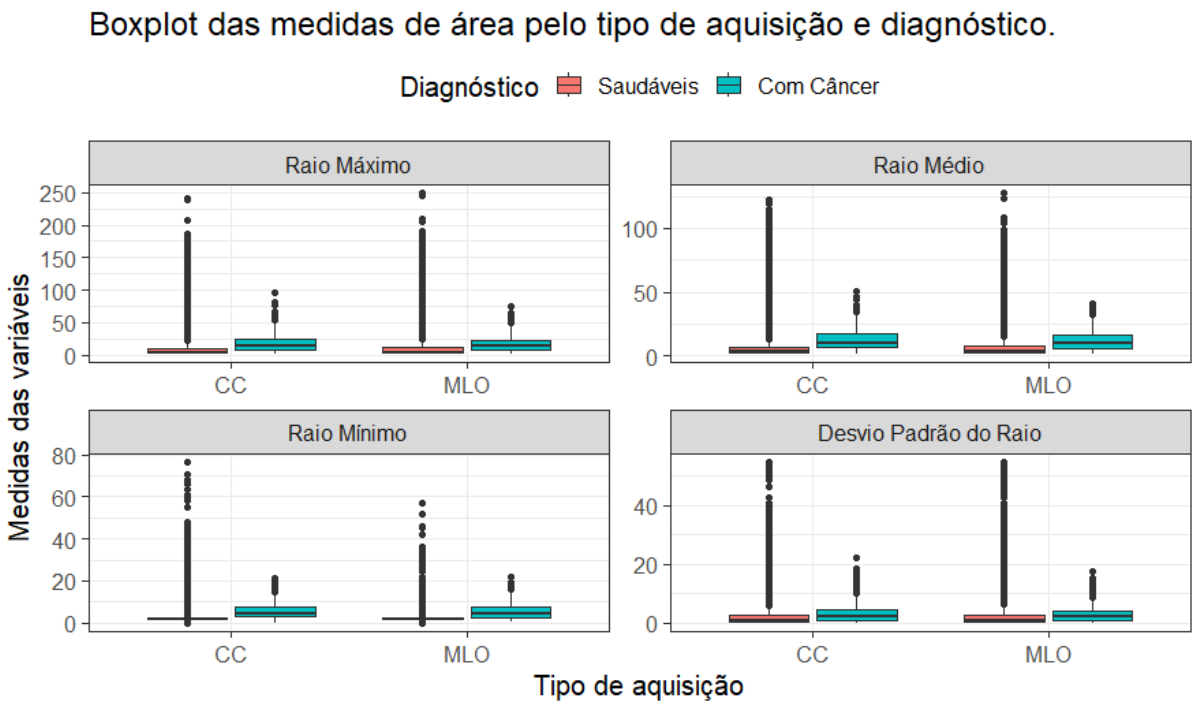


Figura 15: Variação das medidas de área segundo o tipo de aquisição e o diagnóstico.

Este achado levanta a necessidade de uma cuidadosa seleção de variáveis para a construção dos modelos, uma vez que nem todas as métricas analisadas demonstraram ser igualmente úteis ou discriminantes. Assim, a utilização de critérios de seleção torna-se imperativa para identificar as variáveis mais informativas e relevantes, contribuindo para a eficácia e interpretabilidade dos modelos propostos. Na sessão de resultados será detalhada as variáveis que foram selecionadas e as excluídas em cada cenário avaliado. Veja a Figura 16, o qual é um mapa de correlação entre todas as variáveis utilizadas. Este mapa é uma matriz que relacionada as variáveis entre si, as bolinhas representam o valor de correlação entre as variáveis, e a escala de cor específica o quanto correlacionadas então as variáveis. Como esperado, muitas variáveis são muito correlacionadas positivamente e algumas negativamente. Veja que principalmente os descritores de textura entre si como Haralick, Shanmugam e Dinstein (1973) mencionam em seu artigo.

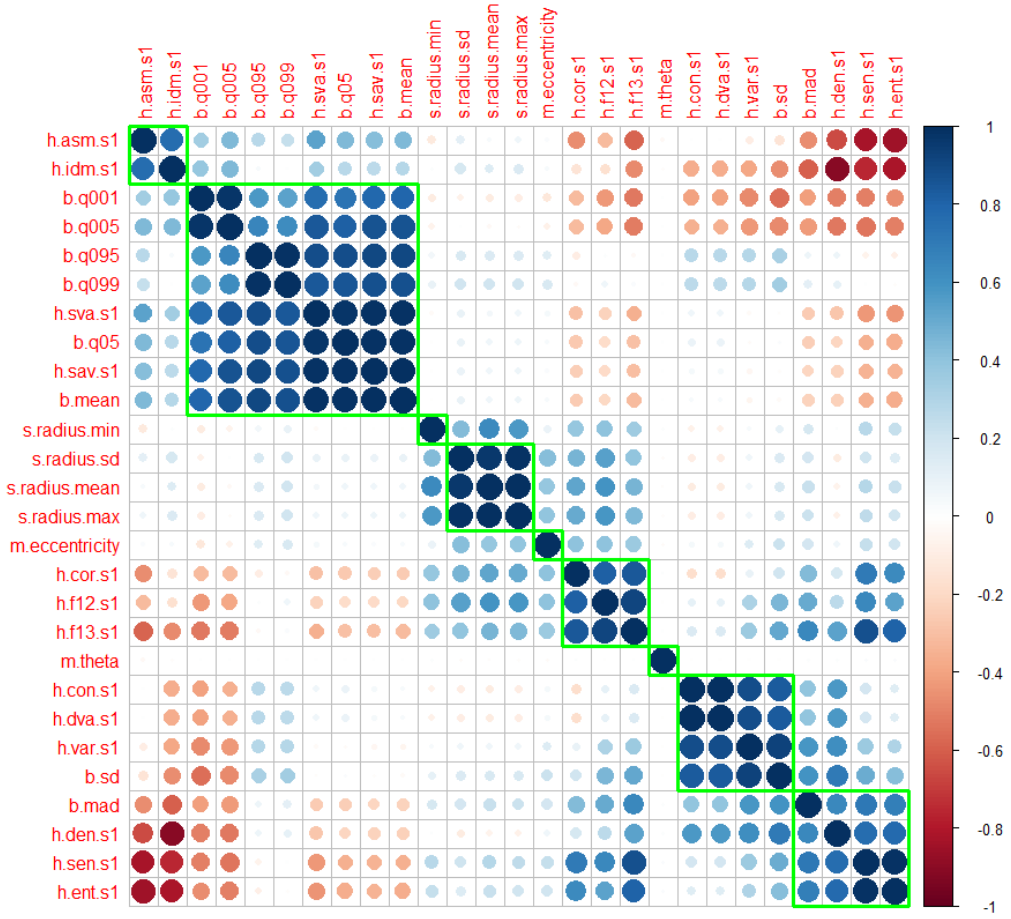


Figura 16: Heatmap: Mapa de correlação entre as variáveis.

3.2 Previsão dos modelos.

Este enfoque duplo na análise de diferentes projeções mamográficas visa fornecer uma visão mais completa e refinada da eficácia do modelo de detecção de câncer de mama, aprimorando sua aplicabilidade clínica e potencial impacto na prática médica. Foram retiradas medidas de área, perímetro e posição central do objeto antes do pré-processamento, visando deixar os modelos mais robustos para aplicar em imagens de tamanhos diferentes.

Como pré-processamento para os modelos apresentados, e a base construída neste trabalho foi aplicado a padronização nos dados e a exclusão de variáveis com variância zero ou quase zero e variáveis correlacionadas. As variáveis correlacionadas com correlação em módulo, maior ou igual a 0,95 foram excluídas da análise. As variáveis excluídas, na comparação entre os modelos MLO e CC são quase as mesmas, veja na Tabela 5.

Tabela 5: Variáveis usadas (x) e excluídas (-) nos modelos, avaliação utilizando todas as variáveis.

Variáveis	Aquisição		Variáveis	Aquisição	
	MLO	CC		MLO	CC
Segundo Momento Angular [asm]	(x)	(x)	Raio Médio	(-)	(-)
Correlação [cor]	(x)	(x)	Percentil 5%	(-)	(-)
Soma dos Quadrados: Variância [var]	(x)	(x)	Percentil 50%	(-)	(x)
Momento de Diferença Inversa [idm]	(x)	(x)	Percentil 99%	(-)	(-)
Entropia [ent]	(x)	(x)	Raio Máximo	(-)	(-)
Diferença de ariância [dva]	(-)	(x)	Entropia da Soma [sen]	(-)	(-)
Medidas de correlação [f12]	(x)	(x)	Média de Soma [sav]	(-)	(-)
Excentricidade	(x)	(x)	Diferença de entropia [den]	(-)	(-)
Ângulo	(x)	(x)	Contraste [con]	(-)	(-)
Percentil 1%	(x)	(x)	Medidas de correlação [f13]	(-)	(-)
Percentil 95%	(x)	(x)	Média	(-)	(-)
Desvio Padrão do Raio.	(x)	(-)	Desvio Padrão	(-)	(-)
Raio Mínimo	(x)	(x)	Variância da soma [sva]	(-)	(-)
Desvio Absoluto Médio	(x)	(x)	Eixo maior elíptico	(-)	(x)
Desvio Padrão do Raio	(x)	(-)			

Para todos os modelos mencionados foi utilizada a metodologia de *cross-validation* com 15 *folds* com 5 repetições. A Tabela 6 a quantidade de dados selecionados para treinamento e teste. A base para treinamento e teste, para ambos os tipos de imagens foram desbalanceadas. Como explicado na Figura 12, foram no total 564 objetos com câncer, onde 280 eram do tipo de aquisição CC e 284 da aquisição MLO. Essas quantidades foram separadas em 2 bases de dados. Para completar as bases de dados avaliando todas as variáveis, foram adicionados o dobro de objetos saudáveis, selecionados aleatoriamente para cada base com seu respectivo tipo de aquisição. Assim como tinha sido explicado na Figura 12 foram gerados 9.972 objetos saudáveis, onde 5.089 são do tipo MLO e 4.319 são do tipo CC. Assim, 568 objetos saudáveis selecionados aleatoriamente completaram a base de dados para o tipo de imagem MLO. E 560 objetos saudáveis selecionados também aleatoriamente completaram a base de dados para o tipo de imagem CC.

Tabela 6: Quantidade de dados separados para treino e teste, para a base desbalanceada com as medidas extraídas pré-processadas.

Label	MLO - Treino	MLO - Teste	CC - Treino	CC - Teste
Sem Câncer	455	113	448	112
Com Câncer	228	56	224	56

Observe a Tabela 8, onde são selecionados os dois melhores modelos para cada tipo de aquisição da imagem MLO e CC. É notório que para ambos os tipos de aquisição a sensibilidade dos modelos ficaram abaixo de 77%. O que é significativamente razoável, pois o objetivo é que os modelos acertem os casos em que os objetos segmentados são tumores, ou seja, queremos que a sensibilidade seja alta. O melhor modelo para as imagens de aquisição MLO foi o GLM e para aquisição CC foi o SVM com *kernel* polinomial. Contudo, podemos concluir que as medidas utilizadas para a criação destes modelos tiveram dificuldades em classificar os objetos com a doença. Veja abaixo na Tabela 7 os hiperparâmetros utilizados em cada modelo de classificação, utilizando a base desbalanceada e todas as variáveis como mencionado.

Tabela 7: Hiperparâmetros utilizados nos modelos de classificação.

Modelo	Nº de árvores	Profundidade	Boostrap	Aquisição
Floresta Aleatória	500	6	Usado	MLO
Floresta Aleatória	500	10	Usado	CC
Adaboost	100	9	Usado	MLO
Adaboost	100	9	Usado	CC
modelo	Nº Máximo de interações	Profundidade Máxima	Método de avaliação	Aquisição
XGboost	100	50	AUC	MLO
XGboost	100	50	AUC	MLO
Modelo	Familia	Função de ligação	Parâmetro de regularização	Aquisição
Regressão Logística	Binomial	Logit	Não usado	MLO
Regrssão Logística	Binomial	Logit	Não usado	CC
Modelo	Parâmetro de regularização	Grau da função	Kernel	Aquisição
SVM	1	segundo grau	Polinomial	MLO
SVM	1	segundo grau	Polinomial	CC
SVM	0.5	-	Radial	MLO
SVM	1	-	Radial	CC
SVM	1	-	Liner	MLO
SVM	1	-	Liner	CC

Tabela 8: Resultado dos modelos para a base desbalanceada com todas as variáveis.

Modelo	Acurácia	Sensibilidade	Especificidade	Aquisição
Adaboost	84,61	67,85	92,92	MLO
Adaboost	84.52	76,78	88,39	CC
Floresta Aleatória	82.84	58.93	94,69	MLO
Floresta Aleatória	85.71	73.21	91,96	CC
KNN	84,62	69,64	92,04	MLO
KNN	82,74	73,21	87,50	CC
Regressão Logística (GLM)	87,57	75,00	93,81	MLO
Regressão Logística (GLM)	85,71	76,79	90,18	CC
Svm Linear	85,80	69,64	93,81	MLO
Svm Linear	84,52	76,79	83,39	CC
Svm Radial	85,21	64,29	95,58	MLO
Svm Radial	85,71	75,00	91,07	CC
Svm Polinomial	82,84	64,29	92,04	MLO
Svm Polinomial	86,90	76,79	91,66	CC
XGboost	83,43	66,07	92,04	MLO
XGboost	83,93	71,43	90,18	CC

A Tabela 10 apresenta a quantidade de dados selecionados para o treino e para o teste. Visando o mal desempenho dos modelos na Tabela 8, selecionamos para nova análise somente as variáveis de textura de Haralick, Shanmugam e Dinstein (1973). O pré-processamento foi igual para os modelos criados na Tabela 8. Veja a Tabela 9 com as variáveis, que foram utilizadas e excluídas neste cenário.

Tabela 9: Variáveis utilizadas (x) e excluídas (-) nos modelos utilizando somente os descritores de textura de Haralick.

Descritores de Textura	MLO	CC	Descritores de Textura	MLO	CC
Segundo Momento Angular [asm]	x	x	Entropia [ent]	x	x
Contraste [con]	x	x	Correlação [cor]	x	x
Soma dos Quadrados: Variância [var]	x	-	Diferença de variância [dva]	-	-
Momento de Diferença Inversa [idm]	x	x	Diferença de entropia [den]	-	x
Média de Soma [sav]	-	-	Medidas de correlação [f12]	x	x
Variância da soma [sva]	x	x	Medidas de correlação [f13]	-	-
Entropia da Soma [sen]	-	-			

Ao contrário dos modelos testados com todas as variáveis, vimos que se utilizarmos apenas as variáveis de textura melhorou significativamente a sensibilidade da maioria dos modelos utilizados, mas também devemos mencionar que alguns modelos ficaram com a especificidade reduzida. Entretanto, obtivemos um modelo melhor para cada tipo de aquisição com a sensibilidade relativamente boa.

Analogamente visto na Figura 12, foram no total 564 objetos com câncer, onde 280 foram para a base CC e 284 para a base MLO. Como havia 9.972 objetos saudáveis, onde 5.089 são do tipo MLO e 4.319 são do tipo CC, Balanceamos a base utilizando o método *downsampling*, onde a base terá a mesma quantidade de instâncias para ambas as classes. Assim, 568 objetos completam a base de dados para o tipo de imagem MLO. E 560 objetos completam a base de dados para o tipo de imagem CC. A separação para treino e teste se encontram na Tabela 10.

Tabela 10: Quantidade de dados separados para treino e teste para a base balanceada com os descritores de textura de Haralick.

Label	MLO - Treino	MLO - Teste	CC - Treino	CC - Teste
Sem Câncer	228	56	224	56
Com Câncer	228	56	224	56

Veja na Tabela 12 com os resultados obtidos. Mais, para as imagens de aquisição MLO o melhor modelo foi o *Adaboost* e para as imagens CC o KNN. Note que o modelo SVM com *kernel* radial, também teve resultados muito próximos ao KNN. E também veja a Tabela 11 com os hiperparâmetros utilizados para cada modelo para esta base balanceado utilizando somente os descritores de textura de Haralick.

Tabela 11: Quantidade de dados separados para treino e teste para a base balanceada com os descritores de textura de Haralick.

Modelo	Nº de árvores	Profundidade	Boostrap	Aquisição
Floresta Aleatória	500	3	Usado	MLO
Floresta Aleatória	500	3	Usado	CC
Adaboost	100	9	Usado	MLO
Adaboost	100	9	Usado	CC
modelo	Nº Máximo de interações	Profundidade Máxima	Método de avaliação	Aquisição
XGboost	100	50	AUC	MLO
XGboost	100	50	AUC	MLO
Modelo	Familia	Função de ligação	Parâmetro de regularização	Aquisição
Regressão Logística	Binomial	Logit	Não usado	MLO
Regressão Logística	Binomial	Logit	Não usado	CC
Modelo	Parâmetro de regularização	Grau da função	Kernel	Aquisição
SVM	1	segundo grau	Polinomial	MLO
SVM	1	segundo grau	Polinomial	CC
SVM	1	-	Radial	MLO
SVM	1	-	Radial	CC
SVM	1	-	Liner	MLO
SVM	1	-	Liner	CC

Tabela 12: Resultado dos modelos com a base balanceada e utilizando os descritores de textura de Haralick.

Modelo	Acurácia	Sensibilidade	Especificidade	Aquisição
Adaboost	85,71	87,50	83,93	MLO
Adaboost	81,25	85,14	76,78	CC
Floresta Aleatória	70,54	69,64	71,43	MLO
Floresta Aleatória	85,71	87,50	83,93	CC
KNN	66,96	76,79	57,14	MLO
KNN	88,39	91,07	85,71	CC
Regressão Logística (GLM)	72,32	67,86	76,79	MLO
Regressão Logística (GLM)	82,14	89,29	75,00	CC
Svm Linear	72,32	69,64	75,00	MLO
Svm Linear	83,04	89,29	76,79	CC
Svm Radial	76,79	75,00	78,57	MLO
Svm Radial	87,50	91,07	83,93	CC
Svm Polinomial	70,54	71,43	69,64	MLO
Svm Polinomial	84,82	89,29	80,36	CC
XGboost	74,11	71,43	76,79	MLO
XGboost	83,93	83,93	83,93	CC

4 Conclusão

Ao encerrar esta pesquisa, é crucial refletir sobre os objetivos delineados e os resultados obtidos. Embora os desempenhos alcançados pelos modelos propostos não tenham atingido os patamares ideais, é evidente, que existe um potencial significativo para aprimoramentos substanciais. Além disso, os objetivos deste estudo foram alcançados.

A revisão bibliográfica que encaminhou este estudo, destacou metodologias promissoras para a detecção de câncer de mama. Em particular, o trabalho de Roy et al. (2021) ressaltou a eficácia do método CatBoost, alcançando uma acurácia de 92,5% e uma precisão média de 93,4%. Outra pesquisa significativa, conduzida por Gonçalves (2021), utilizou SVM e reamostragem para atingir 99,11% de acurácia e sensibilidade máxima. Já o estudo de Marques et al. (2017) empregou características de textura para identificar regiões de massa, obtendo 91,83% de acurácia, 92,28% de sensibilidade e 94,37% de especificidade.

Nossa pesquisa se baseou no artigo do Marques et al. (2017) introduzindo a ideia de encontrar região nas quais poderíamos descrevê-la quantitativamente além de classificá-las. Ou seja, analisar as regiões encontradas extraindo características de textura, forma, medidas baseadas em histograma entre as demais utilizadas neste trabalho. É importante mencionar que, ao contrário desses estudos descritos na revisão bibliográfica, nossa abordagem subdividiu a base em MLO e CC e não utilizou reamostragem. Isso tudo para evitando duplicidade de objetos dos dois tipos de aquisição MLO e CC do mesmo paciente na base de treino e na de teste.

Assim, portanto, permitindo uma análise mais refinada, considerando as peculiaridades de cada tipo de exame. É notório que nosso estudo não alcançou medidas melhores que o artigo de Marques et al. (2017), essa diferença pode ser explicada por diversos fatores. Um deles pelo fato de que a segmentação utilizada é diferente a que este trabalho se propos a realizar. O artigo de (MARQUES et al., 2017), nos inspirou a realizar essa análise quantitativa dos objetos segmentados utilizando a morfologia matemática e

a segmentação binária.

Os modelos mais promissores emergiram ao lidar com a base de imagens CC, destacando-se pelo alcance de uma sensibilidade desejável. Especificamente, o modelo KNN demonstrou desempenho superior, alcançando uma acurácia de 88,39%, sensibilidade de 91,07% e especificidade de 85,71%. No caso da base MLO, o modelo *Adaboost* destacou-se com acurácia 85,71%, sensibilidade de 87,50%, e especificidade de 83,93%.

Veja a Figura 17, com um exemplo ilustrativo da imagem original com a classificação do melhor modelo. Na imagem original segmentada, as áreas de roxo são os objetos segmentados pelo algoritmo. Já na imagem com a classificação do modelo, é apresentado os objetos pintados de roxo, esses objetos foram os que receberam a classificação positiva para câncer segundo o modelo. Note que apenas o objetos segmentado que está na área contornada feita pelo profissional é o de verdadeiro positivo (o que realmente tem a doença). Os demais objetos foram classificados erroneamente pelo modelo (Falso positivo). Os objetos que tiveram a classificação que não possuíam a doença, não são pintados na imagem com a classificação do modelo.

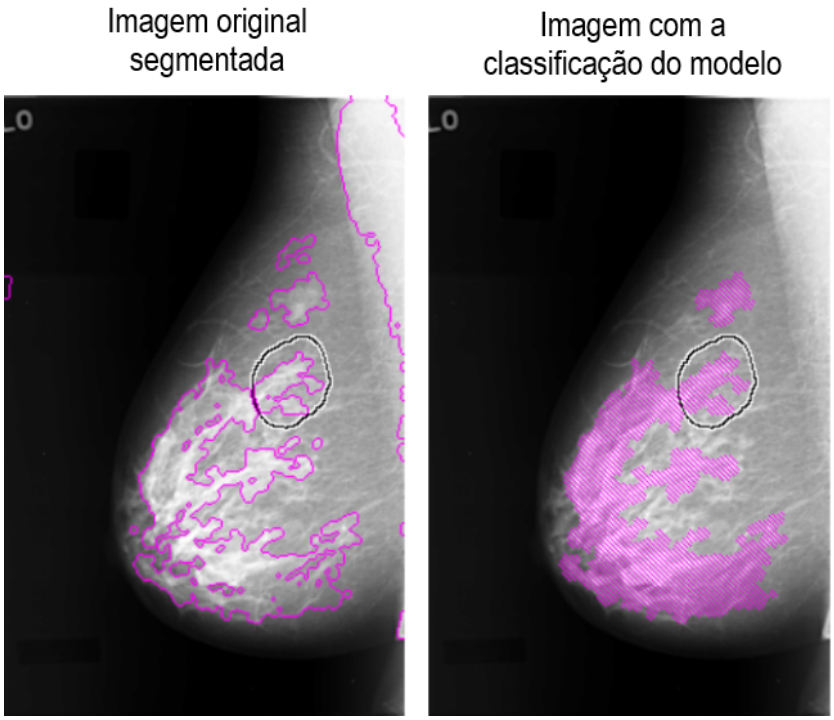


Figura 17: Exemplo ilustrativo do resultado da classificação do melhor modelo, visualizando os objetos segmentados que tiveram a classificação de câncer como positiva.

As dificuldades encontradas durante o desenvolvimento deste estudo acrescentam nu-

ances importantes para a interpretação dos resultados. A segmentação automática utilizada apresentou desafios ao tentar identificar com precisão áreas problemáticas em algumas imagens, resultando na exclusão dessas instâncias da análise. Na Figura 18, é apresentado duas instâncias que quando ocorridas foram excluídas análise. Porque a segmentação automática mostrou-se suscetível a variações nos formatos e tamanhos dos tecidos mamários, divergindo das marcações feitas por profissionais, assim impactando negativamente na rotulação da base construída neste trabalho. Ou seja, na área delimitada pelo profissional, muitas vezes a segmentação segmentava mais de uma área na região problemática ou às vezes não segmentava nenhuma área do problema. Por isso, esses casos foram excluídos do estudo, sendo aceito somente os objetos que cobriam boa parte ou toda a lesão. Veja a ilustração na Figura 18.

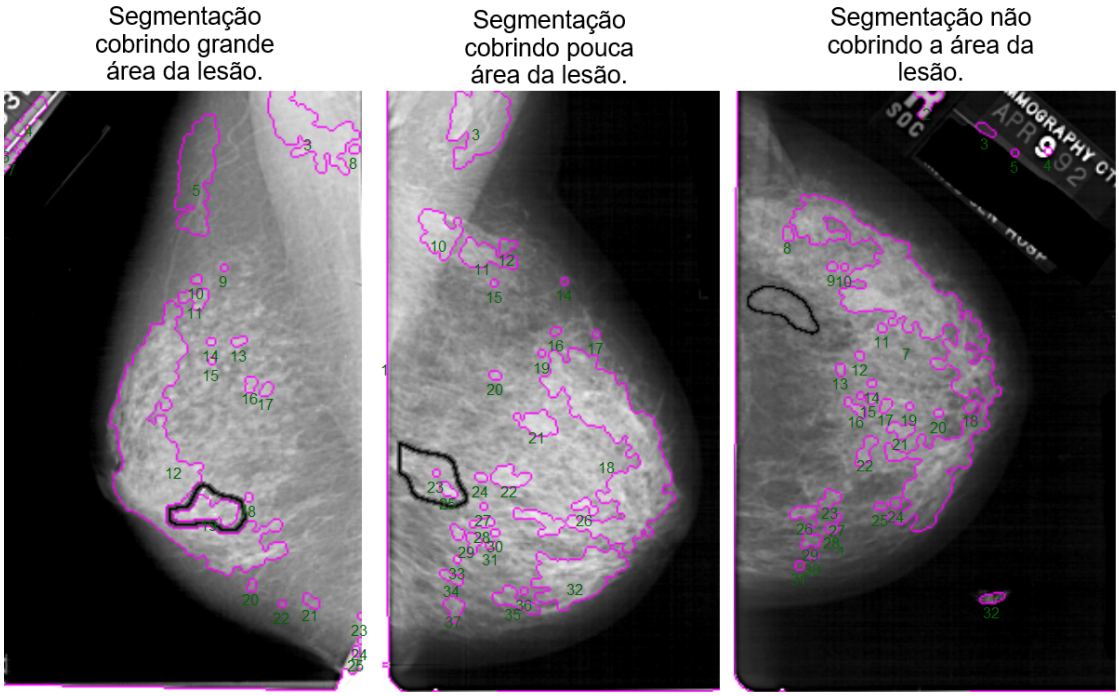


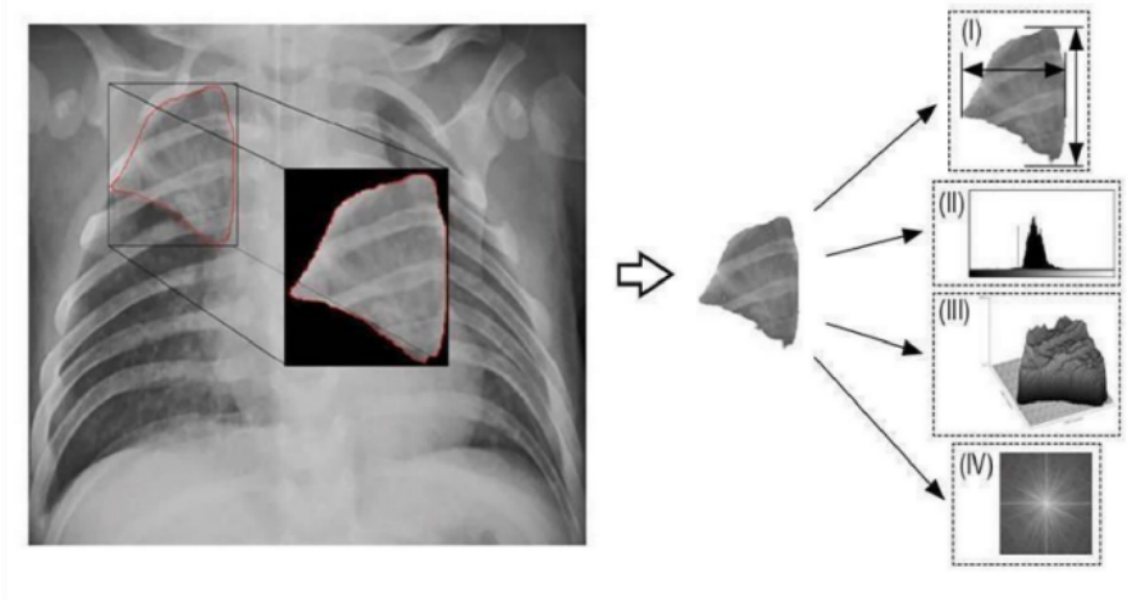
Figura 18: Exemplos de objetos e imagens excluídos da análise.

A quantificação das imagens, ou seja, expressar aquela imagem em medidas quantitativas que a representa, revelou-se uma tarefa desafiadora, culminando na dificuldade de criar modelos robustos para a classificação. O tempo utilizado na seleção dos objetos segmentados também foi uma consideração relevante, visto que a segmentação automática podia gerar múltiplos objetos em uma única região marcada pelos profissionais. A escolha adequada dos objetos gerados tornou-se crucial na montagem da base de dados. Porque a falta de qualificação para essa tarefa poderia acarretar resultados imprecisos e classificações equivocadas dos modelos de classificação utilizados. Para sanar este problema

adotamos em selecionar os maiores objetos quando essa situação ocorria, e descartar da análise as imagens no qual a segmentação gerava objetos duvidosos e de tamanhos semelhantes, dificultado a rotulação, como mencionado antes.

Para trabalhos futuros, sugerimos explorar abordagens mais avançadas de segmentação automática ou semi-automática, visando uma melhor delimitação dos tecidos mamários nas imagens. A otimização dos modelos pode ser potencializada por meio da investigação de outras técnicas de aprendizado de máquina e arquiteturas de redes neurais convolucionais. O aumento da base de dados, com mais exemplos de áreas segmentadas com a doença.

Além disso, uma análise radiômica mais abrangente pode ser incorporada, explorando um conjunto mais diversificado de medidas para quantificar as características das imagens, veja a Figura 19 de exemplo. Essa abordagem mais ampla proporcionaria uma compreensão mais profunda dos padrões radiológicos e, potencialmente, contribuiria para o refinamento dos modelos de detecção de câncer de mama.



Fonte: PEREIRA, et al. 2021. Padrão de consolidação (lobo superior direito) sugestiva de pneumonia bacteriana. Esse é um exemplo em que a lesão poderia ser caracterizada por atributos de forma (I), histograma (II), textura (III) e espectro (IV).

Figura 19: Exemplo de uma análise radiômica.

Além de quantificar uma imagem em diversas medidas, incluindo as que foram utilizadas neste trabalho, que foram apenas descritores de textura e algumas medidas baseadas em forma e histograma. Ainda existem outras medidas envolvendo espectros de Wavelet e Fourier e mais 4 categorias de descritores de textura. Isso porque, neste estudo, as me-

didas de textura foram significativas para explicar o problema abordado. Veja na Figura 20 as medidas utilizadas na análise de radiômica. Utilizando esta análise radiômica, com as outras medidas não utilizadas, pode ser que os modelos consigam melhorar cada vez mais.

Principais Atributos extraídos na Radiômica		
Grupo	Atributos	
Forma e tamanho	Máximo diâmetro 3D (maior distância euclidiana entre voxels de um VOI), volume (número de voxels de uma região multiplicado pelo tamanho do voxel), área da superfície, compacidade, desproporção esférica, esfericidade, densidade, entre outros.	
Estatísticas de primeira ordem (histograma)	Média, mediana, variância, desvio padrão, desvio absoluto médio, entropia, obliquidade, curtose, entre outros.	
Estatísticas de segunda ordem ou textura	GLCM	Autocorrelação, energia, contraste, entropia, inverso da variância, MCC, entre outros.
Estatísticas de segunda ordem ou textura	GLRLM	GLN, percentual de execução, variância dos níveis de cinza, variância executada, entre outros.
Estatísticas de segunda ordem ou textura	GLSZM	Ênfase de menor área, ênfase de maior área, zona de variância, zona de entropia, entre outros.
Estatísticas de segunda ordem ou textura	NGTDM	Granulação, contraste, complexidade, força, entre outros.
Estatísticas de segunda ordem ou textura	GLDM	Ênfase de pequena e grande dependência, dependência não uniforme, dependência de entropia, entre outros.
Estatísticas de ordem superior	Transformada de Wavelet, de Fourier, entre outros.	Decomposição dos níveis de cinza.

Fonte: PEREIRA, et al. 2021. VOI: volume de interesse; GLCM: matriz de concorrência dos níveis de cinza; MCC: coeficiente de correlação máximo; GLRLM: matriz dos comprimentos de sequências dos níveis de cinza; GLN: não uniformidade dos níveis de cinza; GLSZM: matriz de tamanho das zonas de níveis de cinza; NGTDM: matriz da diferença dos tons de cinza da vizinhança; GLDM: matriz de dependência do nível de cinza.

Figura 20: Medidas extraídas em uma análise radiômica de uma imagem.

Em síntese, embora os resultados deste estudo apontem para desafios e limitações, acreditamos que as sugestões e aprimoramentos propostos oferecem uma trilha promissora para futuras pesquisas, direcionando esforços para aperfeiçoar a precisão e confiabilidade dos modelos de detecção de câncer de mama.

Referências

AMORIM, F. S. Previsão de indícios de fraude em fundos de pensão utilizando modelos de aprendizado de máquina supervisionados e técnicas de balanceamento de dados. 2022.

BASTOS, D.; NASCIMENTO, P.; LAURETTO, M. Proposta e análise de desempenho de dois métodos de seleção de características para random forests. In: SBC. *Anais do IX Simpósio Brasileiro de Sistemas de Informação*. [S.l.], 2013. p. 49–60.

CAI, H. et al. An online mammography database with biopsy confirmed types. *Scientific Data*, Nature Publishing Group UK London, v. 10, n. 1, p. 123, 2023.

CHEN, T. et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, v. 1, n. 4, p. 1–4, 2015.

CONCI, A. *Co-ocorrência: Um Descritor para o Reconhecimento de Padrões de Textura*. 2017. Disponível em: ⟨http://www.ic.uff.br/~aconci/co-ocorrencia.pdf⟩.

GONçALVES, G. S. *Predição de diagnóstico de câncer de mama por aprendizado de máquina*. Muriaé, 2021. Trabalho de Conclusão de Curso (Graduação em Ciências Biológicas). Disponível em: ⟨https://www.conic-semesp.org.br/anais/files/2020/trabalho-1000006031.pdf⟩.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. H. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, IEEE, n. 6, p. 610–621, 1973.

HASTIE, T. et al. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer, 2009. v. 2.

HEATH, M. et al. Current status of the digital database for screening mammography. *Digital Mammography: Nijmegen, 1998*, Springer, p. 457–460, 1998.

KUHN, M. Caret: classification and regression training. *Astrophysics Source Code Library*, p. ascl–1505, 2015.

KUMAR, R. M.; SREEKUMAR, K. A survey on image feature descriptors. *Int J Comput Sci Inf Technol*, Citeseer, v. 5, p. 7668–7673, 2014.

LOEWKE, N. *Haralick Texture Analysis for Stem Cell Identification*. 2013.

MARQUES, R. C. d. S. et al. Um estudo sobre vetores de descritores localmente agregados para diagnóstico de câncer de mama. *Jornal Brasileiro de Tecnologia Aplicada nas Ciências da Saúde*, v. 2, n. 2, p. 13–22, 2017. Disponível em: ⟨http://sistemas.deinf.ufma.br/anaisjim/artigos/2016/201606.pdf⟩.

MINUSSI, J. A.; DAMACENA, C.; JR, W. L. N. Um modelo de previsão de solvência utilizando regressão logística. *Revista de Administração Contemporânea*, SciELO Brasil, v. 6, p. 109–128, 2002.

MIYAMOTO, E.; MERRYMAN, T. Fast calculation of haralick texture features. *Human computer interaction institute, Carnegie Mellon University, Pittsburgh, USA. Japanese restaurant office*, 2005.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003.

OTSU, N. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, IEEE, v. 9, n. 1, p. 62–66, 1979.

PAU, G. et al. Ebimage—an r package for image processing with applications to cellular phenotypes. *Bioinformatics*, Oxford University Press, v. 26, n. 7, p. 979–981, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014. Disponível em: ⟨http://www.R-project.org/⟩.

ROY, S. D. et al. Computer aided breast cancer detection using ensembling of texture and statistical image features. *Sensors*, MDPI, v. 21, n. 11, p. 3628, 2021.

SCHWARTZ, W. Segmentaçao de imagens baseada em dependência espacial utilizando o campo aleatório de markov associado com caracterısticas de texturas. *Mestrado em Informática, Universidade Federal do Paran*, 2005.

SEMOLINI, R. et al. Support vector machines, inferência transdutiva e o problema de classificação. *Campinas, SP*, 2002.

URBACH, E. R.; WILKINSON, M. H. Efficient 2-d grayscale morphological transformations with arbitrary flat structuring elements. *IEEE Transactions on Image Processing*, IEEE, v. 17, n. 1, p. 1–8, 2007.

VEZHNEVETS, A.; VEZHNEVETS, V. Modest adaboost-teaching adaboost to generalize better. In: CITESEER. *Graphicon*. [S.l.], 2005. v. 12, n. 5, p. 987–997.