RICARDO WANDRÉ DIAS PEDRO

Utilização de gramáticas estocásticas para reconhecimento de nódulos em mamogramas e validação de contornos nodulares gerados a partir de técnicas de processamento de imagens

São Paulo
2023

RICARDO WANDRÉ DIAS PEDRO

Utilização de gramáticas estocásticas para reconhecimento de nódulos em mamogramas e validação de contornos nodulares gerados a partir de técnicas de processamento de imagens

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

São Paulo
2023

RICARDO WANDRÉ DIAS PEDRO

Utilização de gramáticas estocásticas para reconhecimento de nódulos em mamogramas e validação de contornos nodulares gerados a partir de técnicas de processamento de imagens

Versão Corrigida

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

Área de Concentração:

Engenharia de Computação

Orientador:

Profa. Dra. Fátima de Lourdes dos Santos Nunes Marques

São Paulo
2023

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

São Paulo, 04 de Outubro de 2023

Assinatura do autor: Ricardo Wandré Dias Pedro

Assinatura do orientador: _____

Catalogação-na-publicação

Pedro, Ricardo Wandré Dias
Utilização de gramáticas estocásticas para reconhecimento de nódulos em mamogramas e validação de contornos nodulares gerados a partir de técnicas de processamento de imagens / R. W. D. Pedro -- versão corr. -- São Paulo, 2023.
169 p.

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

1.APRENDIZADO COMPUTACIONAL 2.RECONHECIMENTO DE PADRÕES 3.GRAMÁTICAS FORMAIS POR COMPUTADOR 4.INFORMÁTICA MÉDICA I I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

AGRADECIMENTOS

Agradeço a todos que participaram e que me ajudaram nessa jornada. Em especial, agradeço à minha família e aos meus amigos pela apoio, às minhas orientadoras Profa. Fátima e Profa. Ariane pelos ensinamentos e ao programa de pós-graduação da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo pelo apoio financeiro que permitiu minha participação no *32nd IEEE CBMS International Symposium on Computer-Based Medical Systems* que ocorreu em Córdoba na Espanha.

RESUMO

O câncer de mama é um dos tipos de câncer mais comuns entre as mulheres, representando cerca de 15% de todas as mortes decorrentes de câncer no mundo. A mamografia é considerada o método mais efetivo na detecção precoce desta doença. No entanto, embora existam algumas regras que podem ser seguidas para diferenciar os casos benignos dos malignos, apenas cerca de 15 a 30% dos casos levados à biópsia são nódulos malignos. Para ajudar no diagnóstico da doença, vários sistemas foram desenvolvidos nas últimas décadas para servir como segunda opinião aos médicos radiologistas. Em geral, imagens de nódulos com os contornos demarcados são servidas aos sistemas de auxílio ao diagnóstico (CADx) que fornecem uma classificação desses nódulos, por exemplo, benigno ou maligno. Existem bases públicas nas quais é possível encontrar mamogramas com a indicação da posição dos nódulos, mas esta indicação nem sempre é precisa. A teoria das linguagens formais pode ser utilizada como método sintático na compreensão e criação de conteúdo de imagens, em especial para reconhecimento e construção de objetos. No entanto, métodos sintáticos não têm sido investigados com frequência para reconhecimento de nódulos encontrados nas imagens mamográficas e tampouco para geração de imagens sintéticas de nódulos. O objetivo deste projeto é empregar métodos sintáticos para classificar nódulos benignos e malignos, aplicar técnicas de processamento de imagens para gerar imagens sintéticas de nódulos e, em seguida, utilizar as gramáticas aprendidas de forma supervisionada para validar as imagens geradas. Desta forma, este projeto apresenta uma proposta inicial para geração de imagens sintéticas de nódulos que podem ser utilizadas na composição de uma nova base de imagem. A criação de uma nova base de imagens é útil devido à escassez de bases de imagens públicas com segmentação adequada dos nódulos. A posição das bordas dos nódulos são conhecidas, facilitando a utilização da base de imagens por pesquisadores das áreas de engenharia/computação e para treinamento de profissionais da área de saúde. Os resultados obtidos pelos classificadores sintáticos na classificação dos nódulos considerando as classes benigno e maligno são similares aos do estado da arte. Os resultados obtidos com a geração de imagens sintéticas de nódulos indicaram que técnicas mais elaboradas são necessárias para obtenção de realismo.

Palavras-Chave – Reconhecimento de padrões, classificação de nódulos, câncer de mama, mamografia, geração de imagens sintéticas, gramáticas estocásticas.

ABSTRACT

Breast cancer is one of the most common types of cancer among women, accounting for about 15% of all cancer-related deaths worldwide. Mammography is considered the most effective method for early detection of this disease. However, although there are some rules that can be followed to differentiate between benign and malignant cases, only about 15 to 30% of cases taken for biopsy are malignant nodules. To help diagnose the disease, several systems have been developed in recent decades to serve as a second opinion for radiologists. In general, images of nodules with demarcated contours are served to computer aided diagnosis systems (CADx) that provide a classification of these nodules, for example, benign or malignant. There are public databases in which it is possible to find mammograms with an indication of the position of the nodules, but this indication is not always accurate. The theory of formal languages can be used as a syntactic method in understanding and creating image content, especially for object recognition and construction. However, syntactic methods have not been frequently investigated for recognizing nodules found in mammographic images, nor for generating synthetic images of nodules. The goal of this project is to employ syntactic methods to classify benign and malignant nodules, apply image processing techniques to generate synthetic images of nodules and then use the grammars learned in a supervised way to validate the generated images. Thus, this project presents an initial proposal for the generation of synthetic images of nodules that can be used to create a new image database. Creating a new image database is useful due to the scarcity of public image databases with adequate nodule segmentation. The position of the contours of the nodules is known, facilitating the use of the image base by researchers in the areas of engineering/computing and for training doctors and healthcare professionals. The results obtained by the syntactic classifiers in the classification of nodules considering the benign and malignant classes are similar to the state of the art. The results obtained from the generation of synthetic images of nodules indicate that more complex techniques are necessary to obtain realism.

Keywords – Pattern recognition, masses classification, breast cancer, mammogram, generation of synthetic images, stochastic grammars.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

1D Unidimensional

1DB Dimensão fractal 1D *Box*

1DR Dimensão fractal 1D *Ruler*

2D Bidimensional

2B Dimensão fractal 2D *Box*

2R Dimensão fractal 2D *Ruler*

3D Tridimensional

AC Acutância

ACC A. C. Camargo Cancer Center

ANN *Artificial neural networks*

AT Acutância tradicional

AUC *Area under the ROC curve*

B Benigno

BC Benigno circunscrito

BE Benigno espiculado

BI-RADS *Breast Imaging Report and Data System*

BnE Benigno não espiculado

CAD *Computer-aided diagnosis*

CADe *Computer-aided detection*

CADx *Computer-aided diagnosis*

CC Craniocaudal

CF Concavidade fracionada

CO Contraste

CP Compacidade

CV Coeficiente de variação

DBT *Digital breast tomosynthesis*

DDSM *Digital Database for Screening Mammography*

DGAN *Deep generative adversarial networks*

EM *Expectation-Maximization*

FF Fator de Fourier

FFDM *Full-field digital mammography*

FN Falso negativo

FP Falso positivo

GAN *Generative adversarial networks*

GLC Gramática livre de contexto

GLCM *Gray level co-occurrence matrix*

GR Gramática regular

GSC Gramática sensível ao contexto

HOG *Histogram of oriented gradient*

IE Índice de espiculação

INCA Instituto Nacional do Câncer

KNN *K-nearest neighbors*

LDA *Linear discriminant analysis*

LGBM *Light Gradient Boosting Model*

M Maligno

MC Maligo circunscrito

ME Maligno espiculado

MIAS *Mammographic Image Analysis Society*

MLO Médio-lateral-oblíquo

MnE Maligno não espiculado

OMS Organização Mundial da Saúde

PB Peter Borne

PCA *Principal Component Analysis*

RDP Ramer-Douglas-Peucker

RF *Random Forest*

RMS *Root-mean-squared*

ROC *Receiver operating characteristic*

ROI *Region of interest*

SFFS *Sequential forward floating selection*

ST *Screen Test: Alberta Program for the Early Detection of Breast Cancer*

SVM *Support Vector Machine*

TGR *Tree Grammar Reestimator*

VN Verdadeiro negativo

VP Verdadeiro positivo

UNIFESP Universidade Federal de São Paulo

XGBoost eXtreme Gradient Boosting

SUMÁRIO

PARTE I

INTRODUÇÃO E DEFINIÇÕES

1 INTRODUÇÃO

De acordo com a Organização Mundial da Saúde (OMS), o câncer de mama é um dos tipos de câncer mais comuns entre as mulheres. A cada ano, mais de 1,5 milhão de mulheres sofrem desta doença que possui o maior número de mortes relacionadas ao câncer no mundo. Em 2020, cerca de 685.000 mulheres morreram em decorrência do câncer de mama, o que representa 15% de todas as mortes decorrentes de câncer (OMS, 2023). Segundo o Instituto Nacional do Câncer (INCA), o câncer de mama também pode acometer homens, porém é bastante raro, representando cerca de 1% do total de casos da doença. A estimativa de novos casos para 2022 no Brasil foi de 73.610 (INCA, 2023).

Ainda segundo o INCA, o sintoma mais comum de câncer de mama é a presença de nódulo endurecido, muitas vezes indolor, embora também haja nódulos que apresentam consistência menos rígida. Também são sinais da doença: edema cutâneo que se assemelha à casca de laranja, dor, retração cutânea, descamação do mamilo e secreção papilar (secreção que sai dos mamilos em períodos fora da gestação ou do puerpério), em especial quando ocorre de forma espontânea e unilateral. A secreção associada a este tipo de câncer, em geral é transparente, mas também pode ser rosada ou avermelhada devido à presença dos glóbulos vermelhos. Além desses sintomas, também pode haver o surgimento de linfonodos palpáveis na axila (INCA, 2023).

A mamografia é considerada o método mais confiável e efetivo na detecção precoce de câncer de mama nos estágios iniciais (LI et al., 2017). Embora haja algumas regras para diferenciar entre os casos benignos e malignos, apenas 15-30% dos nódulos que são levados a biópsias são malignos (MOHANTY et al., 2013). Realizar biópsias em situações desnecessárias pode levar a vários problemas tais como o custo financeiro do procedimento, a dor física à qual a mulher é submetida, além da ansiedade e do estresse até o diagnóstico final (TODD; NAGHDY, 2004; KELEs; KELEs; YAVUZ, 2013). Técnicas computacionais desenvolvidas para processar imagens

produzidas por mamografia podem contribuir para diminuir biópsias desnecessárias e auxiliar o diagnóstico; consequentemente, podem levar à detecção precoce com mais precisão.

1.1 Motivação

Para ajudar os médicos proverem diagnósticos mais precisos, vários sistemas de *computer-aided detection* (CADe) e *computer-aided diagnosis* (CADx) foram propostos durante as últimas décadas para, respectivamente, detectar e classificar os achados nos mamogramas (AZOUR; BOUKERCHE, 2022). Muitas técnicas de reconhecimento de padrões e de aprendizado de máquina foram desenvolvidas e aplicadas nesses dois tipos de sistemas.

Em meados de 1950 foi apresentada a teoria das linguagens formais com o objetivo de desenvolver modelos matemáticos relacionados a linguagens naturais. Entretanto, percebeu-se que esta teoria também seria importante para estudos relacionados a linguagens artificiais, em especial, linguagens oriundas da informática e computação (MENEZES, 2008). Desde então, o estudo de linguagens formais vem sendo aplicado na análise sintática de linguagens de programação, para modelar circuitos lógicos, sistemas biológicos etc. Nas últimas décadas, é importante destacar aplicações nas áreas de hipertextos e hipermídias, sistemas de animação, tratamento de linguagens não lineares, tais como linguagens n-dimensionais e linguagens espaciais (MENEZES, 2008), além de aplicações em reconhecimento de padrões estruturais e sintáticos (LI et al., 2021; QI et al., 2021).

Esta área do conhecimento lida com problemas sintáticos das linguagens. Sintaxe de uma linguagem é o nome dado à sua representação visual, à sua forma e à maneira como suas cadeias são estruturadas, sem que seja considerada qualquer informação acerca do seu significado (RAMOS; NETO; VEGA, 2009). Esta teoria possui definições matemáticas bem definidas e universalmente reconhecidas, por exemplo, as gramáticas da hierarquia de Chomsky (CHOMSKY, 1959).

Compreender o conteúdo de imagens sempre foi considerado um dos principais problemas na área de visão computacional (BROWNLEE, 2019). Como as imagens podem ser estruturadas em seus componentes constituintes, a utilização de métodos sintáticos é uma das abordagens utilizadas para tentar resolver este problema. Em especial, nas últimas décadas vários trabalhos publicados utilizaram gramáticas para

lidar com problemas de reconhecimento de padrões em imagens (PEDRO; NUNES; MACHADO-LIMA, 2013). Muitos destes trabalhos tiveram como foco o reconhecimento de objetos nas imagens (PARAG et al., 2012; SOLTANPOUR; HOSSEIN, 2010), reconhecimento de textura (FERREIRA; SANTOS; MONTEIRO, 2007), construção de objetos (PRUSINKIEWICZ; LINDENMAYER; HANAN, 1988; SUN et al., 2009), mudança de escala das imagens (WANG; BAHRAMI; ZHU, 2005), segmentação de imagens e reconhecimento de *layouts* de páginas e documentos (HAMDI; ABDALLAH; BEDOUI, 2012; KANUNGO; MAO, 2003).

Durante a revisão bibliográfica para identificar o estado da arte com relação a classificação de nódulos reais e geração de imagens sintéticas de nódulos, apresentada no Capítulo 3, foi encontrado apenas um trabalho que fez uso de métodos sintáticos com objetivo de reconhecimento de padrões de nódulos encontrados em mamografias (TAHMASBI; SAKI; SHOKOUHI, 2011). Não foram encontrados trabalhos que tenham utilizado gramáticas para criação de imagens sintéticas de nódulos. Diferentemente de Tahmasbi, Saki e Shokouhi (2011), que utilizaram o resultado de análises sintáticas como entrada de uma Rede neural para fazer a classificação dos nódulos, o presente projeto fez uso de gramáticas para classificar os nódulos sem o auxílio de outra técnica de reconhecimento de padrões ou aprendizado de máquina. Adicionalmente, foi proposta a utilização de gramáticas para validar contornos de nódulos criados mediante uso de técnicas de processamento de imagens que poderiam ser utilizados na criação de uma base de dados de imagens sintéticas de nódulos. A base de imagens gerada poderia ser utilizada por pesquisadores das áreas de computação/engenharia e no treinamento de profissionais da área de saúde. Como as bases públicas de imagens de mamogramas possuem algumas imagens com os contornos dos nódulos demarcados de forma não adequada, a base de imagens gerada possuiria os nódulos demarcados corretamente o que facilitaria sua utilização. Neste sentido, este trabalho apresenta ineditismo ao propor um método usando gramáticas para classificação de nódulos reais e para validar contornos de nódulos em imagens sintéticas, conforme descrito na Seção 1.2

1.2 Objetivos

1.2.1 Objetivo geral

O objetivo principal deste projeto de pesquisa é a utilização de métodos sintáticos para fazer a classificação de imagens de nódulos reais encontrados em mamogramas e validar imagens sintéticas de nódulos.

1.2.2 Objetivos específicos

Para que o objetivo principal deste projeto seja alcançado, foram estabelecidos os seguintes objetivos específicos:

1. definição das características utilizadas para representação dos nódulos;
2. representação hierárquica e criação das gramáticas para realizar a classificação dos nódulos considerando as classes Benigno e Maligno;
3. comparação dos resultados obtidos pelos modelos gramaticais na classificação dos nódulos com resultados obtidos pelos modelos não gramaticais;
4. utilização de métodos de síntese de imagens que façam uso de métodos gramaticais.

1.3 Hipóteses

As hipóteses a serem investigadas neste projeto de pesquisa podem ser declaradas da seguinte forma:

1. é possível fazer uso de métodos sintáticos para realizar a classificação de nódulos considerando as classes Benigno e Maligno obtendo resultados similares aos do estado da arte;
2. é possível utilizar métodos sintáticos para validar contornos de nódulos sintéticos criados a partir de técnicas de processamento de imagens ou de aprendizado de máquina.

1.4 Justificativa

Quando o câncer de mama é detectado nos estágios iniciais, as chances de cura são maiores (DOMÍNGUEZ; NANDI, 2009; INCA, 2023). Com o intuito de apoiar os profissionais da área de saúde na tomada de decisão, muitas técnicas para classificação de nódulos, de forma automática, foram propostas nas últimas décadas. Em especial, as técnicas de reconhecimento de padrões e aprendizado de máquina mais utilizadas foram Redes neurais artificiais, Máquinas de vetores de suporte e K-vizinhos mais próximos, conforme percebido durante a revisão da literatura apresentada no Capítulo 3.

A proposta deste trabalho se concentra, inicialmente, na utilização de gramáticas para fazer a classificação dos nódulos nas classes Benigno e Maligno e, em seguida, utilizar as gramáticas criadas anteriormente para validar imagens sintéticas de nódulos geradas por meio de técnicas de processamento de imagens e aprendizado de máquina. A utilização de gramáticas se justifica pelos seguintes motivos:

1. as informações que caracterizam os nódulos podem ser representadas utilizando estruturas hierárquicas. Uma das vantagens da utilização dessas estruturas é a explicabilidade do modelo devido à visualização mais objetiva que pode contribuir para compreensão da forma como as características extraídas dos nódulos são organizadas. Este fato é importante, pois tanto profissionais da área de engenharia/computação, quanto da área médica são capazes de entender o modelo com relativa facilidade, diferentemente do que ocorre, por exemplo, com as Redes neurais artificiais (LARASATI, 2022);

2. trata-se de um trabalho inédito, pois não foram encontrados trabalhos que tenham feito uso de gramáticas, criadas a partir de uma estrutura hierárquica, para classificar nódulos encontrados em mamogramas sem auxílio de outras técnicas de aprendizado de máquina, conforme pode ser verificado no Capítulo 3;

3. no trabalho proposto foi verificada a possibilidade de utilizar gramáticas para validar as imagens sintéticas de nódulos. Estas imagens poderiam compor uma nova base de imagens e ser disponibilizada às áreas de interesse. Embora haja bases de dados de imagens públicas disponíveis, por exemplo, *Digital Database for Screening Mammography* (DDSM) e *Mammographic Image Analysis Society* (MIAS), essas bases de dados não contam com os nódulos segmentados. A DDSM possui, para cada imagem, um arquivo as coordenadas cartesianas do

contorno do nódulo, entretanto, para várias imagens esses pontos não correspondem exatamente ao local onde os nódulos se encontram. A base de dados MIAS fornece apenas as coordenadas cartesianas do centro de massa do nódulo e o tamanho do raio (em pixels) do círculo que engloba o nódulo. Desta forma, para a correta utilização destas bases de dados, é necessário um processo de segmentação inicial executado de forma automática ou manual por um especialista, fazendo com que a utilização dessas bases de dados apresente limitações.

No único trabalho encontrado que fez uso de gramáticas (TAHMASBI; SAKI; SHOKOUHI, 2011), os autores utilizaram uma gramática, desenhada por médicos especialistas, mais complexa que aquelas definidas no decorrer do presente projeto. Aqui, a complexidade está relacionada a um maior número de regras gramaticais, bem como regras de produção recursivas. Além das características utilizadas como entrada para o classificador, existem ainda outras diferenças entre o trabalho aqui proposto e o descrito em (TAHMASBI; SAKI; SHOKOUHI, 2011). A primeira é que foram utilizadas gramáticas diferentes para representar os nódulos malignos e benignos, enquanto que em (TAHMASBI; SAKI; SHOKOUHI, 2011), apenas uma gramática foi utilizada. A segunda diferença é que no trabalho proposto, toda a classificação foi feita com base nas gramáticas criadas, enquanto que em (TAHMASBI; SAKI; SHOKOUHI, 2011) foi construído um modelo que utiliza as saídas obtidas após a análise sintática como entrada de uma Rede neural artificial que é responsável por fazer a classificação. Outra diferença é que no trabalho proposto não há necessidade de um especialista criar a gramática, pois elas são criadas a partir do modelo hierárquico. Por fim, no presente projeto de pesquisa foram utilizadas gramáticas estocásticas que podem ser sempre aprimoradas a medida que são fornecidos novos exemplos rotulados corretamente, enquanto que Tahmasbi, Saki e Shokouhi (2011) utilizaram uma gramática determinística.

1.5 Organização do texto

Além da Introdução, este trabalho possui os seguintes capítulos:

Capítulo 2 - Aspectos Conceituais: são discutidos os principais conceitos sobre o câncer de mama, tais como fatores de risco, prevenção, sintomas e detecção precoce. Também são apresentados conceitos sobre mamografias e a classificação

BI-RADS. Ainda, são apresentados conceitos referentes a gramáticas e sua utilização quando aplicadas a imagens;

Capítulo 3 - Revisão Bibliográfica: apresenta os principais resultados acerca da revisão bibliográfica sistemática realizada para investigar o estado da arte em relação à classificação dos nódulos utilizando técnicas de reconhecimento de padrões e de aprendizado de máquina. Os resultados da revisão bibliográfica realizada para detectar quais técnicas são empregadas para geração de nódulos sintéticos também são expostos neste capítulo;

Capítulo 4 - Materiais e métodos: descreve de forma detalhada os materiais e métodos utilizados durante este projeto de pesquisa para atingir os objetivos propostos;

Capítulo 5 – Classificação de Nódulos: descreve o processo realizado para construção dos classificadores sintáticos para a classificar dos nódulos considerando as classes Benigno e Maligno;

Capítulo 6 – Geração de imagens sintéticas de nódulos: apresenta as abordagens utilizadas para a geração de imagens sintéticas de nódulos e sua validação utilizando métodos gramaticais com o intuito de prover uma base de imagens que possa ser utilizada para testes em sistemas auxílio ao diagnóstico;

Capítulo 7 - Resultados da Classificação dos Nódulos: descreve os resultados obtidos com a abordagem proposta para classificação dos nódulos encontrados nos mamogramas e comparações com outras técnicas utilizadas nesta área de pesquisa;

Capítulo 8 - Resultados da Geração de Imagens Sintéticas de Nódulos: apresenta os resultados obtidos na geração de nódulos sintéticos, sua validação utilizando métodos gramaticais e os principais desafios encontrados para realizar esta tarefa;

Finalizando o trabalho, tem-se a **Conclusão** (Capítulo 9), as **Referências** utilizadas nesta pesquisa, além do **Apêndice**.

2 ASPECTOS CONCEITUAIS

2.1 Câncer de mama

Câncer é o nome dado a um conjunto de mais de 100 doenças que se caracteriza por um crescimento desordenado de células que acabam invadindo tecidos e órgãos (INCA, 2023). As divisões dessas células ocorrem de forma muito rápida, desordenadas e tendem a ser bastante agressivas e incontroláveis, formando nódulos malignos que podem se espalhar para várias partes do corpo. Suas causas podem ser externas ou internas ao organismo. As causas externas ocorrem devido à relação do indivíduo com o meio ambiente, hábitos ou costumes. As causas internas são, geralmente, pré-determinadas geneticamente estando ligadas à incapacidade do organismo de se defender de agressões externas (INCA, 2023).

O câncer de mama é o segundo tipo de câncer mais comum entre as mulheres no Brasil, respondendo por cerca de 28% dos novos casos todos os anos. Também pode acometer homens, mas é bastante raro, representando apenas 1% do total de casos da doença (INCA, 2023).

Existem vários tipos de câncer de mama, sendo que alguns evoluem de forma mais rápida que outros. Ainda segundo o INCA, a estimativa de novos casos de câncer de mama é de 73.610 em 2022. O número de mortes em 2020 foi 18.032, dos quais 207 foram homens e 17.825, mulheres.

2.2 Mamografia

A mamografia é uma forma de radiografia da mama. O objetivo do exame é gerar, por meio da utilização dos raios-X, imagens das estruturas mamárias de tal forma que seja possível detectar algumas doenças, por exemplo, nódulos, calcificações e distorção das mamas, lesões não palpáveis e câncer de mama (MITCHELL-JR; BASSETT,

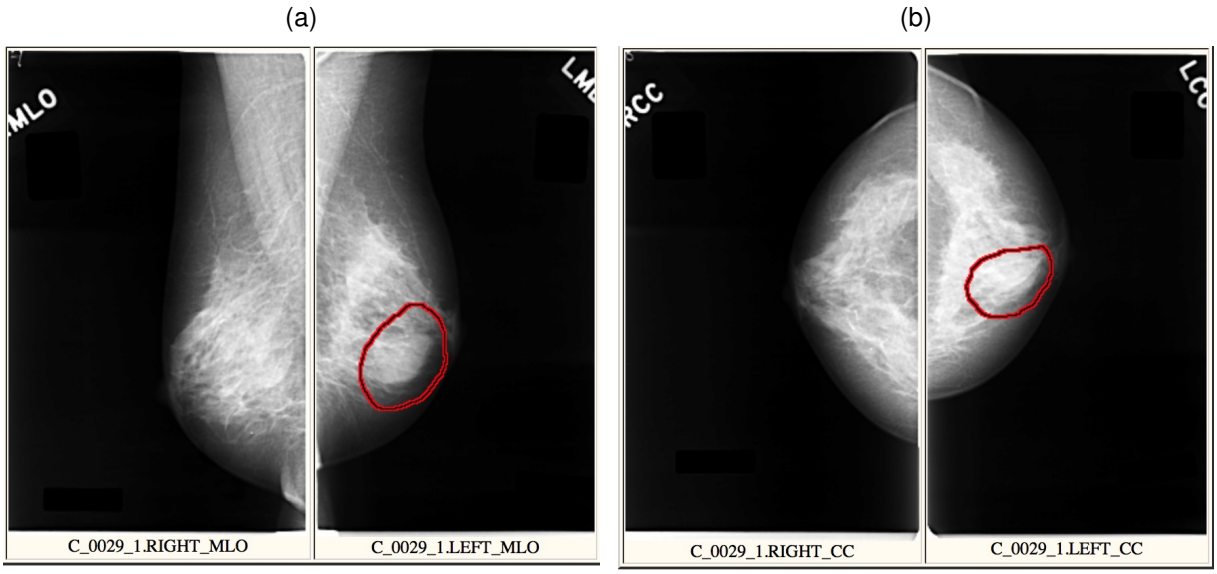
1988).

De acordo com o INCA, no Brasil, a recomendação a partir de 2015 é que mulheres com idade entre 50 e 69 anos façam uma mamografia de dois em dois anos. Essa recomendação vai ao encontro com a adotada em grande parte dos países que implantaram o rastreamento do câncer de mama e que reduziram a mortalidade por essa doença. São benefícios dessa prática a possibilidade de detectar a doença no início e, portanto, possibilitar um tratamento menos agressivo (INCA, 2023). Outro benefício é a diminuição do risco de morrer em decorrência da doença em função de um tratamento adequado. Embora altamente recomendada, a mamografia de rastreamento também implica em certos riscos (INCA, 2023):

- resultados incorretos: quando há suspeita da doença, outros exames podem ser requeridos, sem que a doença seja confirmada no final. Esse resultado é conhecido como falso-positivo e pode deixar o paciente em estado de estresse e ansiedade. É chamado de falso-negativo quando a doença existe, mas o resultado do exame não a indica. Este último tipo de erro pode retardar o tratamento;
- sobrediagnóstico e sobretratamento: o paciente pode ser tratado com cirurgia, quimioterapia e radioterapia de um câncer que não lhe traria risco de morte;
- aumento da exposição aos raios-X: embora seja raro, pode aumentar o risco de desencadear um câncer quando a exposição é mais frequente.

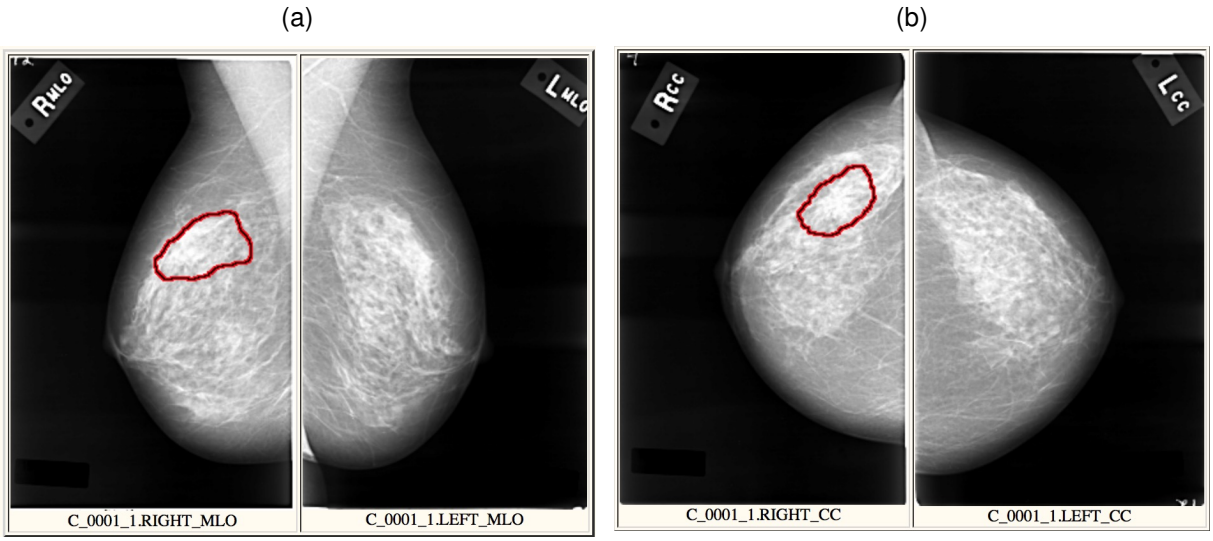
Ainda segundo o INCA, em uma mamografia são indispensáveis uma visão médio-lateral-oblíqua (MLO) e uma visão crânio-caudal (CC) da mama. As Figuras 1 e 2 exibem as visões MLO e CC de um nódulo benigno e de um nódulo maligno, respectivamente. A incidência MLO é mais eficaz que a CC, uma vez que exibe uma maior quantidade de tecido mamário, além de incluir estruturas mais profundas do prolongamento axilar e do quadrante supéro-externo. O objetivo da incidência CC é incluir todo o material póstero-medial de forma a complementar a incidência MLO. Além disso, a incidência CC permite uma maior compressão da mama, pois não inclui a axila, o que resulta em uma definição superior da estrutura mamária e de possíveis lesões (SANTOS, 2002).

Figura 1: (a) Visão médio-lateral-oblíqua de um nódulo benigno. (b) Visão crânio-caudal de um nódulo benigno. As posições dos nódulos estão marcadas em vermelho.



Fonte: Digital Database for Screening Mammography.

Figura 2: (a) Visão médio-lateral-oblíqua de um nódulo maligno. (b) Visão crânio-caudal de um nódulo maligno. As posições dos nódulos estão marcadas em vermelho.



Fonte: Digital Database for Screening Mammography.

2.2.1 Classificação BI-RADS

De acordo com Bassett (1997), *Breast Imaging Report and Data System* (BI-RADS™) é o produto de um esforço colaborativo de vários comitês da *American College of Radiology* com a cooperação do *National Cancer Institute*, dos *Centers for Disease Control and Prevention*, da *Food and Drug Administration*, da *American Medical Association*,

do *American College of Surgeons*, e do *College of American Pathologists*, que serve como um guia para padronizar e organizar as informações referentes à mamografia, tais como descrição e categorização dos achados. O relatório provido, segundo o BI-RADS, deve ser conciso, organizado e ter a seguinte estrutura: i) razão para o exame; ii) descrição sucinta da composição da mama; iii) descrição dos achados seguindo uma terminologia padronizada; e iv) impressão geral e uma avaliação da categoria (BASSETT, 1997).

Os nódulos encontrados nas mamografias são descritos de acordo com sua forma, borda e densidade. A forma pode ser redonda, oval ou irregular. Nódulos ovais ou redondos geralmente estão associados a tumores benignos, enquanto que um nódulo com formato irregular tem uma maior probabilidade de ser maligno. As bordas dos nódulos também são um forte indicativo de probabilidade de malignidade. As bordas podem ser descritas como circunscritas, microlobuladas, obscurecidas, indefinidas ou espiculadas (BASSETT, 1997). A Figura 3 esquematiza os tipos de formas e de bordas dos nódulos. Um nódulo com borda circunscrita tem baixa probabilidade de ser maligna (menos de 2%). Borda microlobulada aumenta a probabilidade do nódulo ser maligno. Um nódulo que possui borda indefinida é suspeito de ser maligno, enquanto que a borda espiculada tem uma alta probabilidade de malignidade (BASSETT, 1997).

O relatório que segue a padronização BI-RADS termina com uma conclusão na qual os achados mais importantes são sumarizados e o exame é categorizado em uma das sete categorias de avaliação existentes conforme exibido na Tabela 1. Se a classificação final for **inconclusiva**, a classificação é BI-RADS categoria 0, se for **negativa** a classificação é BI-RADS categoria 1, se for **achado benigno** a classificação é BI-RADS categoria 2 e se for **achado provavelmente benigno** a classificação é BI-RADS 3; nesses casos a interpretação final é *negativa*. Se a categorização for **anormalidade suspeita** a classificação é BI-RADS 4, se for **altamente sugestivo de malignidade** a classificação é BI-RADS 5; nesses últimos casos a interpretação final é *positiva*. Quando a malignidade do nódulo é confirmada por meio de biópsia a classificação BI-RADS é 6.

2.3 CADe e CADx

O diagnóstico apoiado por computador (do inglês - *Computer-Aided Diagnosis* - CAD) pode ser definido como um diagnóstico fornecido por um radiologista auxiliado

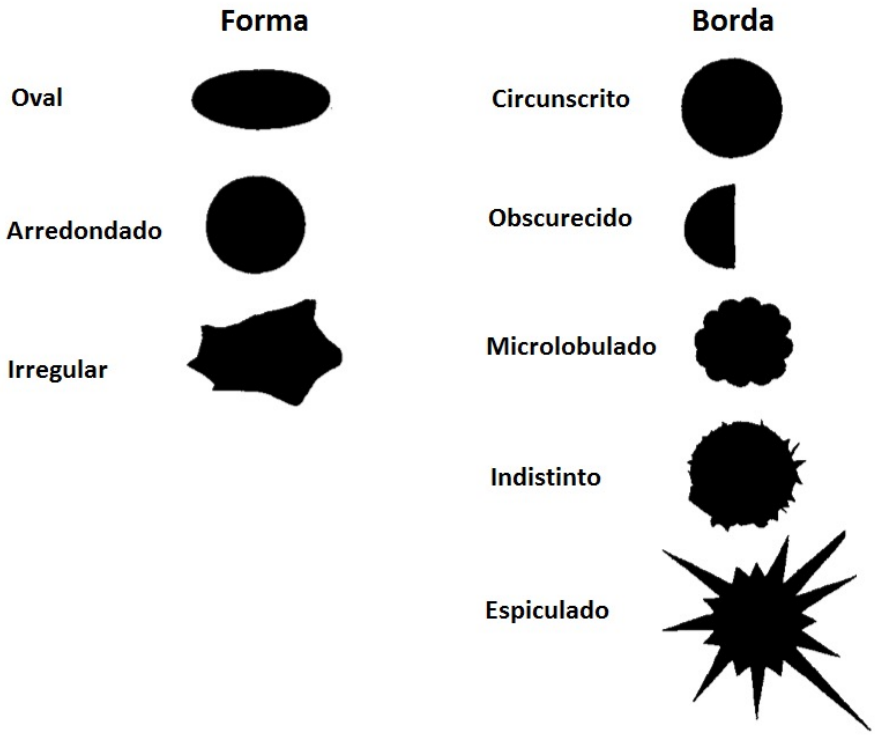


Figura 3: Terminologia utilizada para descrever as formas e as bordas de nódulos. Fonte: Adaptado de (BASSETT, 1997).

pelo resultado das análises das imagens realizadas por um sistema computacional (NISHIKAWA, 2010). Desde que a *Food and Drug Administration* autorizou seu uso, os sistemas CAD vêm sendo utilizados em vários países servindo como uma segunda opinião para o médico radiologista (SCHIABEL, 2014). No caso de mamografias, geralmente o sistema analisa as imagens obtidas e fornece marcações de locais suspeitos para o médico, indicando quais tipos de achados são prováveis em cada local (GIGER; CHAN; BOONE, 2008).

Considerando uma nomenclatura mais recente, os sistemas CAD podem ser divididos em duas categorias: *Computer-Aided Detection* (CADe) e *Computer-Aided Diagnosis* CADx (AZOUR; BOUKERCHE, 2022):

- esquemas CADe auxiliam o médico na detecção de possíveis regiões de interesse, chamando a atenção do médico radiologista para estruturas de interesse clínico. Este, por sua vez, confirma ou não se a estrutura demarcada é de fato algo suspeito. Em geral, esses sistemas demarcam regiões com suspeitas de calcificações, nódulos e áreas com densidade anormal;

Tabela 1: BI-RADS - Categorização final dos achados nas mamografias.

Categoria	Classificação	Descrição
0	Exame inconclusivo	Radiologista considerou o exame inconclusivo ou incompleto. Isso pode acontecer devido à baixa qualidade da imagem ou quando há dúvida da existência ou não de alteração na mama
1	Exame normal ou exame negativo	Exame normal. Nenhuma alteração foi encontrada
2	Exame com achados certamente benignos	Um nódulo definitivamente benigno foi encontrado
3	Exame com achados provavelmente benignos	Alta probabilidade do achado ser benigno
4	Exame com achados suspeitos	Alteração na mama que pode ou não ser um câncer. Biopsia deve ser considerada
5	Exame com elevado risco de câncer	Alta probabilidade de câncer. Biópsia deve ser realizada
6	Exame positivo	Malignidade provada por biópsia

Fonte: Baseado em (BASSETT, 1997).

- esquemas CADx possui um sistema de classificação dos achados detectados pelo CADe. A classificação pode variar dependendo do tipo de estrutura encontrada, por exemplo nódulos, calcificações entre outros. Um sistema CADx ideal deveria ser capaz de detectar todos os achados possíveis presentes em uma imagem e classificá-los gerando uma opinião completa sobre o caso em estudo.

Existem vários esquemas CAD disponíveis no mercado, por exemplo *ImageChecker*®, *CADVision*®, *iCAD*®, *M-Vision*® entre outros (SOUSA, 2017). Além desses sistemas comerciais, vários outros trabalhos foram desenvolvidos com o intuito de criar sistemas de CADx que auxiliem os médicos a fornecerem diagnósticos mais precisos no que diz respeito à classificação dos nódulos encontrados nos mamogramas, conforme apresentado no Capítulo 3.

2.4 Gramáticas

Gramáticas podem ser utilizadas para descrever os mais diversos tipos de objetos, como por exemplo, desde sentenças em vários idiomas e linguagens de programação

até outros exemplos menos intuitivos como plantas (SUN et al., 2009), construções (WANG; JIANG, 2009), objetos criados pelo homem (LUO et al., 2009), etc.

Antes da apresentação formal de uma gramática, serão apresentados a seguir alguns conceitos importantes que podem ser encontrados em (AHO; ULLMAN, 1972; SALOMAA, 1973; HOPCROFT; ULLMAN, 1979).

Um **vocabulário** ou **alfabeto** Σ é qualquer conjunto finito de símbolos (letras ou dígitos, por exemplo). Uma **cadeia sobre um conjunto de símbolos** é uma sequência de símbolos de tal conjunto concatenados. Seja Σ um conjunto de símbolos, dizemos que Σ^* é o conjunto de todas as possíveis cadeias formadas com os símbolos de Σ, ou seja, Σ^* representa todas as cadeias que são a concatenação de zero ou mais símbolos do conjunto Σ, sendo o operador * o **fecho de Kleene**, que representa o conjunto de todos elementos que podem ser formados por meio da concatenação de zero ou mais elementos. A **cadeia vazia** é denotada por ε.

Definição 1: Uma **linguagem** sobre um alfabeto Σ é um conjunto de cadeias sobre este alfabeto.

Definição 2: Uma **gramática** é uma quádrupla $G = (V_N, V_T, R, S)$, na qual V_N é o conjunto de símbolos não terminais, V_T é o conjunto de símbolos terminais, R é o conjunto de regras de produção (ou simplesmente conjunto de produções) e $S \in V_N$ é o símbolo inicial da gramática.

O conjunto V_N é composto por símbolos auxiliares utilizados na elaboração das regras gramaticais, sendo o símbolo S o símbolo que define a primeira regra gramatical a ser utilizada. As produções R representam as regras sintáticas propriamente ditas. Quando uma regra de produção é aplicada a uma cadeia substitui-se uma ocorrência do lado esquerdo da regra (esquerdo ao símbolo $\rightarrow$) por uma do lado direito (direito ao símbolo $\rightarrow$). O conjunto V_T é formado pelo alfabeto Σ da linguagem descrita, ou seja, todos os símbolos que aparecem nas cadeias que se deseja representar.

Exemplo: $G = (V_N, V_T, R, S)$, sendo:

$V_N = \{S\}$

$V_T = \{a, b\}$

$S = S$

$R = \{S \rightarrow \varepsilon, S \rightarrow a, S \rightarrow b, S \rightarrow aSa, S \rightarrow bSb\}$

A gramática do exemplo anterior será utilizada como base nas definições seguin-

tes.

Dizemos que a **gramática** é **livre de contexto** (*GLC*) se suas produções são da forma:

- $A \rightarrow \beta$, $A \in V_N$, $\beta \in (V_T \cup V_N)^*$

Definição 3: Dadas duas cadeias c_1 e c_2, dizemos que c_2 é uma **derivação direta** de c_1, denotada por $c_1 \Rightarrow c_2$ se $c_1 = \alpha\rho\beta$ e $c_2 = \alpha\gamma\beta$ sendo α, $\beta \in (V_T \cup V_N)^*$ e $\rho \rightarrow \gamma$ uma regra de produção de G.

Exemplo: c_1 = *aaSaa* e c_2 = *aabaa*. Neste caso, $\alpha = \beta =$ *aa* e $\rho \rightarrow \gamma$ é representada pela regra S $\rightarrow$ b.

Definição 4: Uma cadeia c_2 é uma **derivação** de uma cadeia c_1, c_1, $c_2 \in (V_T \cup V_N)^*$, denotada por $c_1 \Rightarrow^* c_2$, se existem zero ou mais derivações diretas entre c_1 e c_2.

Exemplo: c_1 =*aaSaa*, c_{11} = *aabSbaa*, c_{12} = *aabaSabaa*, c_2 = *aababababaa*. Neste caso, temos $c_1 \Rightarrow c_{11} \Rightarrow c_{12} \Rightarrow c_2$.

Definição 5: Uma cadeia c_1 é uma **sentença** gerada por uma gramática $G = (V_N, V_T, R, S)$ se $c_1 \in \{c \in V_T^* \mid S \Rightarrow^* c\}$, ou seja, uma sentença é uma cadeia composta de símbolos terminais que seja uma derivação do símbolo inicial da gramática.

Exemplo: A cadeia *babbab* pode ser obtida utilizando as seguintes regras gramaticais na sequência: S $\rightarrow$ bSb; S $\rightarrow$ aSa; S $\rightarrow$ bSb; S $\rightarrow \varepsilon$.

Definição 6: Denotamos por ***L(G)*** o conjunto de todas as sentenças geradas por uma gramática *G*, chamado de **linguagem gerada** por *G*.

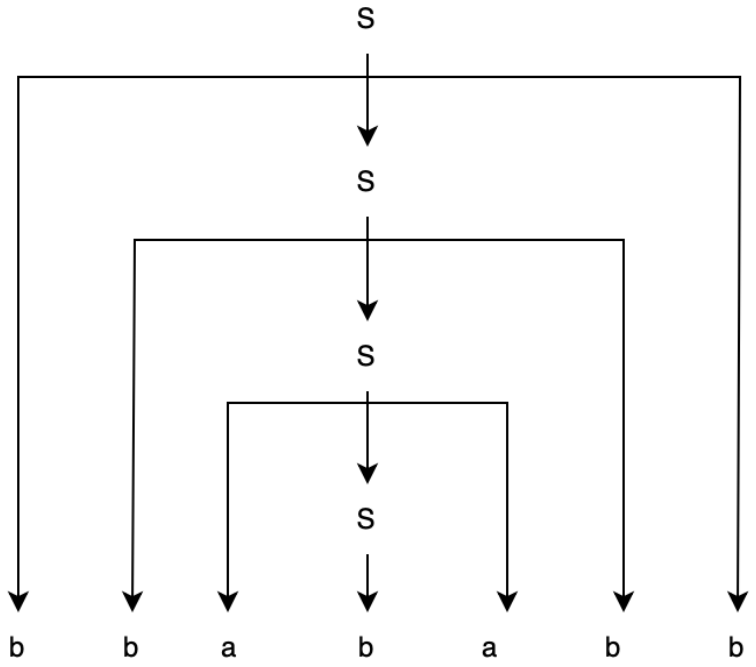
Nota: A gramática utilizada na **Definição 2**, gera a linguagem dos palíndromos, que é quando uma cadeia pode ser lida da esquerda para a direita ou da direita para a esquerda sem alterações na sequência dos símbolos e no seu significado.

Definição 7: Duas gramáticas são **equivalentes** se elas geram a mesma linguagem.

A derivação de uma cadeia segundo uma gramática pode ser representada por meio de uma árvore de derivação ou árvore sintática, conforme pode ser vista na Figura 4 e definida na **Definição 8**.

Definição 8: Dada uma sentença *c* gerada por uma gramática *G*, diz-se que uma árvore é uma **árvore sintática** ou **de derivação** se:

Figura 4: Árvore de derivação ou árvore sintática da cadeia *bbababb*.



Fonte: O autor (2023).

- cada nó da árvore possui um rótulo pertencente ao conjunto V_N ou V_T de *G*;
- o rótulo do nó raiz é o símbolo inicial de *G*;
- se um determinado nó *m*, cujo rótulo é *A*, possui filhos f_1, f_2, ..., f_k na ordem da esquerda para a direita e seus rótulos são X_1, X_2, ..., X_k, respectivamente, então $A \in V_N$ e a produção $A \rightarrow X_1X_2...X_k$ faz parte do conjunto de produções da gramática.

Definição 9: Uma gramática *G* é **ambígua** se existe uma sentença $c \in L(G)$ que possui duas ou mais árvores de derivação distintas.

Definição 10: O **analisador sintático** de uma gramática *G* é um programa que, dada uma cadeia, consegue fornecer uma árvore sintática para tal cadeia caso ela pertença à *L(G)*, ou uma mensagem de erro, caso contrário. Caso a gramática seja ambígua, quando uma sentença possuir mais de uma árvore de derivação, o analisador sintático pode fornecer todas as suas árvores ou apenas uma delas, a depender do algoritmo.

Uma gramática estocástica G_s é capaz de gerar diferentes sentenças *w* com pro-

babilidade P(*w*|G_s). A seguir seguem algumas definições formais acerca de gramáticas estocásticas.

Definição 11: Uma gramática estocástica é uma quíntupla G_s = (*V_N, V_T, R, S, P*), na qual:

- *V_N*, *V_T*, *S* e *R*, têm o mesmo significado que em gramáticas não estocásticas;
- *P* é responsável por associar uma probabilidade p ($0 \leq p \leq 1$) a cada regra de produção do tipo $\alpha \rightarrow \beta$.
- Para cada α que seja o lado esquerdo de uma produção, considerando todas as produções $\alpha \rightarrow \beta_i$ e a probabilidade associada p_i, temos que $\sum_{i=1} p_i$ = 1;

Definição 12: O analisador sintático para uma gramática estocástica *G_s*, é um programa em que dada uma cadeia *c*, se esta cadeia pertencer a *L(G_s)*, o analisador fornece a árvore de derivação e a probabilidade de *c*, segundo sua gramática. Caso contrário, fornece uma mensagem de erro. As seguintes regras são válidas para as gramáticas estocásticas:

- a probabilidade de uma cadeia *c* pertencer à linguagem gerada por uma gramática estocástica *G_s*, considerando uma árvore sintática *t* ($P(c, t|G_s)$), é o produto das produções utilizadas em *t*;
- a probabilidade de uma cadeia *c* pertencer à linguagem gerada por uma gramática estocástica *G_s* considera todas suas árvores sintáticas, sendo o somatório das probabilidades de cada uma das árvores de derivação t_i que podem ser derivadas para a cadeia *c* a partir da gramática *G_s* ($P(c|G_s) = \sum_{i=1} P(c, t_i|G_s)$).

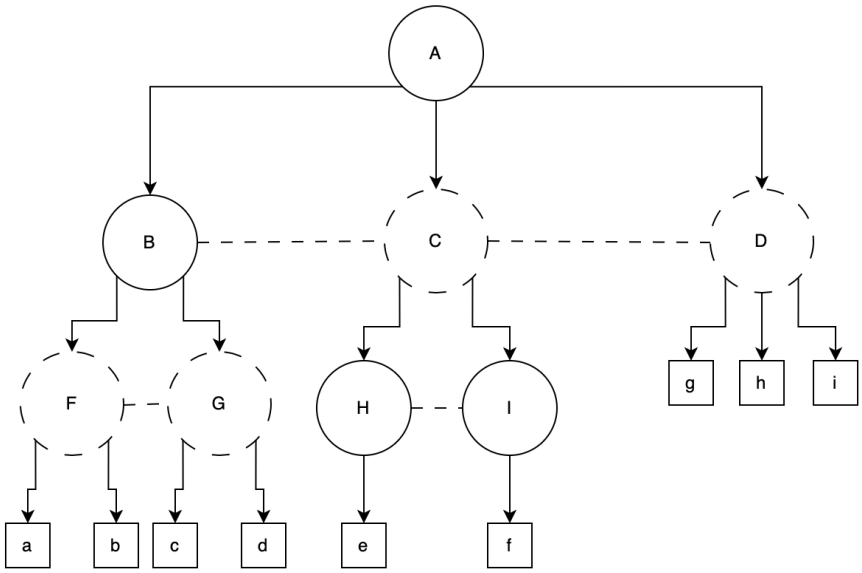
2.5 Grafo AND-OR

Um grafo AND-OR é definido como uma tripla $\mathcal{G} = (V, R, P)$ na qual $V = V^{AND} \cup V^{OR} \cup V^T$ consiste de um conjunto disjunto de nós *AND, OR* e *Terminais*, respectivamente, *R* é um conjunto de relações entre os nós ou subgrafos representando o processo de geração a partir de um nó pai para os seus filhos e *P(r)* é a probabilidade de expansão para cada relação $r \in R$ (XIONG et al., 2016). Enquanto os nós

AND representam a decomposição de uma entidade em suas partes, os nós *OR* fornecem alternativas de escolhas com relação a qual sub-árvore será utilizada (ZHU; MUMFORD, 2006).

Quando o grafo AND-OR é também uma árvore AND-OR ele pode ser interpretado como uma gramática livre de contexto. A Figura 5 apresenta um exemplo simples de uma árvore AND-OR na qual os nós AND estão representados por círculos sólidos, os nós OR estão representados por círculos tracejados e os nós folhas (terminais) estão representados por quadrados. Os nós AND correspondem às regras gramaticais do tipo A $\rightarrow$ BCD, B $\rightarrow$ FG, H $\rightarrow$ e, e I $\rightarrow$ f. Os nós OR representam regras gramaticais alternativas para o mesmo lado esquerdo (utilizando-se o símbolo "|" para separar as alternativas do lado direito), como por exemplo C $\rightarrow$ H | I, D $\rightarrow$ g | h | i, F $\rightarrow$ a | b, e G $\rightarrow$ c | d. Os nós folhas representam os símbolos terminais. As linhas tracejadas entre os filhos dos nós AND apresentam algum tipo de relacionamento entre os nós, mas esses relacionamentos não foram utilizados neste projeto.

Figura 5: Exemplo de grafo AND-OR.



Fonte: O autor (2023).

2.6 Classificador Bayesiano

A teoria de decisão bayesiana é uma abordagem estatística utilizada no problema de tomada de decisão. O teorema de Bayes está expresso na Equação 2.1, na qual A e B são variáveis aleatórias.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2.1}$$

O classificador Bayesiano pode ser utilizado para classificar um objeto como pertencente a uma determinada classe. Neste trabalho o objeto a ser classificado é um nódulo e as classes consideradas Benigno e Maligno, sendo que tanto o objeto quanto as classes são variáveis aleatórias para o classificador. Para decidir a qual classe o objeto pertence, o classificador faz uso das curvas de distribuição de probabilidade, sendo que cada curva corresponde a uma classe. Neste sentido, quando queremos saber a probabilidade de um objeto x para uma determinada classe c_i, queremos saber a probabilidade condicional $P(x|c_i)$ (probabilidade do objeto x dada a classe c_i), também chamada de verossimilhança. A $P(c_i)$, sem considerar o objeto a ser classificado, é a probabilidade a *priori*. A atualização dessa probabilidade ocorre assim que o objeto é considerado e é dada por $P(c_i|x)$ (probabilidade da classe c_i dado o objeto x) sendo chamada de probabilidade a *posteriori*. Fazendo uso do teorema de Bayes é possível combinar as probabilidades a *priori* que temos sobre as classes com as probabilidades condicionais dos objetos dentro das classes, para atualizarmos a probabilidade da classe da *priori* para a *posteriori*, e então classificar um objeto x para a classe c_i que possuir a maior probabilidade *a posteriori*.

De modo geral, a fórmula de Bayes pode ser utilizada para classificar um objeto x em uma classe *c*, de acordo com a Equações 2.2 e 2.3, ou seja, ***posteriori* = verossimilhança****priori*/**evidência**. Desta forma, a classificação de um objeto seria dada pela Equação 2.4.

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)} \tag{2.2}$$

$$P(x) = \sum_j P(x, c_j) = \sum_j P(x|c_j)P(c_j) \tag{2.3}$$

$$\begin{cases} c_1, & \text{se} \quad P(c_1|x) > P(c_2|x) \\ c_2, & \text{caso contrário.} \end{cases} \tag{2.4}$$

Como $P(x)$ é uma constante em relação às classes, podemos resumir a regra à Equação 2.5. Assumindo $P(c_1) = P(c_2)$, a decisão se resume à comparação das verossimilhanças.

$$\begin{cases} c_1, & \text{se} \quad P(x|c_1)P(c_1) > P(x|c_2)P(c_2) \\ c_2, & \text{caso contrário.} \end{cases} \tag{2.5}$$

2.7 Considerações

O câncer de mama é um dos tipos de câncer mais comuns que acomete principalmente mulheres. Diversos fatores estão relacionados ao seu surgimento, tais como fatores endócrinos ou relativos à vida reprodutiva, fatores comportamentais ou relacionados ao ambiente e fatores genéticos. A detecção precoce aumenta as chances de um tratamento adequado e de cura desta doença.

O objetivo da mamografia é registrar imagens da mama para que seja possível detectar indícios de anomalias, tais como nódulos, calcificações e distorção das mamas. Para auxiliar os médicos nas análises das imagens, vários sistemas CAD foram propostos durante os últimos anos tanto para a detecção quanto para classificação dos achados.

Sistemas CAD podem empregar várias abordagens de aprendizado de máquinas e de reconhecimento de padrões durante a análise das imagens. Uma das abordagens que podem utilizadas e que foi pouco explorada neste contexto são os métodos sintáticos, conforme pode ser visto no Capítulo 3.

3 REVISÃO BIBLIOGRÁFICA

Para este projeto, duas revisões de literatura foram feitas. A primeira foi uma revisão bibliográfica sistemática que teve como objetivo identificar quais técnicas são utilizadas para classificação de nódulos, bem como para compreender o estado da arte e as possíveis lacunas desta área de pesquisa. Esta revisão sistemática possibilitou a publicação do artigo (PEDRO; MACHADO-LIMA; NUNES, 2019a) e está resumida na Seção 3.1. Entretanto, como durante a revisão sistemática foram analisados os artigos até 2017, outra revisão da literatura recente foi conduzida e está detalhada na Seção 3.1.2. A segunda revisão bibliográfica foi uma revisão com o objetivo de encontrar quais métodos são utilizados para criação de imagens sintéticas contendo nódulos benignos ou malignos, apresentada na Seção 3.2.

3.1 Classificação de nódulos

A revisão sistemática conduzida foi dividida em três fases: 1) planejamento, com elaboração de um protocolo utilizado como guia para a pesquisa; 2) busca e seleção de artigos de interesse de acordo com critérios de inclusão/exclusão definidos no protocolo; e 3) análise dos artigos selecionados para compreensão do estado da arte desta área de pesquisa.

O protocolo criado definiu três perguntas de pesquisas:

- Quais são as técnicas utilizadas para classificar nódulos nos mamogramas?
- Quais características dos nódulos são utilizados como entrada para os classificadores?
- Quais resultados foram obtidos pelos diversos classificadores e características utilizados?

Para responder estas questões, foram realizadas duas buscas nas bases de dados

em momentos distintos. A primeira foi em setembro de 2016 e a segunda foi em maio e julho de 2017.

Em setembro foram pesquisadas as bases de artigos PubMed[1], que contém mais de 26 milhões de citações relacionadas à literatura biomédica, e Periódico Capes [2], que é uma biblioteca virtual brasileira que disponibiliza a instituições de pesquisa no Brasil o melhor da produção científica mundial. Para estas buscas foi utilizada a *string* "**mammogra* AND (classi* OR recognition) AND (mass OR nodule)**" que retornou 1965 artigos somando-se as duas bases.

Após remover os artigos duplicados, o próximo passo foi aplicar os critérios de inclusão/exclusão a seguir para inclusão do artigo na revisão sistemática.

Critério utilizado para incluir um artigo:

- artigos que lidaram com o problema de classificação de nódulos na mama de forma automática ou semiautomática utilizando mamogramas.

Critérios utilizados para excluir um artigo:

- artigos que exclusivamente lidaram com o problema de detecção de nódulos nos mamogramas;
- artigos que focaram em apresentar técnicas de segmentação de mamogramas;
- artigos que apresentaram estudos de classificação em mamogramas, mas não relacionados a nódulos;
- artigos que utilizaram outras modalidades de imagens, tais como ultrassom ou ressonância magnética.

Após os critérios de inclusão/exclusão terem sido aplicados, foram incluídos 86 estudos nesta revisão sistemática.

Em maio de 2017 foi feita a mesma busca na base PubMed, desta vez limitando aos anos de 2016 e 2017, e foram retornados 62 estudos. Adicionalmente, novas *strings* foram criadas baseadas nas seguintes palavras **mammogram**, **classification**, **recognition**, **mass**, **nodule**, and **NOT microcalcification**. Estas novas *strings* foram

[1] https://www.ncbi.nlm.nih.gov/pubmed
[2] https://www.periodicos.capes.gov.br

utilizadas diretamente nas bases IEEE Xplorer Digital Library[3], Springer Link[4], Elsevier[5], SPIE Digital Library[6], e Medical Physics[7] em julho de 2017 retornando 309 trabalhos no total. Após remoção dos artigos duplicados e dos critérios de inclusão/exclusão terem sido aplicados, foram adicionados mais 43 artigos à revisão sistemática que, somados aos 86 trabalhos adicionados anteriormente, totalizam um total de 129 estudos incluídos. A Tabela 2 mostra as *strings* de busca utilizadas em cada biblioteca digital utilizada.

Tabela 2: Bibliotecas digitais utilizadas para buscar por artigos.

Biblioteca	***String***	**Data**	**Quantidade de artigos**
PubMed	mammogra* AND (classi* OR recognition) AND (mass OR nodule)	Setembro - 2016	945
Periódicos Capes	mammogra* AND (classi* OR recognition) AND (mass OR nodule)	Setembro - 2016	1020
PubMed	mammogra* AND (classi* OR recognition) AND (mass OR nodule)	Maio - 2017	62
IEEE Xplorer	(“Document Title”:mammogra* AND (classification) AND (mass OR nodule) NOT (calcification OR microcalcification))	Julho - 2017	182
Elsevier	title(mammogra* AND (classification) AND (mass OR nodule) AND NOT (calcification OR microcalcification))	Julho - 2017	25
Springer Link	mammogram AND classification AND (mass OR nodule) AND NOT (calcification AND microcalcification)	Julho - 2017	46
SPIE Digital Library	mammogram AND (mass OR nodule) AND (classification) NOT (calcification OR microcalcification)	Julho - 2017	37
Medical Physics	mammogram AND (mass OR nodule) AND (classification) NOT (calcification OR microcalcification) in Article Titles	Julho - 2017	19

Fonte: O autor (2023).

[3]IEEE Xplorer Digital Libray: http://ieeexplore.ieee.org/Xplore/home.jsp
[4]Springer Link: https://link.springer.com
[5]Elsevier: http://www.sciencedirect.com
[6]SPIE Digital Libray: https://www.spiedigitallibrary.org/?SSO=1
[7]Medical Physics: http://aapm.onlinelibrary.wiley.com/hub/journal/10.1002/(ISSN)2473-4209/

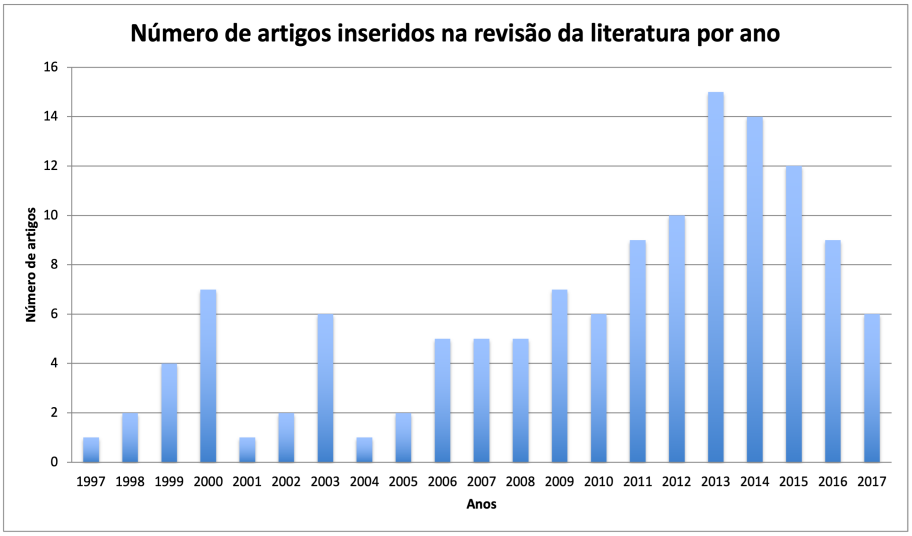
3.1.1 Resultados e discussões da revisão sistemática

Nesta seção são apresentados os resultados encontrados durante a execução da revisão bibliográfica.

3.1.1.1 Análise global

A Figura 6 mostra que esta área de pesquisa tem apresentado um aumento no número de artigos publicados durante as últimas duas décadas, sendo que a maioria dos artigos analisados foi publicada na última década (98 artigos - 76%). Este aumento no número de artigos publicados pode ser resultado de maior poder computacional, algoritmos mais sofisticados para extração das características e classificação, além do aumento no número de imagens disponíveis para serem utilizadas nos estudos. Desde 2014, o número de artigos publicados nesta área vem decrescendo, o que pode ser um indício do início de estagnação desta área de pesquisa.

Figura 6: Quantidade de artigos inseridos na revisão bibliográfica sistemática.

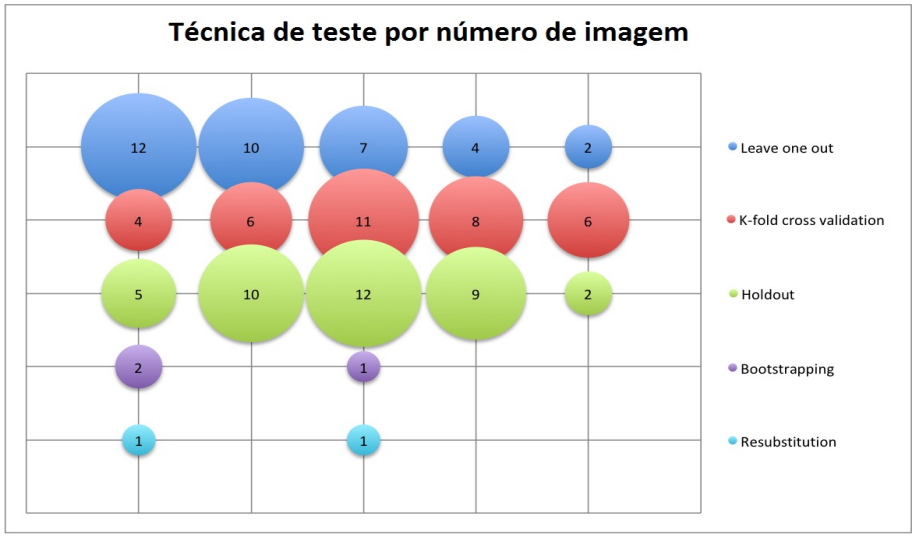


Fonte: O autor (2023).

A abordagem *holdout* foi a técnica de teste e validação mais utilizada (29% dos artigos). Em seguida, as técnicas de validação cruzada utilizando *k-fold* e *leave-one-out* foram apresentadas em 27% dos estudos. Quinze artigos (12%) não mencionaram como o conjunto de imagens foi dividido para o treinamento e teste do classificador. O método *holdout* foi ao mais citado provavelmente porque é o método que consome menos tempo de se utilizar. Ela consiste em dividir o conjunto de dados em duas partes, uma para treinamento e outra para teste, sendo que o conjunto de treinamento

é, tipicamente, maior que o conjunto de testes. A Figura 7 mostra que, quando há um pequeno número de imagens (inferior a 100), os pesquisadores preferem utilizar a técnica *leave-one-out*, mas conforme o número de imagens aumenta, eles tendem a preferir o *holdout* ou validação cruzada utilizando *k-fold*. *Bootstrapping* e resubstituição foram pouco utilizadas.

Figura 7: Técnicas utilizadas para treinamento e testes para validar os classificadores desenvolvidos.



Fonte: O autor (2023).

A maioria dos artigos (78%) considerou apenas as classes “Benigno” e “Maligno” para fazer a classificação dos nódulos. Entretanto, outros trabalhos fizeram a classificação utilizando, por exemplo, as classes Normal, Redondo, Lobular, Oval, Irregular, Estrelado, BI-RADS, etc.

A acurácia foi a medida mais comum utilizada na avaliação dos estudos incluídos, aparecendo em 80 artigos (62%). A segunda medida mais utilizada (74 artigos - 57%) foi a área sob curva ROC (*area under the ROC curve - AUC*) que é a área sob a curva desenhada considerando a sensibilidade em função de 1 - especificidade. Cada ponto da curva ROC representa o par (sensibilidade, especificidade) correspondente a uma determinada escolha de parâmetro. A AUC é utilizada para mensurar quão bom é um classificador, considerando os vários valores de parâmetros, para distinguir entre duas classes (Benigno *vs.* Maligno, Doença *vs.* Normal, etc.). As próximas medidas foram as mais comumente empregadas: sensibilidade e especificidade, que apareceram em 38 artigos (29%) e 37 artigos (29%), respectivamente. A Figura 8 exibe a frequência com que cada medida aparece nos estudos. Na Tabela 3 pode-se ver a definição de cada medida, na qual *P* é o número de exemplos positivos, *N* é o

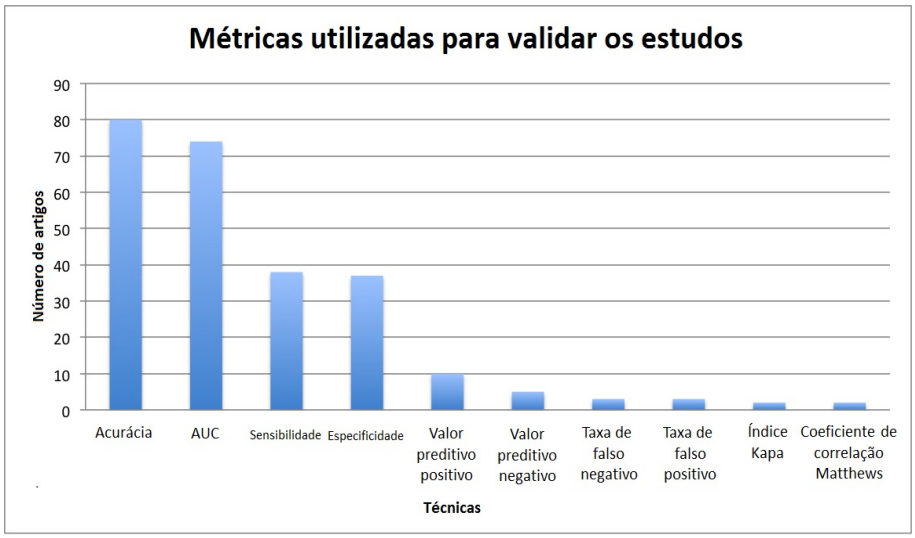
número de exemplos negativos, *VP* é o número de resultados verdadeiros positivos, *VN* é o número de resultados verdadeiros negativos, *FP* são os resultados falsos positivos e *FN* representa os resultados falsos negativos.

Tabela 3: Definição das medidas utilizadas para validar os estudos.

Medida	Definição
Acurácia (ou taxa de classificação)	$\frac{VP+VN}{P+N}$
Sensibilidade (ou taxa verdadeiro positivo)	$\frac{VP}{P}$
Especificidade (ou taxa verdadeiro negativo)	$\frac{VN}{N}$
Precisão (ou valor preditivo positivo)	$\frac{VP}{VP+FP}$
Valor preditivo negativo	$\frac{TN}{TN+FN}$
Fall-out (ou taxa de falso positivo)	$\frac{FP}{N}$
Taxa de falso negativo	$\frac{FN}{P}$
Coeficiente de correlação de Matthews	$\frac{VP.VN-FP.FN}{\sqrt{(VP+FP).(VP+FN).(VN+FP).(VN+FN)}}$
Área sob a curva ROC (AUC)	Área entre o eixo *x* e a curva ROC.

Fonte: O autor (2023).

Figura 8: Medidas utilizadas para validar os estudos analisados.



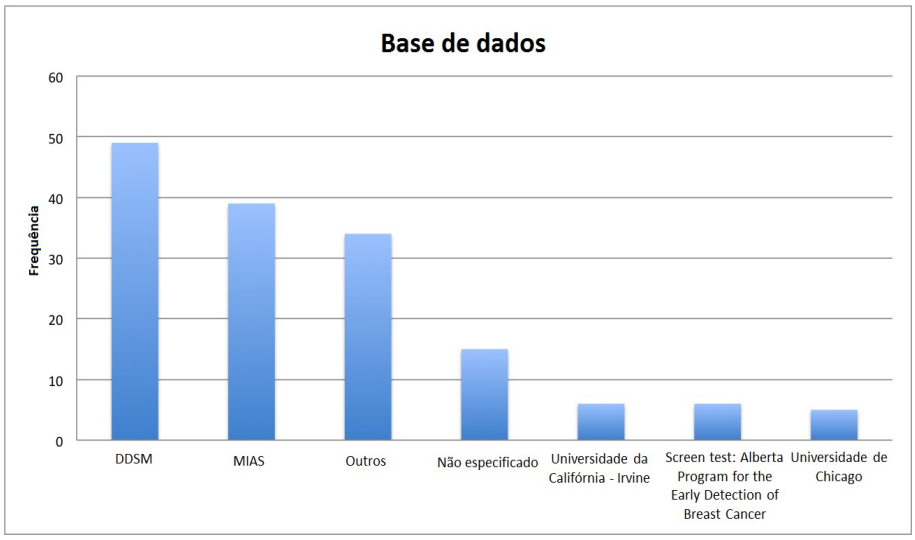
Fonte: O autor (2023).

3.1.1.2 Bases de dados

A Figura 9 apresenta as bases de dados mais comuns utilizadas nos estudos. As duas mais utilizadas foram DDSM, citada em 49 estudos (38%), e MIAS, com 39 estudos (30%). Ainda, 15 estudos (12%) não mencionaram a origem das imagens. DDSM

é uma base de dados pública provida pela *University of South Florida* que contém mais de 2.500 estudos (cada estudo inclui duas imagens de cada mama). As imagens disponíveis estão classificadas como normal, benigna e maligna. Para cada imagem um conjunto de pontos descrevendo o local do nódulo está disponível. MIAS também é uma base de dados pública que contêm 322 imagens (161 pares). As imagens contêm informações das coordenadas cartesianas do centro da anormalidade e o raio do círculo que engloba a anormalidade. Estas duas bases de dados são as mais utilizadas porque provêm um grande número de imagens e são de uso livre desde que sejam respeitados os acordos de licença. Por outro lado, alguns estudos utilizaram bases de dados privadas, tais como as bases de dados fornecidas pelo *Alberta Program for the Early Detection of Breast Cancer* e pela *University of Chicago*. Bases de dados privadas tendem a aparecer menos frequentemente nos estudos, uma vez que o acesso a elas é mais controlado.

Figura 9: Bases de dados mais utilizadas nos estudos analisados.



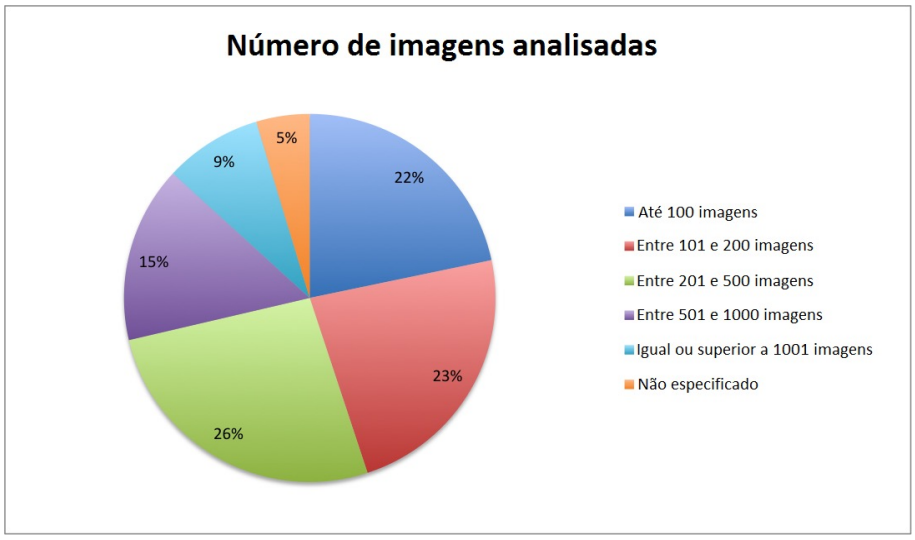
Fonte: O autor (2023).

Foi observado que dos artigos que especificaram o número de bases de dados utilizada, a maioria utilizou apenas uma base de dados (88 artigos - 77%), 20% dos artigos utilizaram duas bases de dados e apenas três (3%) estudos analisados utilizaram imagens provenientes de três ou mais fontes. Ainda, 15 artigos (12%) não especificaram o número de base de dados utilizadas para obtenção das imagens. Uma das dificuldades de trabalhar com diferentes bases de dados reside no fato de as imagens serem adquiridas utilizando equipamentos diversos e serem armazenadas com diferentes formatos e resoluções espaciais. Por isso, cada conjunto de dados demanda um tipo diferente de pré-processamento para que todas as imagens sejam padroniza-

das, o que, por si só, pode ser uma área de pesquisa. Alguns dos artigos analisados apresentaram estudos com relação a essa área, Rangayyan et al. (2010) apresentaram um estudo sobre o efeito da resolução do pixel no processo de classificação e em (HUO et al., 2000) foi exibido um estudo com relação a robustez do método de classificação quando exposto a imagens adquiridas a partir de diferentes aparelhos. Esta área deveria ser mais explorada, pois, dessa forma, as técnicas de reconhecimento de padrões poderiam ser utilizadas com imagens provenientes de diferentes fontes e ainda assim apresentar resultados consistentes.

A Figura 10 exibe o número de imagens utilizadas nos estudos. Pode-se observar que 28 artigos (22%) utilizaram até 100 imagens para treinar os classificadores e realizar os testes. Além disso, 30 artigos (23%) utilizaram entre 101 e 200 imagens e 34 artigos (26%) utilizaram entre 200 e 500 imagens. Ainda, seis artigos (5%) não especificaram o número de imagens utilizadas. Foi observado que 45% dos trabalhos utilizaram até 200 imagens. Por um lado, este fato pode mostrar que não é necessário um grande número de imagens para treinar os classificadores para resolver o problema de classificação de nódulos. Em (MCLEOD; VERMA, 2013) foram utilizadas 200 imagens e a acurácia obtida foi 98%. Por outro lado, classificadores construídos utilizando poucas imagens podem não ser genéricos suficientes para lidar com uma variedade de outras imagens que não foram consideradas no estudo.

Figura 10: Número de imagens analisadas por artigo.



Fonte: Fonte: O autor (2023).

Em geral, os estudos tentam balancear o número de imagens de cada classe para criar classificadores com menos viés. Entretanto, isto não é possível em todas as soluções. Para superar este problema de ter que lidar com número desbalanceado de

imagens, foram utilizadas em (LIMA; FILHO; SANTOS, 2016) combinações lineares com pesos aleatórios para gerar instâncias sintéticas de casos benignos e malignos para balanceamento do conjunto de dados. Outro trabalho que fez uso de imagens sintéticas foi (TRALIC; BOZEK; GRGIC, 2011), uma vez que os autores utilizaram imagens manualmente desenhadas por especialistas.

3.1.1.3 Características

Aspectos de textura e de forma foram as duas categorias de características mais utilizadas para discriminação dos nódulos. Características de textura podem ser extraídas utilizando várias técnicas como *gray-level co-occurrence matrix* (MOHANTY et al., 2013), *gray-level run-length matrix* (HAPFELMEIER; HORSCH, 2011), *gray-level aura matrix* (KANADAM; CHEREDDY, 2016), *run-length statistics matrix* (SAHINER et al., 1998) etc. Características de forma tentam descrever a forma do nódulo, por exemplo, sua área, perímetro, circularidade, concavidade ou convexidade, índice de espiculação etc. Os nódulos malignos tendem a ser mais irregulares e espiculados, enquanto que os benignos tendem a ser mais arredondados e ovalares (BASSETT, 1997). Devido a este fato, as características de forma são muito utilizadas no processo de classificação. Entretanto, esta categoria de característica requer um processo de segmentação eficiente, que seja capaz de separar corretamente o nódulo do fundo da região de interesse, preservando a borda do mesmo.

A segmentação automática dos nódulos pode ser uma tarefa bastante desafiadora, especialmente em imagens densas. Por esta razão, alguns nódulos precisam ser manualmente segmentadas por radiologistas experientes. Para evitar este problema, muitos trabalhos utilizaram apenas características de textura (MOHANTY et al., 2013; MISHRA; RANGANATHAN, 2014; KANADAM; CHEREDDY, 2016). Quando decide-se utilizar apenas características de forma, além de ser necessário um método eficiente de segmentação, é difícil lidar com cenários nos quais os nódulos malignos têm formato ovalar e os nódulos benignos têm formato irregular. Por outro lado, utilizar apenas características de textura não requer um método tão preciso de segmentação, mas as importantes informações da forma podem ser perdidas. Para que informações importantes não sejam perdidas, alguns autores combinam características de textura e de forma para discriminar os nódulos (DELOGU et al., 2007; HADJIISKI et al., 2001; DONG et al., 2015).

Outros autores tentaram uma abordagem diferente e não utilizaram características

extraídas diretamente das imagens. Como exemplo, em (MCLEOD; VERMA, 2013; WU et al., 2013) foram utilizadas informações providas pelos radiologistas que analisaram as imagens. Esta informação é fornecida na forma de descritores BI-RADS ou informações do paciente, por exemplo, idade, histórico de câncer na família, tratamento hormonal etc. Entretanto, alguns artigos apresentaram a combinação de características de forma, textura e textual para classificar os nódulos (GEORGIOU et al., 2007; VERMA, 2008; VELTHUIZEN; GANGADHARAN, 2000). Mesmo que a utilização de descritores BI-RADS possa melhorar a acurácia do classificador, é desencorajada a criação de um sistema de classificação baseado apenas nesta categoria de característica (ou a combinação com outras informações do paciente) para evitar que o sistema seja muito dependente do ser humano (PANCHAL; VERMA, 2006).

Alguns pesquisadores utilizaram características baseadas em *wavelets* para alimentar os classificadores e atingiram bons resultados. Entretanto, apesar dos bons resultados, um relativo alto custo de processamento está envolvido quando esta categoria de características é utilizado (BRUCE; KALLERGI; MENDOZA, 1999). A comparação entre a utilização de *wavelet* e *curvlet* foi apresentada em (ELTOUKHY; FAYE; SAMIR, 2010), sendo que as características baseadas em *curvelet* apresentaram uma melhor taxa de classificação. É importante notar que Eltoukhy, Faye e Samir (2010) foram os únicos pesquisadores que utilizaram características baseadas em *curvlet* e, como a taxa de acerto foi de 94.07%, esta característica poderia ser mais explorada.

Características baseadas em análise fractal foram utilizadas em alguns trabalhos (YANG et al., 2005; BEHESHTI et al., 2014). Estas características são úteis para classificar nódulos malignos em estágios iniciais, podendo ajudar os radiologistas a proverem um diagnóstico de câncer de mama mais precocemente (PANCHAL; VERMA, 2006).

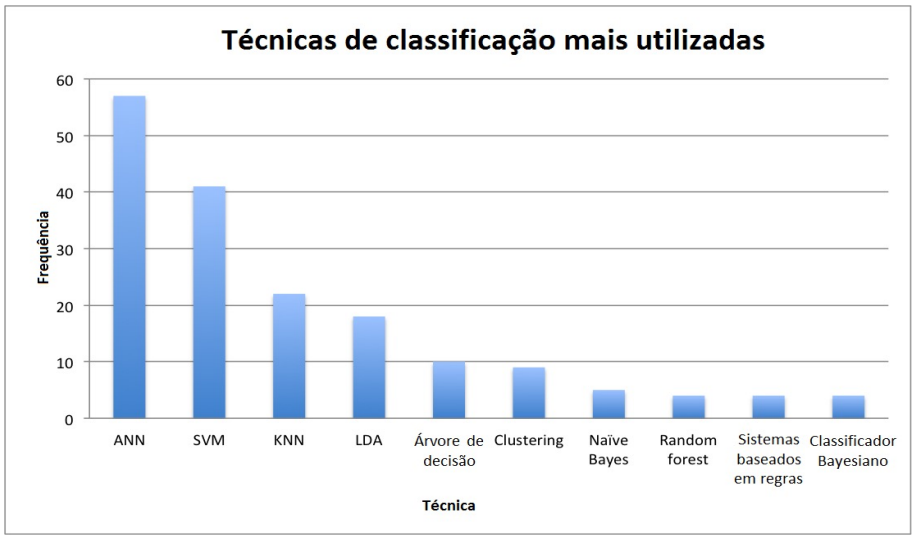
Conforme o número de características aumenta, é comum ocorrer o problema conhecido como “maldição da dimensionalidade”. Nesta situação, o modelo pode não apenas consumir mais memória e tempo para aprender um novo padrão, mas também sua acurácia pode diminuir se o número de exemplos não for grande o suficiente quando comparado ao número de características. Por se tratar de um problema bastante conhecido na área de aprendizado de máquina e reconhecimento de padrões, vários artigos apresentaram técnicas para selecionar as características mais úteis para discriminar os nódulos (KHAN et al., 2016; TAN; PU; ZHENG, 2014b; HAPFELMEIER; HORSCH, 2011; KHAN et al., 2017).

Algoritmos genéticos foram os mais utilizados para reduzir a dimensionalidade das características (8 artigos - 6%), seguidos por procedimentos de seleção que utilizam o critério *Wilks' lambda* e *Principal Component Analysis* (PCA) (6 artigos cada - 5%). Em (TAN; PU; ZHENG, 2014a), os autores implementaram um algoritmo chamado *Sequential Forward Floating Selection* (SFFS) que tem uma melhor eficiência com relação ao tempo (3% - 5%) quando comparado aos algoritmos genéticos. Os autores sugerem que, embora algoritmos genéticos sejam uma ferramenta poderosa para serem utilizados em sistemas CAD, eles são métodos computacionais bastante intensivos e a utilização do SFFS pode melhorar a eficiência da etapa de seleção de características.

3.1.1.4 Técnicas de classificação

As Figuras 11 e 12 mostram as técnicas mais comumente utilizadas nos artigos analisados. Redes neurais artificiais (*Artificial neural networks - ANN*) com suas variações (57 artigos - 44%) e Máquina de vetores de suporte (*Support vector machine - SVM*) com suas variações (41 artigos - 34%) são as duas técnicas mais empregadas. K-vizinhos mais próximos (*K-nearest neighbors - KNN*), a terceira mais frequente apareceu em 22 artigos (17%), seguida por *Linear discriminant analysis* (LDA) (18 artigos - 14%) e Árvores de decisão (10 estudos - 8%).

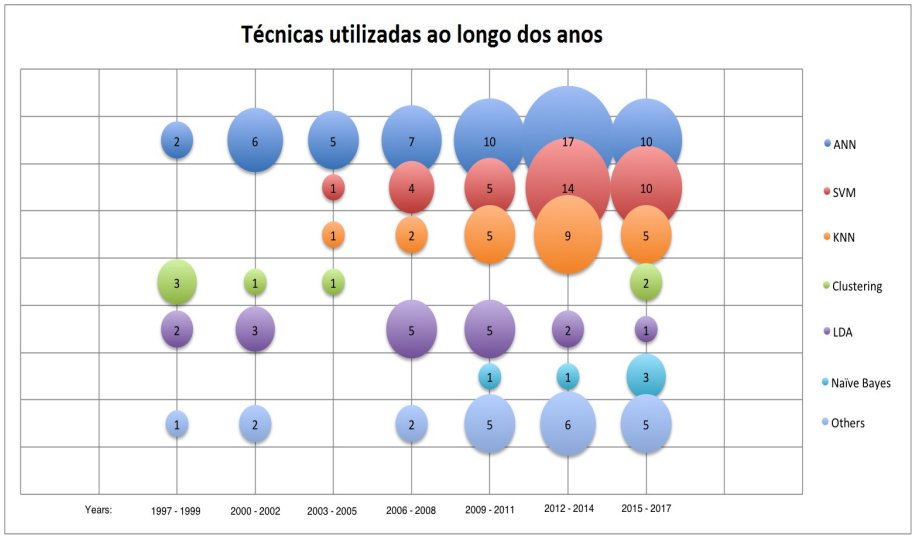
Figura 11: Técnicas de reconhecimento de padrões e aprendizado de máquina mais utilizadas.



Fonte: O autor (2023).

Apesar dos bons resultados obtidos pelas técnicas mais populares, por exemplo,

Figura 12: Técnicas utilizadas para discriminar nódulos ao longo dos anos.



Fonte: O autor (2023).

ANN foram utilizadas em (MCLEOD; VERMA, 2013) com uma acurácia de 98%; SVM foi empregado em (KHAN et al., 2016) e obteve a AUC de 0,948; KNN apareceu em (AROQUIARAJ; THANGAVEL, 2014) com AUC de 0.973, porém outras técnicas também foram capazes de atingir resultados similares. Por exemplo, a técnica Naïve Bayes foi explorada em apenas cinco estudos, mas em (BENNDORF et al., 2015) bons resultados foram encontrados em termos de AUC (0,935 e 0,876 dependendo das características utilizadas). Outro exemplo é a utilização de sistemas baseado em regras que apareceu apenas em (JAVADI; FAEZ, 2012; AL-NAJDAWI; BILTAWI; TEDMORI, 2015) com sensibilidade e especificidade superiores a 90%, o que mostra que existe oportunidade para esta técnica ser mais explorada. As técnicas *radial basis network/function, mixture model*, regressão logística, *artificial immune systems*, funções discriminantes com classificador Bayesiano, árvores de inferência condicional, programação genética e *squared discriminant analysis* também foram utilizadas nos estudos analisados, mas receberam pouca atenção neste contexto e poderiam ser mais exploradas.

Apesar de muitas técnicas terem sido utilizadas para fazer a classificação dos nódulos, não foram encontrados artigos que citam a utilização de gramáticas ou de métodos sintáticos para resolver tal problema. Especialmente na última década, gramáticas têm sido aplicadas no reconhecimento de padrões de imagens para reconhecimento e construção de objetos, em reconhecimento de *layouts* e segmentação (PEDRO; NUNES; MACHADO-LIMA, 2013). Gramáticas já foram aplicadas em imagens médicas para lidar com visualização em 3D de vasos coronários (TRZUPEK; OGIELA; TADEU-

SIEWICZ, 2011) e para fazer a análise de fratura de ossos da perna (OGIELA; TADEUSIEWICZ; OGIELA, 2008). Fazendo uma análise exploratória visando a atualizar o conhecimento sobre o estado da arte da área (Seção 3.1.2), foi encontrado apenas o artigo (TAHMASBI; SAKI; SHOKOUHI, 2011) que utilizou gramáticas em conjunto com ANN para discriminar nódulos alcançando uma acurácia de 91,38% e AUC de 0,858.

3.1.2 Literatura recente

Esta seção apresenta uma análise exploratória da literatura recente realizada nas bases SPIE Digital Library, IEEE Xplorer, Springer Link e Medical Physics sobre os trabalhos publicados na área de classificação de nódulos encontrados nas mamografias.

Rodríguez-Esparza et al. (2020) apresentaram uma abordagem de segmentação multinível baseada no algoritmo de meta-heurística *minimum cross-entropy threshold - Harris Hawks Optimization (MCET - HHO)* para identificar regiões na mama (ROI) que possuíssem tecido anormal. Em seguida, essa região era classificada de forma automática por um classificador baseado na técnica *bag-of-visual-words (BoVW)* para identificar se a ROI era de um tecido saudável, um nódulo benigno ou um nódulo maligno. A abordagem utilizada obteve uma acurácia média de 86% com a amostra de treinamento e de 73% com a amostra de teste.

Foram encontrados alguns trabalhos que fizeram a classificação dos nódulos utilizando SVM (HEIDARI et al., 2020; SONG; LI; WANG, 2020; PEZESHKI et al., 2021). Heidari et al. (2020) extraíram características das imagens MLO e CC considerando toda a região da mama (não apenas a região de interesse) para alimentar um classificador SVM visando a auxiliar os radiologistas a aumentarem a acurácia do diagnóstico na predição de lesões malignas. Utilizar características globais pode ser uma abordagem importante, pois evita que seja necessário segmentar o nódulo de forma automática ou manual. As características foram extraídas utilizando a técnica hierárquica *structural similarity index measure (SSIM)* que é um método para calcular a similaridade entre duas imagens (visões MLO e CC). A AUC provida pelo classificador foi de 0,84. Em (SONG; LI; WANG, 2020) foi apresentada uma estratégia que utiliza não apenas as características extraídas a partir de Redes neurais convolucionais profundas, mas também características extraídas a partir de *gray-level co-occurrence matrix (GLCM)* e de *histogram of oriented gradient (HOG)*. As características foram utilizadas como entrada para um classificador SVM e para um XGBoost (*eXtreme Gradient*

Boosting). O XGBoost apresentou melhores resultados quando comparado ao SVM atingindo uma acurácia de 92%. Pezeshki et al. (2021) apresentaram uma nova forma de segmentar o nódulo avaliando a similaridade dos pixels da região do centro do nódulo e a dissimilaridade dos pixels das regiões das espículas. Em seguida, características baseadas em estatísticas, forma, texturas, níveis de cinza e dimensão fractal foram extraídas e servidas como entrada para um classificador SVM. A AUC obtida foi de 0,96 para imagens do *dataset* MIAS e de 0,97 para imagens do *dataset* DDSM.

Técnicas baseadas em *deep learning* também estão sendo empregadas na classificação dos nódulos. Saber et al. (2021) apresentaram técnicas de *deep learning* e *transfer learning* para detecção e classificação de câncer de mama nas quais as características foram extraídas de imagens provenientes do *dataset* MIAS utilizando Redes convolucionais pré-treinadas, tais como *Inception V3*, *ResNet50*, *Visual Geometry Group networks (VGG)-19*, *VGG-16*, e *Inception-V2 ResNet*. Os resultados experimentais mostraram que o modelo *VGG-16* obteve AUC de aproximadamente 0,99 durante a classificação dos nódulos, sendo superior aos demais modelos. *Deep learning* e *transfer learning* também foram utilizados em (ARORA; RAI; RAMAN, 2020) em uma abordagem de *ensemble* em conjunto com um classificador baseado em rede neural para extração automática das características e classificação dos nódulos encontrados nos mamogramas. As características extraídas são otimizadas em um vetor de características e então servidas como entrada para o classificador. A acurácia e a AUC obtidas com a abordagem proposta foram de 88%. Em (MALEBARY; HASHMI, 2021) é apresentado um modelo que combina várias técnicas diferentes para classificar os nódulos como normal, benigno ou maligno. Para calcular os pesos de cada característica foi utilizada uma Rede neural recorrente do tipo *Long Short-Term Memory*. Uma Rede neural convolucional foi integrada ao modelo para extrair as características enquanto a técnica *transfer learning* foi introduzida para melhorar a Rede convolucional pré-treinada. Para reduzir a variância e generalização do erro foi empregada a técnica de *Random Forest* e de *boosting*. O modelo proposto foi utilizado na classificação das imagens disponíveis nos *datasets* DDSM e MIAS, sendo as acurácias obtidas 96% e 95% respectivamente.

A combinação de duas Redes neurais convolucionais para extração das características das imagens que foram utilizadas na classificação dos nódulos benignos e malignos foi proposta em (NIU et al., 2021). As características foram extraídas considerando *patches* globais e locais. Um método de atenção foi utilizado nas duas redes de extração de características, dando um maior foco para as características mais úteis

e suprimindo as menos úteis. O modelo proposto atingiu uma acurácia de 96% com imagens da DDSM e de 95% nas imagens da MIAS.

A análise dos artigos mais recentes possibilitou perceber que o uso de técnicas baseadas em *deep learning* tem crescido nesta área de pesquisa. Em especial, estas abordagens são bastante utilizadas para a extração de características uma vez que elas não requerem uma segmentação prévia do nódulo a ser classificado. Algo bastante comum quando o *deep learning* é aplicado é a necessidade de utilizar alguma técnica para aumentar o número de imagens, pois com um pequeno número de imagens não é possível aprender, de forma adequada, a importância de cada característica extraída. Entretanto, técnicas tradicionais de aprendizado de máquina continuam sendo bastante empregadas, destacando-se SVM.

3.1.3 Lacunas e desafios

Como foi percebido durante a revisão bibliográfica realizada, a maioria dos trabalhos utilizou apenas uma base de dados. Uma possível lacuna nesta área de pesquisa está na criação de sistemas mais robustos que sejam capazes de manipular uma grande variedade de base de dados, aumentando sua utilização por profissionais da área de saúde no dia a dia.

Outra lacuna está relacionada à criação de sistemas que podem aprender ou melhorar seu conhecimento de forma *on-line*. Os classificadores citados nos artigos só podem aprender um padrão de forma *off-line* ou *batch*. A criação de sistemas capazes de aprender novos padrões conforme novas imagens de nódulos são apresentadas de forma *on-line* seria bastante útil para esta área de pesquisa.

Nenhum dos artigos analisados apresentou a utilização de gramáticas, métodos sintáticos ou grafos para a classificação dos nódulos. Muitos pesquisadores têm utilizado estes métodos para lidar com problemas de classificação em imagens, mas estas técnicas não estão sendo exploradas para a classificação de imagens médicas e, em especial, em imagens mamográficas contendo nódulos. Esta lacuna pode ser explorada.

Ao mesmo tempo que algumas lacunas foram encontradas, também é possível perceber problemas recorrentes em alguns dos trabalhos analisados. O primeiro problema é a falta de padronização com relação às medidas de desempenho utilizadas para validar o estudo. Enquanto a acurácia e a AUC foram as medidas mais utiliza-

das, também foram encontradas várias outras medidas, por exemplo, sensibilidade, especificidade, precisão, taxa de falso positivo, taxa de falso negativo etc. O problema identificado é que mesmo que as medidas se relacionem de alguma forma, é difícil comparar estudos que apresentaram os resultados em termos de sensibilidade e especificidade com outros trabalhos que apresentaram os resultados na forma de AUC. Esta dificuldade acaba tornando bastante difícil encontrar o estado da arte nesta área de pesquisa.

Outro problema bastante comum é a ausência de comparação entre os resultados obtidos pelo classificador e os resultados obtidos pelos médicos. Por exemplo, uma acurácia de 90% obtida por um classificador pode ser boa ou ruim dependendo da acurácia que seria obtida por um médico analisando as mesmas imagens, caso contrário o sistema não poderia ser utilizado no dia a dia por profissionais da área de saúde. Adicionalmente, muitos artigos não deixam claro quais métodos foram utilizados durante a avaliação dos modelos. Em alguns casos é até mesmo difícil saber se as mesmas imagens foram utilizadas tanto na fase de treinamento quanto na fase de testes. Esta informação é bastante útil, principalmente porque, quando os testes são executados com as mesmas imagens utilizadas durante o treinamento, os resultados tendem a ser melhores.

Dada a grande variedade de métodos e técnicas, é bastante difícil fazer uma comparação quantitativa dos resultados de um artigo com os resultados de outros artigos, uma vez que eles variam nos mais diferentes aspectos. Para que uma comparação adequada seja realizada, deveria ser utilizada uma base de dados comum de tal forma que as mesmas imagens fossem empregadas na validação das diferentes abordagens. Ainda, além de utilizar a mesma base de dados em comum, as mesmas imagens devem ser utilizadas na etapa de treinamento dos diferentes modelos e, da mesma forma, as mesmas imagens devem ser utilizadas na etapa de validação desses modelos (exatamente os mesmos *folds* deveriam ser utilizados no caso de uma validação cruzada, por exemplo).

3.2 Representação de nódulos

A revisão da literatura foi realizada em junho de 2018 e novamente em julho de 2021 e teve como objetivo responder as seguintes perguntas:

- Quais técnicas são utilizadas para representar nódulos benignos e malignos em

mamogramas?

- Mamogramas sintéticos são úteis para a comunidade científica?
- Gramáticas estão sendo utilizadas em imagens médicas para modelar estruturas do corpo humano?

Para responder as perguntas foram utilizadas *strings* de busca com as seguintes palavras: **"breast cancer"**, **synthetic**, **mammography**, **breast**, **"synthetic image"**, **generation**, **phantoms**, **grammar** e **images**. Para viabilizar a análise dos artigos, foram considerados os 100 primeiros artigos de cada biblioteca digital ordenados por relevância.

3.2.1 Análise global

Em (KIM; KIM; RO, 2014) foi criada uma abordagem para geração de imagens sintéticas bidimensionais (2D) a partir de um volume tridimensional (3D) obtido por meio de uma tomossíntese digital da mama (*digital breast tomosynthesis - DBT*). De acordo com os autores, o objetivo almejado com a criação de imagens sintéticas foi reduzir o nível de radiação necessário para aquisição dos dois tipos de imagens, a DBT e a *full-field digital mammography* (FFDM). A abordagem desenvolvida faz uso da conspicuidade (medida de quanto um objeto é discernível do fundo no qual ele está inserido) dos voxels do volume 3D da DBT para melhorar a conspicuidade das lesões nas imagens 2D geradas sinteticamente. A conspicuidade de uma lesão mostra o quanto uma dada lesão pode ser discernida do seu entorno. Para validar o estudo, os autores extraíram quatro índices de qualidade da imagem (*global sharpness, sharpness of mass boundary, contrast* e *contrast to noise ratio*) e verificaram que a abordagem proposta obteve melhores resultados que a abordagem conhecida como *maximum intensity projection*. Estudos semelhantes foram apresentados em (MARISCOTTI et al., 2017; CHOI et al., 2018). Em (MARISCOTTI et al., 2017), os autores concluíram que as análises realizadas utilizando as imagens sintéticas apresentam um desempenho interpretativo similar ao FFDM, confirmando seu papel importante como uma alternativa ao FFDM em mulheres que fazem tomossíntese mamária. Em (CHOI et al., 2018), os resultados indicaram que o diagnóstico utilizando imagens sintéticas e mamografia digital, com ou sem o auxílio das imagens DBT, para a avaliação de microcalcificações, não foram significativamente diferentes.

Outro trabalho que criou mamogramas sintéticos (2D) a partir de volumes obtidos de DBT foi (SCHIE et al., 2013). O objetivo foi criar imagens em 2D nas quais os nódulos e as distorções arquiteturais fossem melhor visualizadas. Inicialmente a abordagem utilizada encontra os pontos relevantes no volume DBT por meio de um sistema de CAD e, em seguida, renderiza um mamograma a partir da intersecção da superfície ajustada desses pontos e do volume DBT. Os resultados indicaram AUC de 0,85, enquanto que a AUC obtida nos mamogramas convencionais foi de 0,81 no que diz respeito à tarefa de detecção dos nódulos.

Em (CHOI et al., 2016) o objetivo foi comparar o desempenho do diagnóstico obtido a partir de imagens sintéticas em 2D construídas a partir de DBT, com o desempenho do diagnóstico obtido a partir de imagens FFDM na detecção de câncer de mama. A avaliação das imagens foi realizada por três radiologistas e os resultados não mostraram diferença significativa em termos de sensibilidade, mas um dos observadores mostrou uma maior especificidade utilizando as imagens sintéticas.

No estudo (BLIZNAKOVA; PALLIKARAKIS, 2009) foi criado um modelo de mama que inclui a forma da mama, o sistema de ductos, os ligamentos de Cooper, o músculo peitoral, a textura da mama e possíveis anormalidades tais como nódulos ou microcalcificações. No modelo proposto, todos os objetos são transformados em valores de voxels em uma matriz de textura. As imagens mamográficas foram sintetizadas utilizando a ferramenta *XRayImagingSimulator*. A avaliação da abordagem foi realizada mediante a comparação de algumas características (*fractal dimension, skewness, kurtosis* e *power law spectral analysis*) obtidos a partir das imagens sintéticas e das imagens reais. Os resultados mostraram que a abordagem utilizada produz modelos de mama bastante realísticos.

Em (TAYLOR; OWENS; INGRAM, 1998) foi apresentada uma abordagem recursiva para simular o crescimento do sistema de dutos dentro do volume da mama, utilizando conceito de crescimento de árvores e sub-árvores. A mama simulada foi projetada no espaço bidimensional com a finalidade de possibilitar uma análise quantitativa comparando-a com a imagem da mama real utilizando o espectro de Fourier. Os resultados mostraram que o modelo sintético apresentou propriedades texturais que são similares a imagens reais.

Em (NAPPI et al., 2001) foi exibido um *framework* para simulação 3D de calcificações na mama. As calcificações criadas podem ser colocadas em uma mamografia real ou simulada de tal forma que possam ser utilizadas como exemplos para radio-

logistas. Para validar o experimento, os autores exibiram as imagens geradas para quatro radiologistas experientes e solicitaram que eles avaliassem as imagens com base nas seguintes respostas: i) sem calcificações; ii) definitivamente calcificações reais; iii) provavelmente calcificações reais; iv) pode ser uma das duas; v) provavelmente calcificações simuladas; vi) definitivamente calcificações simuladas. A maioria (18/30) das calcificações simuladas foram consideradas reais por pelo menos um dos radiologistas.

Além dos trabalhos já mencionados nesta revisão, foram encontrados alguns estudos que utilizam gramáticas para lidar com problemas na área médica, mas que não resolvem problemas relacionados a imagens mamográficas. O trabalho (GALARRETA-VALVERDE et al., 2013) propôs a geração de vasos sanguíneos em 3D utilizando gramáticas estocásticas do tipo *L-systems*, visando a prover material que pudessem ser utilizados para testar novos métodos de segmentação deste tipo de imagem. Os vasos gerados simulam características naturais tais como bifurcações, tamanho médio e diâmetro, além de algumas anomalias como aneurismas e estenoses. Já em (GALARRETA-VALVERDE et al., 2015), *L-systems* foram utilizados para representar árvores vasculares de imagens angiográficas e para descrever uma medida baseada na taxonomia Tokunaga capaz de diferenciar arquiteturas vasculares distintas. Os autores concluíram que a metodologia aplicada não apenas possibilitava uma representação da arquitetura vascular de forma compacta, mas também fornecia uma medida quantitativa da complexidade da bifurcação capaz de caracterizar diferentes tipos de anomalias vasculares.

Trzupek e Ogiela (2014) criaram uma descrição semântica das topologias das artérias coronárias utilizando um grafo gramatical ETPL{k} e uma descrição semântica de seções individuais das artérias coronárias por meio de uma gramática livre de contexto. O objetivo do trabalho foi desenvolver uma abordagem linguística para modelar as artérias coronárias e utilizar esta abordagem em sistemas de recuperação de imagens. Já em (HAMDI et al., 2011) o trabalho consiste em segmentar imagens cardíacas obtidas por meio de ultrassom e fazer estimativa das cavidades do coração mediante o uso de uma abordagem baseada em gramáticas.

Três trabalhos (CAULKIN; ASTLEY, 1999; SAUNDERS et al., 2006; SHEN et al., 2021) encontrados abordaram especificamente o problema de geração de nódulos sintéticos em mamogramas, com propostas diferentes entre si, sendo que somente um deles é recente.

Em (CAULKIN; ASTLEY, 1999) foi realizada uma análise para selecionar as localizações mais realistas onde um nódulo frequentemente aparece e a conclusão foi que a região mais frequente é o quadrante superior externo da mama. O método proposto para geração do nódulo consiste inicialmente em extrair as regiões de interesse que possuem um nódulo, sendo que os níveis de cinza dentro do nódulo devem ser estimados utilizando apenas os pixels que estão fora do nódulo. Para realizar tal inferência é utilizada uma *thin plate spline* para interpolar os pixels no interior do nódulo. O formato do nódulo é definido a partir de um modelo estatístico construído considerando um conjunto de imagens que possuíam um consistente conjunto de pontos de referência. Todos os pontos de referência são alinhados considerando coordenadas comuns de tal forma que cada forma pode ser representada por um vetor que será utilizado na construção do modelo. Os autores afirmam que a abordagem utilizada apresenta limitações como: estruturas que aparecem apenas dentro dos nódulos não são modeladas, lesões muito grandes aparecem de forma mais clara nas imagens, há perda de resolução devido ao processo de amostragem de níveis de cinza quando são consideradas lesões de tamanhos diferentes. Além disso, na técnica proposta as espículas dos nódulos não foram consideradas.

Uma abordagem para simular lesões em mamografias (nódulos e calcificações) foi apresentada em (SAUNDERS et al., 2006). Para a simulação dos nódulos, inicialmente é necessário segmentar a lesão, examinar o perfil de contraste do nódulo em termos do gradiente das bordas e mensurar as propriedades das bordas considerando o perfil de desvio dessas bordas. A segmentação foi realizada utilizando o método *Laplacian of Gaussian* para detecção de bordas. O perfil de contraste do nódulo foi mensurado computando o nível médio de cinza dos pixels pertencentes aos anéis de elipses com o mesmo centro e mesmo ângulo de orientação. Como um nódulo não tem exatamente o mesmo formato de uma elipse, foi criado um perfil de desvio das bordas considerando uma máscara com o formato do nódulo e a correspondente elipse que melhor se encaixa nesta máscara. As imagens geradas foram analisadas e comparadas com imagens por três especialistas e a AUC quando considerado os nódulos benignos e malignos foram 0,68 e 0,65, respectivamente.

O conceito de *Generative adversarial networks* (GAN) foi utilizado em (SHEN et al., 2021) no processo de geração de nódulos sintéticos malignos em mamogramas. O primeiro passo para gerar os nódulos é composto por uma *Deep Convolutional GAN* (DGAN), treinada para aprender o formato de nódulos reais e gerar várias máscaras com formatos semelhantes a partir de entradas aleatórias. No segundo passo, é ex-

traído uma ROI de um mamograma normal onde não há nenhuma lesão. Ao mesmo tempo, a máscara gerada é inserida em um fundo preto com o mesmo tamanho do mamograma normal sendo colocada no mesmo local da ROI extraída na imagem do mamograma normal. Em seguida, a ROI é preenchida com ruídos uniformemente aleatórios no valor de 0 a 255 (nível de cinza). No quarto passo, a ROI com ruído, a máscara gerada e sua borda são combinadas formando uma pilha de três canais que é fornecida para uma GAN treinada para gerar o nódulo que, por sua vez, é inserido no mamograma normal na posição original da ROI. A abordagem proposta foi utilizada para aumentar o número de imagens utilizadas para treinamento de modelos de detecção de nódulos. Os resultados mostraram um incremento de 5,03% na taxa de detecção dos nódulos reais quando o treinamento fez uso das imagens sintéticas se comparado com modelos que fizeram uso apenas das imagens reais.

3.2.2 Discussão e conclusão

Como pode-se perceber nesta revisão, vários trabalhos geraram imagens sintéticas de mamogramas. Os objetivos de alguns destes trabalhos eram desenvolver técnicas para criar as imagens sintéticas a partir de volumes obtidos por meio de tomossíntese digital da mama.

Outros trabalhos tiveram foco na comparação do desempenho da classificação utilizando imagens sintéticas geradas a partir de DBT e de imagens FFDM. Também foram encontrados trabalhos que fazem modelagens 3D da mama (sistema de ductos e calcificações) e depois fazem uma projeção 2D dessas imagens e as submetem para avaliação de radiologistas experientes.

Em especial, foram encontrados alguns trabalhos que geraram nódulos sintéticos em mamogramas com o intuito de enriquecer os *datasets*. Dos trabalhos analisados na revisão bibliográfica o estudo (SHEN et al., 2021) foi o que apresentou melhor resultado, sendo que a utilização das imagens sintéticas de nódulos malignos geradas pela abordagem elevou a taxa de detecção dos nódulos reais.

No que diz respeito à utilização de gramáticas, somente três trabalhos foram encontrados para modelar vasos sanguíneos. É importante notar que, após realizada a revisão bibliográfica, não foram encontrados trabalhos que utilizassem gramáticas nas etapas de criação de nódulos benignos e malignos de forma sintética. O trabalho analisado com o melhor desempenho nesta área fez uso de *deep learning* o que demanda um grande número de imagens para que o padrão possa ser aprendido. Desta

forma, podemos considerar este campo com alto potencial de exploração em trabalhos futuros.

3.3 Considerações

Neste capítulo foram apresentados os resultados da revisão sistemática conduzida com o intuito de compreender o cenário das pesquisas científicas na classificação dos nódulos encontrados nos mamogramas. Também foram disponibilizados os resultados da revisão com vistas a identificar as pesquisas acerca da representação de nódulos sintéticos nos mamogramas.

Para a classificação dos nódulos, as técnicas mais utilizadas atualmente são ANN, SVM e KNN. As imagens da mama criadas de forma sintética são, em geral, o resultado de projeções 2D de imagens DBT ou imagens criadas a partir de *phantoms*. Embora técnicas de *deep learning* estejam sendo mais recentemente exploradas, o número de trabalhos encontrados na área do presente projeto ainda é menor quando comparado com técnicas tradicionais de aprendizado de máquina.

Embora muitos trabalhos tenham sido publicados nos últimos anos considerando estas duas áreas de pesquisa (classificação e representação de estruturas em mamogramas sintéticos), nenhum dos estudos encontrados utilizou métodos sintáticos para atingir seus objetivos. Neste sentido, a exploração da utilização de gramáticas se faz pertinente nesta área de pesquisa. Os resultados compilados da revisão sistemática realizada foram publicados no artigo (PEDRO; MACHADO-LIMA; NUNES, 2019a).

O próximo capítulo detalha os materiais e métodos utilizados neste projeto a fim de atingir os objetivos propostos no Capítulo 1.

PARTE II

PROJETO

4 MATERIAIS E MÉTODOS

O objetivo principal deste trabalho é utilizar métodos sintáticos para classificação de nódulos em imagens mamográficas e geração de imagens sintética de nódulos. Neste sentido, o propósito do presente capítulo é apresentar de forma detalhada os materiais (Seção 4.1) e os passos percorridos para atingir os objetivos desta pesquisa (Seção 4.2).

4.1 Materiais

Esta seção descreve quais materiais estão sendo utilizados neste projeto a fim de atingir seus objetivos.

Foi percebido durante a revisão de literatura que as duas bases de dados mais utilizadas são a DDSM (HEATH et al., 2000) e a MIAS (SUCKLING et al., 1994). Embora essas duas bases de dados sejam públicas, elas não foram utilizadas neste projeto pelos seguintes motivos: i) na base DDSM os nódulos estão demarcados nas imagens, mas em várias imagens a demarcação não é precisa (é fornecido um arquivo texto com a posição espacial de cada pixel referente à demarcação realizada), o que dificultaria o processo de extração das características; ii) a base MIAS provê as imagens, mas as informações sobre os nódulos são insuficientes para reconstruí-los com precisão (são fornecidos a posição espacial do centróide no nódulo e o tamanho do raio cuja circunferência engloba todo o nódulo).

A partir do exposto, neste projeto de pesquisa foi utilizado um *dataset* (*Screen Test: Alberta Program for the Early Detection of Breast Cancer*) fornecido por pesquisadores da *University of Calgary* - Canadá, que a partir de agora será chamado de *dataset ST*. Este *dataset* foi utilizado anteriormente nos trabalhos (MUDIGONDA; RANGAYYAN; DESAUTELS, 1999; RANGAYYAN; MUDIGONDA; DESAUTELS, 2000; MUDIGONDA; RANGAYYAN; DESAUTELS, 2000; NANDI et al., 2006; RANGAYYAN; NGUYEN, 2007) e possui um conjunto de imagens de uma base de dados local e um

conjunto de imagens do banco de dados MIAS com informações precisas dos nódulos descrito na seção 4.1.1. Além disso, foi construído um novo *dataset* com imagens fornecidas por pesquisadores do A. C. Camargo Cancer Center que a partir de agora será chamado de *dataset ACC* descrito na seção 4.1.2. O diagnóstico das imagens de ambos os *datasets* foram obtidos mediante realização de biópsias.

4.1.1 *Dataset* ST

O *dataset* em questão é formado por arquivos em formato de texto (um arquivo para cada imagem) com as coordenadas dos pixels das bordas dos nódulos, além de um único arquivo no formato Microsoft Excel contendo todas as características extraídas. Neste arquivo Microsoft Excel estão disponíveis as seguintes informações: i) um identificador do nódulo; ii) a base de dados utilizada (MIAS ou local); iii) a classe do nódulo (maligno e benigno no caso das imagens da base local, e maligno circunscrito (MC), maligno espiculado (ME), benigno circunscrito (BC) e benigno espiculado (BE) no caso das imagens da base MIAS); iv) características referentes à forma do nódulo; e v) características referentes à textura/gradiente. O conjunto de imagens é formado por 54 imagens da base de dados MIAS e 57 imagens da base de dados local, com tamanho de pixel de 50μm. Os contornos dos nódulos foram demarcados por médicos experientes. Pesquisadores da Universidade Federal de São Paulo (UNIFESP) classificaram o tipo da borda do nódulo das imagens da base de dados local como sendo circunscrito ou espiculado. Desta forma, todas as imagens deste *dataset* possuem as classes MC, ME, BC ou BE. A Tabela 4 exibe a distribuição das imagens de acordo com as classes e tipos de borda.

Tabela 4: Distribuição dos nódulos do *dataset* ST de acordo com sua classe e tipo de borda.

Classe e tipo de borda	**Quantidade**
Benigno circunscrito (BC)	53
Benigno espiculado (BE)	12
Maligno circunscrito (MC)	10
Maligno espiculado (ME)	36

Fonte: O autor (2023)

4.1.2 *Dataset* ACC

O *dataset* ACC foi construído com 202 imagens de nódulos com tamanho de pixel de 70μm fornecidas por pesquisadores do A. C. Camargo Cancer Center. Destas imagens, 104 nódulos são benignos e 98 são malignos e tiveram os contornos demarcados por dois médicos residentes de radiologia, orientados por um médico radiologista do A. C. Camargo Cancer Center com 15 anos de experiência neste tema. Além disso, os médicos radiologistas forneceram a informação sobre a classificação dos contornos, categorizando-os em espiculados ou não espiculados. Desta forma, as nódulos foram classificados como maligno espiculado (ME), benigno espiculado (BE) e benigno não espiculado (BnE). É importante ressaltar que este *dataset* não contém nódulos malignos não espiculado (MnE). A Tabela 5 exibe a distribuição dos nódulos de acordo com suas classes e tipo de borda.

Tabela 5: Distribuição dos nódulos do *dataset* ACC de acordo com sua classe e tipo de borda.

Classe e tipo de borda	**Quantidade**
Benigno não espiculado (BnE)	99
Benigno espiculado (BE)	5
Maligno não espiculado (MnE)	0
Maligno espiculado (ME)	98

Fonte: O autor (2023)

É importante ressaltar a diferenciação dos tipos de borda fornecidos pelos pesquisadores que classificaram as imagens dos *datasets* ST e ACC. Ambos *datasets* possuem nódulos com tipos de borda “espiculado“ gerando as classes BE e ME. O *dataset* ST possui nódulos com tipos de borda “circunscrito“, gerando as classes BC e MC e o *dataset* ACC possui nódulos com tipos de bordas “não espiculado“ gerando as classes BnE e MnE. Por conta desta diferenciação, quando os *datasets* foram utilizados em conjunto no mesmo experimento, foram combinados os nódulos benignos circunscritos (BC) do *dataset* ST com os nódulos benignos não espiculados (BnE) do *dataset* ACC e, de forma semelhante, foram combinados os nódulos malignos circunscritos (MC) do *dataset* ST com os nódulos malignos não espiculados (MnE) do *dataset* ACC.

A Figura 13 exibe as bordas de alguns nódulos benignos e malignos.

Figura 13: Bordas de alguns nódulos benignos e malignos.

Fonte: O autor (2023).

4.2 Métodos

Esta seção descreve os métodos utilizados neste projeto para alcançar os objetos de pesquisa.

4.2.1 Visão geral dos métodos utilizados

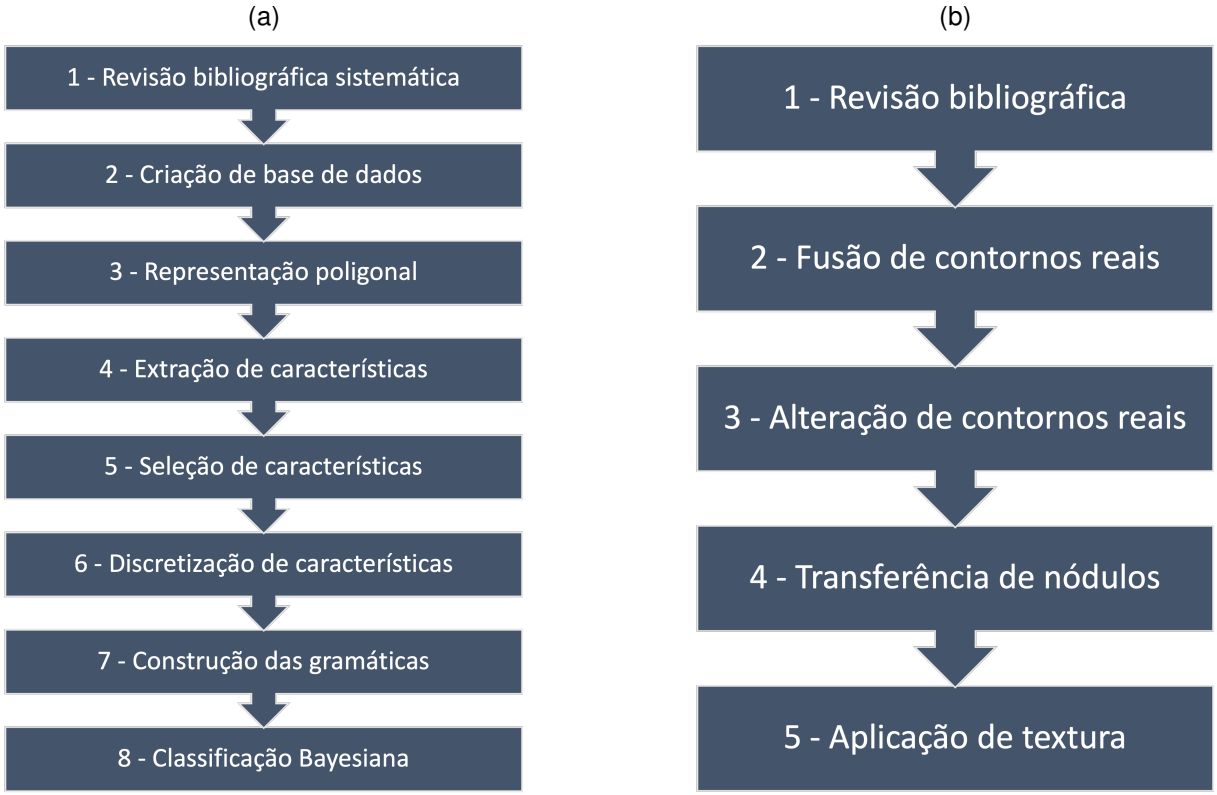
A Figura 14a exibe um diagrama com as atividades realizadas neste projeto de pesquisa no que diz respeito à classificação dos nódulos reais, enquanto a Figura 14b apresenta as atividades realizadas para a geração das imagens sintéticas.

Conforme apresentado na Figura 14a, as seguintes atividades foram executadas para a tarefa de classificação de nódulos:

1 - Revisão bibliográfica sistemática: foi realizada uma revisão sistemática da literatura cujo objetivo foi identificar quais classificadores e características são utilizados e quais resultados foram obtidos na tarefa de classificação dos nódulos. O conteúdo da revisão sistemática gerou a publicação do artigo (PEDRO; MACHADO-LIMA; NUNES, 2019a). A revisão bibliográfica sistemática foi apresentada na Seção 3.1.

2 - Criação de uma nova base de dados de imagens: uma nova base de dados

Figura 14: Atividades para classificação e geração de nódulos: (a) Atividades realizadas na etapa de classificação dos nódulos reais; (b) Atividades realizadas na etapa de geração de imagens sintéticas de nódulo.



Fonte: O autor (2023).

com 202 imagens foi criada durante este projeto de pesquisa, conforme apresentado na Seção 4.1.

3 - Representação poligonal dos nódulos: foi desenvolvido um método para criar uma representação poligonal dos nódulos. A representação poligonal é importante pois reduz a complexidade dos contornos, tornando o processo de extração de características menos custoso em termos de recursos computacionais. Os algoritmos utilizados e as representações poligonais são detalhados na Seção 5.1.

4 - Extração das características: foram extraídas características relacionadas a forma, textura e gradiente das imagens. Além disso, foram extraídos os momentos de Hu como características adicionais para os classificadores. O processo e as definições das características utilizadas são apresentados na Seção 5.2.

5 - Seleção das características: foi desenvolvido um processo baseado na importância de Gini para selecionar as características mais relevantes para a classificação dos nódulos. Os detalhes deste processo estão na Seção 5.3.

6 - Discretização das características: para que as características extraídas das imagens possam ser utilizadas em um classificador baseado em gramáticas, é necessário que elas sofram um processo de discretização. Neste caso, após a discretização, as características que possuíam um valor contínuo passam a serem representadas por um rótulo. Este processo é descrito na Seção 5.4.

7 - Construção das gramáticas: foram criados modelos hierárquicos (árvores) que pudessem ser utilizados para representar os nódulos de tal forma que estes modelos pudessem ser utilizados para a criação das gramáticas. Estas gramáticas, por sua vez, tiveram suas probabilidades estimadas e foram utilizadas para a criação de analisadores sintáticos que são a base dos classificadores (seções 5.5.1 a 5.5.3).

8 - Classificação Bayesiana: para classificar os nódulos como pertencentes à classe Benigno ou Maligno, foi construído um classificador Bayesiano composto por dois analisadores sintáticos. O classificador utilizado é descrito na Seção 5.5.4.

Conforme ilustrado na Figura 14b, as atividades realizadas para a geração de nódulos sintéticos foram:

1 - Revisão bibliográfica: inicialmente foi conduzida uma revisão bibliográfica da literatura para verificar quais técnicas são utilizadas para a representação de nódulos sintéticos em mamogramas, conforme detalhado na Seção 3.2.

2 - Fusão de contornos reais para geração de novos contornos: esta abordagem combinou contornos de nódulos reais para gerar contornos sintéticos. Os contornos mais semelhantes aos contornos de nódulos reais foram então selecionados mediante utilização das gramáticas criadas anteriormente durante o processo de classificação dos nódulos. Esta abordagem é descrita com detalhes na Seção 6.1.

3 - Alteração de contornos reais para geração de novos contornos: outra abordagem utilizada para gerar os contornos dos nódulos sintéticos foi alterar pequenas partes dos contornos dos nódulos reais. Mais precisamente, o contorno real foi dividido em 8 partes e uma ou mais partes do contorno foram alteradas. De forma semelhante à abordagem de fusão de contornos reais, foram utilizadas as gramáticas para selecionar os contornos gerados mais semelhantes aos contornos reais, conforme descrito na Seção 6.2.

4 - Transferência de um nódulo real para um novo mamograma: esta abordagem consiste em transferir um nódulo encontrado em um mamograma para um outro mamograma em uma posição na qual nenhum outro nódulo exista. Mais detalhes

sobre esta abordagem são encontrados na Seção 6.3.

5 - Aplicação de textura para os nódulos sintéticos: as abordagens utilizadas para a aplicação de textura nos nódulos sintéticos gerados são detalhadas na Seção 6.4.

4.2.2 Tecnologias utilizadas

Nesta seção são apresentadas as tecnologias utilizadas durante o desenvolvimento deste projeto.

Representação poligonal dos nódulos: os algoritmos utilizados para criar a representação poligonal dos nódulos foram implementados no formato de *scripts* utilizando a linguagem de programação *Python* na versão 3. Em especial, foi utilizada a biblioteca de código fonte aberto *OpenCV* (BRADSKI, 2000) que já possui uma implementação para *Python* do algoritmo RDP.

Extração das características: para realizar a extração das características dos nódulos representados pelos modelos poligonais, foram criados *scripts* utilizando a linguagem de programação *Python* na versão 3. A escolha desta linguagem de programação se deu devido aos recursos providos pela *OpenCV* (BRADSKI, 2000).

Seleção das características: algoritmo Random Forest foi utilizado para obter a importância de Gini utilizada no mecanismo de seleção de característica. A implementação foi feita utilizando a linguagem de programação *Python* versão 3 e utilizando a biblioteca *scikit-learn* (PEDREGOSA et al., 2011) devido à facilidade de uso e aos recursos providos pela biblioteca.

Discretização das características: os algoritmos para realizar a discretização das características utilizados neste projeto foram implementados utilizando a linguagem de programação *Java* (Ômega) e *Python* (KBinsDiscretizer). O algoritmo Ômega foi implementado em *JAVA* pois ele não estava encapsulado em nenhuma biblioteca disponível publicamente. Posteriormente, o algoritmo KBinsDiscretizer também foi utilizado neste projeto, mas optou-se por utilizar a implementação em *Python* disponível pela biblioteca *scikit-learn* (PEDREGOSA et al., 2011).

Gramáticas: o arcabouço para construção dos modelos gramaticais e para classificação dos nódulos (discretização das características - algoritmo Ômega, construção das gramáticas e classificação Bayesiana) foi implementado utilizando a linguagem *Java*, uma vez que não havia bibliotecas públicas disponíveis para construção dos

modelos e *JAVA* era a linguagem de maior conhecimento no momento da implementação.

Outros classificadores: no presente projeto também foram implementados outros modelos de aprendizado de máquina para que fossem feitas comparações com os métodos gramaticais. Os modelos ANN, KNN, SVM e RF foram implementados utilizados a linguagem *Python* e a biblioteca *scikit-learn* (PEDREGOSA et al., 2011). O modelo LGBM também foi implementado utilizando a linguagem *Python*, mas utilizou o *LightGBM Python-package* (PYTHON-PACKAGE, 2023).

Geração de imagens: todas as implementações feitas para a geração de imagens foram realizadas utilizando a linguagem de programação *Python* e a biblioteca *OpenCV* (BRADSKI, 2000).

4.3 Considerações finais

Neste capítulo foi apresentada uma visão geral dos materiais utilizados e dos métodos empregados durante o desenvolvimento do presente projeto.

Os capítulos a seguir detalham os métodos aqui apresentados. O Capítulo 5 apresenta as definições e o detalhamento da implementação em relação às técnicas desenvolvidas para classificação de nódulos e o Capítulo 6 detalha o processo de geração de imagens sintéticas de nódulos.

5 CLASSIFICAÇÃO DE NÓDULOS

Este capítulo detalha as atividades realizadas para a construção dos classificadores baseados em gramática e a classificação dos nódulos considerando as classes Benigno e Maligno. A representação poligonal dos nódulos é vista na seção 5.1, a extração das características é detalhada na seção 5.2, na seção 5.3 é explicado o processo de seleção de características, na seção 5.4 é exido o processo empregado para discretização das características, o processo classificação dos nódulos utilizando gramáticas é explicado na seção 5.5 e na seção 5.6 é detalhada a técnica utilizada para validação dos modelos criados.

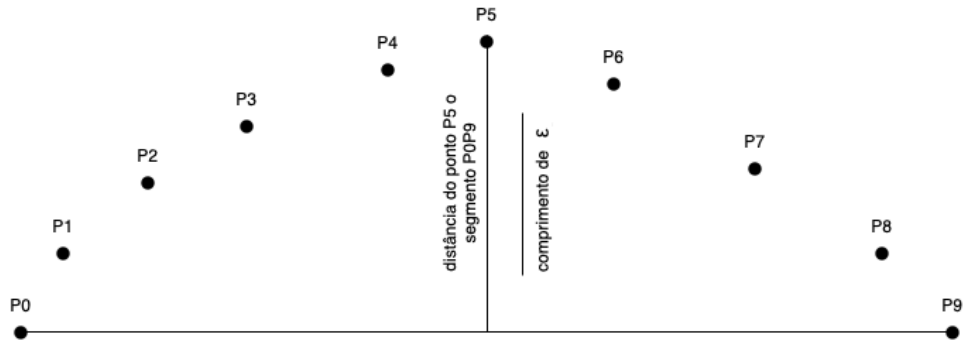
5.1 Representação poligonal dos nódulos

Dois algoritmos foram utilizados na criação dos modelos poligonais: i) Ramer-Douglas-Peucker (RDP) e, ii) the Peter Borne (PB) algorithms. A implementação desses algoritmos foi realizada durante um trabalho de iniciação científica e utilizada neste projeto de pesquisa.

O algoritmo RDP (RAMER, 1972) é capaz de gerar modelos poligonais a partir de um conjunto de pontos. O algoritmo pode ser descrito em quatro etapas: i) considere um conjunto de *N* pontos (10 nesse exemplo - P_0 to P_9, mas no caso dos nódulos, seriam todos os pontos (pixels) que fazem parte da borda do nódulo) e um parâmetro ϵ utilizado como limiar; ii) conecte o primeiro ao último ponto (P_0 e P_9) utilizando um segmento de reta e encontre um ponto pertencente à *N* que possua a maior distância até o segmento $\overline{P_0P_9}$. Caso a distância encontrada seja maior que ϵ, então este ponto será mantido na criação poligonal do modelo. Supondo que o ponto encontrado previamente seja o P_5, divida o segmento de reta $\overline{P_0P_9}$ em dois segmentos $\overline{P_0P_5}$ e $\overline{P_5P_9}$; iii) repita a etapa anterior para cada segmento de reta e, quando as distâncias dos pontos encontrados até o segmento de reta for maior que ϵ, mantenha esses pontos para formar o modelo poligonal, caso contrário, descarte-os; iv) repita esse processo

até que todos os pontos em *N* tenham sido verificados (LIU et al., 2011). Ao término da seleção e descarte dos pontos, conecte os pontos selecionados utilizando segmentos de retas para gerar o modelo poligonal. A Figura 15 ilustra a primeira iteração do algoritmo RDP considerando *N* = 10 na qual o ponto P5 é escolhido para fazer parte do conjunto final de pontos do modelo poligonal, pois a distância do ponto P5 ao segmento $\overline{P_0P_9}$ é maior que ϵ.

Figura 15: Primeira iteração do algoritmo RDP na qual o ponto P5 é escolhido para fazer parte do conjunto de pontos do modelo poligonal.

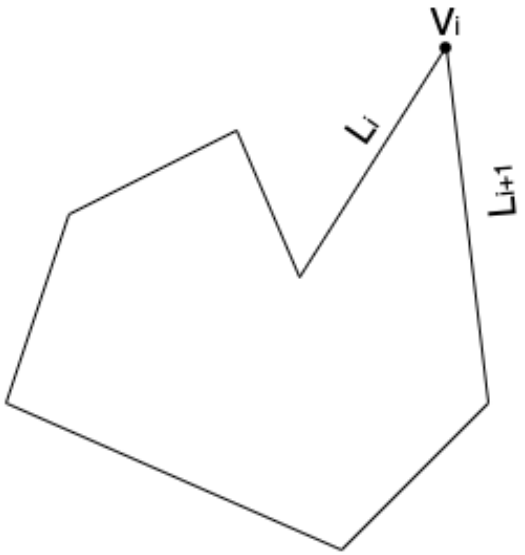


Fonte: O autor (2023).

O modelo poligonal gerado pelo algoritmo PB (BORNE, 2014) simplifica o contorno do nódulo descartando vértices menos importantes - o vértice pode ser qualquer ponto presente no contorno. A importância do vértice Vi é calculada utilizando a equação 5.1, na qual L_i e L_{i+1} são, respectivamente, os vetores correspondentes aos segmentos de retas que se encontram no vértice Vi, e S_i é o produto dos comprimentos desses segmentos. A Figura 16 exemplifica o vértice Vi e os segmentos L_i e L_{i+1} de um dado contorno a ser simplificado. Este algoritmo pode ser descrito considerando as seguintes fases: i) considere um conjunto de pontos *N* (por exemplo, os pontos (pixels) que fazem parte da borda do nódulo) e um número inteiro *T* que é o maior número de pontos que formará o modelo poligonal final; ii) calcule a importância de cada vértice considerando todos os vértices do contorno original. Enquanto o número de vértices no contorno for maior que *T*, selecione o vértice com a menor importância e o remova do contorno; iv) recalcule a importância de cada vértice adjacente aos vértices removidos recentemente; v) repita o passo *ii* até que o número total de vértices no contorno seja igual a *T*.

$$Vi = \arccos\left(\frac{L_i \cdot L_{i+1}}{S_i}\right)S_i \tag{5.1}$$

Figura 16: Exemplo da aplicação do algoritmo PB, no qual são exibidos o vértice Vi e os segmentos de retas L_i e L_{i+1}.



Fonte: O autor (2023).

A Figura 17 exibe alguns exemplos obtidos a partir da aplicação dos algoritmos RDP e PB.

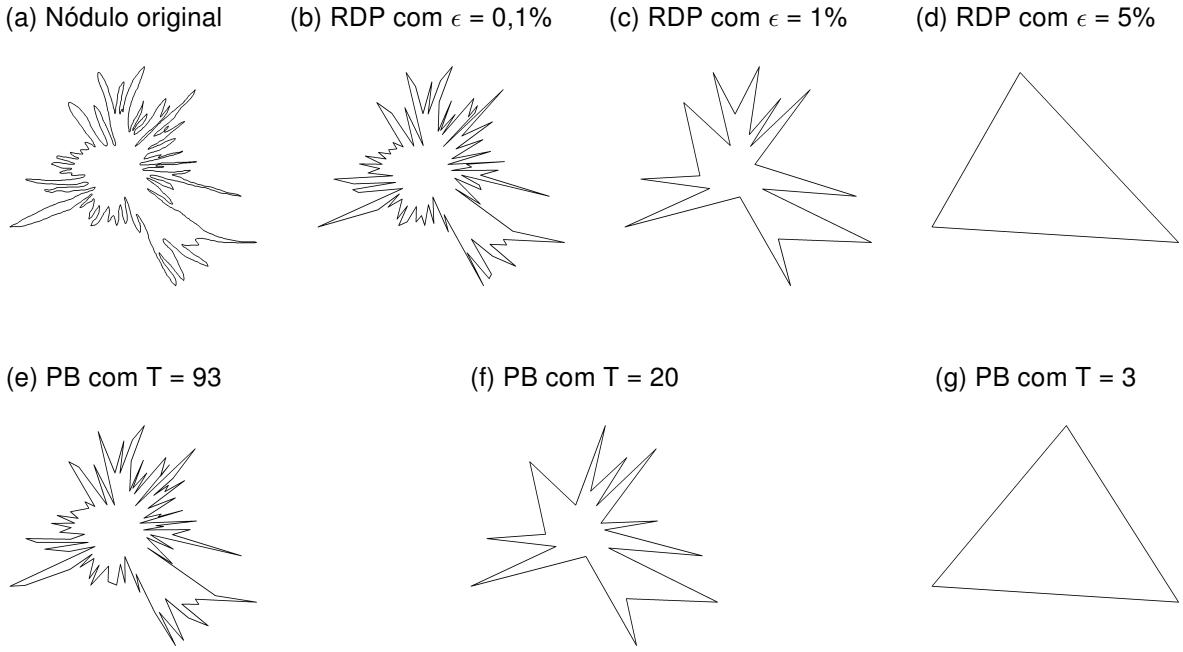
5.2 Extração de características

Em alguns experimentos realizados neste projeto de pesquisa foram utilizadas características extraídas pelos pesquisadores da *University of Calgary* (*dataset ST*). As características relativas à forma do nódulo são: compacidade, índice de espiculação, fração de concavidade, fator de Fourier e dimensões fractais. As características relativas à textura/gradiente dos nódulos são: acutância, coeficiente de variação e contraste.

Após a implementação dos algoritmos para criar os modelos poligonais, foram implementados algoritmos para extrair as mesmas características de forma citadas anteriormente e também os momentos de Hu a partir modelos poligonais criados. Estas características foram extraídas tanto das imagens do *dataset ST* como das imagens do *dataset ACC*.

Durante o restante deste trabalho, será mencionado se as características utilizadas no experimento foram extraídas pelos pesquisadores da *University of Calgary* ou pelos algoritmos implementados durante este projeto. A Seção 5.2.1 detalha a definição das características extraídas.

Figura 17: Representação poligonal dos nódulos. (a) O contorno original do nódulo demarcado por especialistas; (b) - (d) modelos poligonais utilizando o algoritmo RDP com ϵ = 0,1%, 1% e 5%, respectivamente. (e) - (f) modelos poligonais utilizando o algoritmo PB com 93, 20 e 3 pontos, respectivamente, respectively.



Fonte: (HIRAMA et al., 2020).

5.2.1 Características extraídas das imagens

As características extraídas das imagens e que estão sendo utilizadas neste projeto podem ser divididas em três grupos: características de **forma**, de **textura**/**gradiente** e **momentos de Hu**.

5.2.1.1 Características de forma

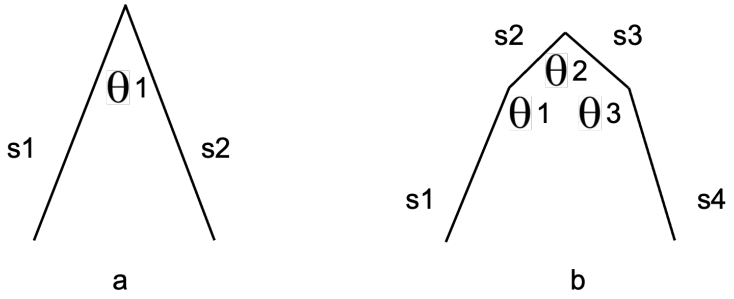
Compacidade: Esta característica mede a eficiência de um contorno em conter uma determinada área, sendo comumente descrita conforme Equação 5.2, na qual *P* e *A* são o perímetro e a área do nódulo, respectivamente. Entretanto, a medida *compacidade* pode também ser representada pela Equação 5.3, sendo esta a equação utilizada neste trabalho. Com esta expressão, *C = 0* para um círculo e seu valor aumenta conforme a forma vai se tornando mais complexa (irregular) até atingir um valor máximo de 1 (RANGAYYAN; MUDIGONDA; DESAUTELS, 2000).

$$C = \frac{P^2}{A} \tag{5.2}$$

$$C = 1 - \frac{4\pi A}{P^2} \tag{5.3}$$

Índice de espiculação: O carcinoma invasivo é um tipo de câncer de mama que promove uma invasão dos tecidos ao seu redor formando distorções estreitas e estelares em sua borda. Dada esta característica do nódulo, foi utilizado o *índice de espiculação* para medir o grau de espiculação da borda de um nódulo (RANGAYYAN; MUDIGONDA; DESAUTELS, 2000). As espículas candidatas são identificadas como pertencendo à borda do nódulo e delimitadas por pares de sucessivas inflexões. A Figura 18 exibe duas espículas candidatas.

Figura 18: Exemplos de espículas e seus ângulos. (a) Espícula formada pelos segmentos s1 e s2 ($M = 2$), ângulo de espiculação $\Theta = \Theta 1$ e S = s1 + s2. (b) Espícula formada pelos segmentos s1, s2, s3 e s4 ($M = 4$). Como $M > 2$, Θ_{th} = $(\Theta 1 + \Theta 2 + \Theta 3)/3$. Como $\Theta 1 > \Theta_{th}$ e $\Theta 3 > \Theta_{th}$ eles podem ser descartados. Como $\Theta 2 < \Theta_{th}$ temos que $\Theta = \Theta 2$. e S = s1 + s2 + s3 + s4.



Fonte: O autor (2023).

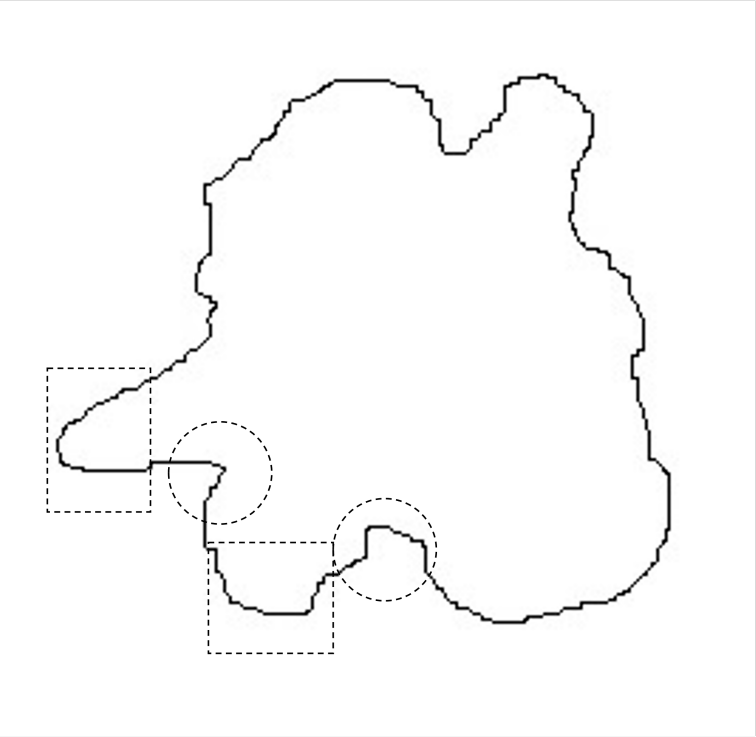
Seja *i* uma espícula de um conjunto de *N* espículas formada por S_i e θ_i, *i = 1, 2, 3, ..., N*, na qual S_i é a soma dos segmentos que formam a espícula e θ_i o ângulo considerado entre os segmentos, o *índice de espiculação* é computado conforme Equação 5.4.

$$SI = \frac{\sum_{i=1}^{N}(1 + \cos\theta_i)S_i}{\sum_{i=1}^{N} S_i} \tag{5.4}$$

Fração de concavidade: A maioria dos nódulos benignos possui maiores quantidades de macrolobulações convexas, embora algumas possam ter poucas concavidades e espículas. Por outro lado, nódulos malignos tipicamente possuem segmentos côncavos e convexos além de microlobulações e espículas proeminentes (RAN-

GAYYAN; MUDIGONDA; DESAUTELS, 2000). A Figura 19 exibe um nódulo maligno no qual foram demarcados alguns segmentos côncavos e convexos.

Figura 19: Partes côncavas e convexas de um nódulo maligno espiculado. Duas partes côncavas destacadas por círculos tracejados e duas partes convexas destacadas por retângulos tracejados.



Fonte: O autor (2023).

O valor da característica *fração de concavidade* deve ser menor para os nódulos benignos do que para os malignos, pois os nódulos benignos possuem menos segmentos côncavos. Detalhes sobre o cálculo desta característica podem ser encontrado em (RANGAYYAN; MUDIGONDA; DESAUTELS, 2000; MUDIGONDA; RANGAYYAN; DESAUTELS, 1999).

Seja um nódulo formado por *M* segmentos e seja $S_i, i = 1, 2, 3, ..., M$ o comprimento de um segmento que compõe o nódulo. O comprimento total T_l da borda do nódulo é computado conforme Equação 5.5. Seja $CC_i, i = 1, 2, 3, ...P$, o comprimento de *P* segmentos côncavos. O comprimento acumulado dos segmentos côncavos é dado pela Equação 5.6. A característica *fração de concavidade* é calculada a partir da Equação 5.7.

$$T_l = \sum_{i=1}^{M} S_i \tag{5.5}$$

$$CC_l = \sum_{i=1}^{P} CC_i \tag{5.6}$$

$$F_{cc} = \frac{CC_l}{T_l} \tag{5.7}$$

Fator de Fourier: É uma medida relacionada à presença de rugosidade ou componentes de alta frequência[1] em um contorno. A vantagem de utilizar o *Fator de Fourier* é que ele é limitado ao intervalo [0, 1], além de não ser sensível a ruído, ser invariante à translação, à rotação, ao ponto inicial, ao tamanho do contorno e seu valor aumenta conforme o contorno do objeto se torna mais complexo e rugoso. É esperado que o valor desta característica seja menor para os nódulos benignos do que para os malignos. Esta medida é obtida por meio da Equação 5.8, na qual $Z_0(k)$ são os descritores de Fourier normalizados (Equação 5.9) e $Z(k)$ são os descritores de Fourier (Equação 5.10, na qual $k = -N/2, ..., -1, 0, 1, 2, ...N/2 - 1$ e $z(n) = x(n) + jy(n), n = 0, 1, ..., N - 1$ representa a sequência das coordenadas de pixels do contorno). Mais informações sobre a extração do *Fator de Fourier* pode ser visto em (RANGAYYAN; NGUYEN, 2007).

$$FF = 1 - \frac{\sum_{K=-N/2+1}^{N/2} |Z_0(k)|/|k|}{\sum_{K=-N/2+1}^{N/2} |Z_0(k)|} \tag{5.8}$$

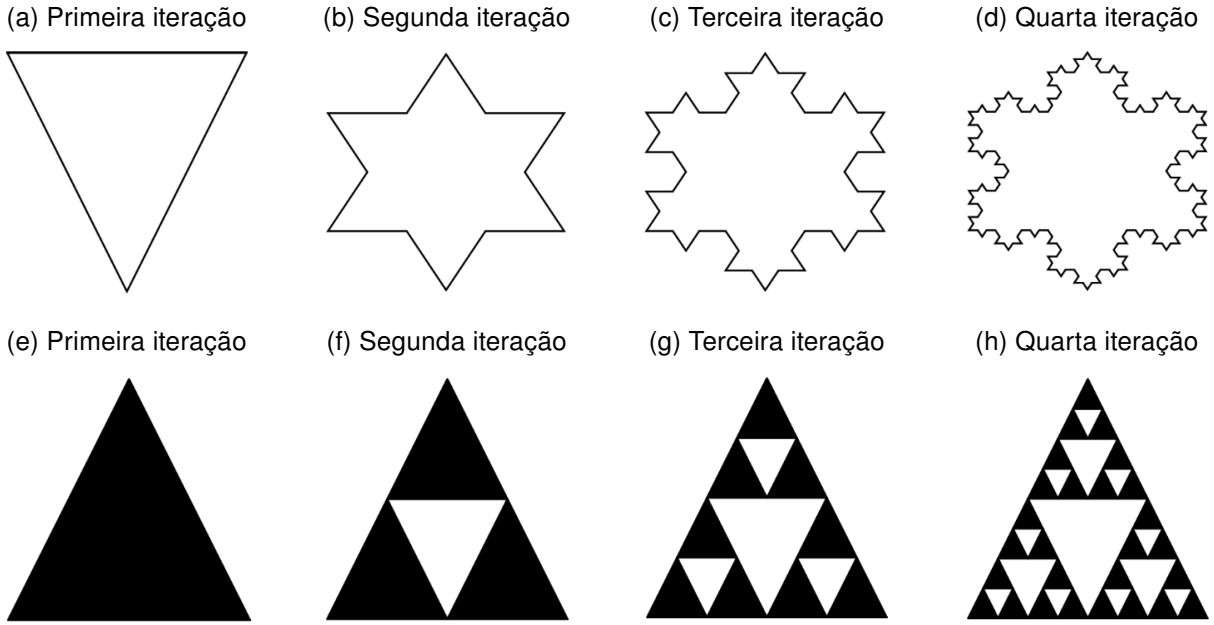
$$Z_0(k) = \begin{cases} 0, & k = 0; \\ \frac{Z(k)}{|Z(1)|}, & \text{caso contrário.} \end{cases} \tag{5.9}$$

$$Z(k) = \frac{1}{N} \sum_{n=0}^{N-1} z(n) exp\left[-j \frac{2\pi}{N} nk \right] \tag{5.10}$$

Dimensão Fractal: Um fractal é uma função ou padrão que possui autossimilaridade em todas ou em várias escalas ou níveis de ampliação. São objetos geométricos irregulares com aninhamento estrutural infinito em todas as escalas. Dois exemplos conhecidos de fractais são a curva de Koch e o triângulo de Sierpinski, que podem ser vistos na Figura 20. Estes exemplos podem ser gerados a partir da repetição de um padrão básico em um processo recursivo ou iterativo. Com o intuito de fazer uma análise quantitativa da forma de um objeto, pode-se utilizar uma medida cha-

[1] No domínio de Fourier, as altas frequências são geradas pelos detalhes das imagens, por exemplo, bordas, lados e transições de nível de cinza.

Figura 20: Exemplos de fractais. (a) - (d) quatro iterações do fractal conhecido como curva de Koch. (e) - (h) quatro iterações do triângulo de Sierpinski.



Fonte: O autor (2023).

mada de dimensão fractal para caracterizar a autossimilaridade, a complexidade de aninhamento ou a propriedade de preenchimento de espaços, podendo ser estendida para caracterizar a complexidade de um padrão em geral. Nódulos malignos possuem um certo grau de aleatoriedade associado com o seu crescimento e têm, tipicamente, forma complexa e irregular, sendo possível utilizar a dimensão de fractal para medir a complexidade de sua forma. Como os nódulos malignos têm uma forma mais rugosa, é esperado que as dimensões fractais destes nódulos sejam maiores que a dos nódulos benignos, pois estes últimos possuem contornos relativamente suaves (RANGAYYAN; NGUYEN, 2007).

Seja *a* o número de peças autossimilares considerando um fator de redução *n* e *D* a dimensão fractal, a relação entre *a*, *n* e *D* é dada pela Equação 5.11. Com base na Equação 5.11, a dimensão fractal *D* é definida conforme Equação 5.12. Portanto, a inclinação da reta aproximada de um gráfico *log(a) vs. log(n)* pode prover uma estimativa de *D* (RANGAYYAN; NGUYEN, 2007).

$$a = n^D \tag{5.11}$$

$$D = \frac{log(a)}{log(n)} \tag{5.12}$$

A forma mais comumente utilizada para calcular a dimensão fractal é o método *box counting*. Este método consiste em particionar o padrão ou o espaço de uma imagem em quadrados de tamanhos iguais, e contar o número de quadrados que contêm uma parte da imagem (pelo menos um pixel). O processo é repetido mediante particionamento do espaço da imagem em pedaços cada vez menores.

Outra forma clássica de calcular a dimensão fractal é conhecida como *ruler method*. Utilizando diferentes comprimentos de segmentos de retas (réguas), o tamanho total de um contorno ou padrão pode ser estimado com diferentes níveis de acurácia. Quando é utilizada uma régua mais larga, os detalhes de um determinado contorno são ignorados; em contrapartida, quando uma régua menor é utilizada, estes detalhes são levados em consideração. Desta forma, a estimativa do tamanho melhora conforme o tamanho da régua diminui. Mais detalhes sobre o cálculo das dimensões fractais utilizando *box counting* e *ruler method* podem ser vistos em (RANGAYYAN; NGUYEN, 2007).

Tanto o *box counting* quanto o *ruler method* podem ser diretamente aplicados a uma assinatura unidimensional (1D) quanto a contornos 2D. Na abordagem utilizada neste projeto, os contornos 2D foram transformados em assinaturas 1D, definidos como a distância radial de cada ponto do contorno ao centroide do nódulo com uma função do índice do ponto do contorno. Como os nódulos benignos são mais arredondados, as assinaturas dos contornos são mais suaves. Em contrapartida, os nódulos malignos são mais rugosos e, portanto, têm uma assinatura mais complexa.

Neste trabalho foram utilizadas quatro medidas de dimensões fractais. São elas: *dimensão fractal 1D ruler*, *dimensão fractal 2D ruler*, *dimensão fractal 1D box counting* e *dimensão fractal 2D box counting*.

5.2.1.2 Características de gradiente e textura

Acutância: É uma medida utilizada para representar nitidez ou a mudança de densidade em uma borda de um nódulo. É computada utilizando derivadas direcionais ao longo da normal em cada um dos pontos da borda do nódulo, levando em consideração as diferenças das intensidades internas e externas normalizada para a distância de um pixel unitário (MUDIGONDA; RANGAYYAN; DESAUTELS, 2000). A acutância pode ser calculada de acordo com a Equação 5.13, na qual f_{max} e f_{min} são os valores de máximo local e mínimo local dos pixels extraídos ao redor da borda, e *N* é o número de pixels ao longo da borda. O *root-mean-squared* (RMS) gradiente d_i no i-ésimo

ponto da borda é obtido por meio da Equação 5.14, na qual $f_i(j), j = 0, 1, 2, ..., n_i$, são os $(n_i + 1)$ números de pixels ao longo da perpendicular no i-ésimo ponto da borda incluindo o próprio ponto da borda. O valor n_i é limitado a no máximo 160 pixels (80 pixels de cada lado da borda).

$$A = \frac{1}{f_{max} - f_{min}} \frac{\sum_{i=1}^{N} d_i}{N} \tag{5.13}$$

$$d_i = \sqrt{\frac{\sum_{j=0}^{n_i-1} [f_i(j) - f_i(j+1)]^2}{n_i}} \tag{5.14}$$

Duas versões de *acutância* foram utilizadas. A primeira, chamada *acutância tradicional* foi obtida por meio do cálculo da diferença entre a intensidade dos pixels ao longo da perpendicular traçada com relação à borda do nódulo. A segunda medida, chamada apenas de *acutância*, foi computada de uma forma similar, mas considerando as diferenças entre os pixels adjacentes ao longo das normais. Mais informações sobre estas características estão descritas em (MUDIGONDA; RANGAYYAN; DESAUTELS, 2000).

Coeficiente de variação: O objetivo desta característica é investigar a variabilidade na nitidez do nódulo em torno de sua borda, além de avaliar a nitidez média com a medida *acutância*. Variância é uma medida estatística da força de um sinal e pode ser utilizada como detector de bordas, pois responde às bordas entre as regiões de diferente cores (MUDIGONDA; RANGAYYAN; DESAUTELS, 2000). A variância σ_ω^2 localizada é calculada em uma janela móvel ω com um número ímpar de pixels *M* na direção perpendicular do pixel da borda conforme Equação 5.15, na qual $M = 5$, $f_i(n), n = 0, 1, 2, ..., n_i$ são os pixels considerados no i-ésimo ponto da borda na direção perpendicular e μ_ω é definido de acordo com a Equação 5.16. O valor máximo de variância computado é escolhido como a 'força' da borda no ponto sendo processado. Em seguida o *coeficiente de variação* é calculado levando em consideração a 'força' da borda para todos os pontos. Mais detalhes sobre esta medida podem ser vistos em (MUDIGONDA; RANGAYYAN; DESAUTELS, 2000).

$$\sigma_\omega^2 = \frac{1}{M} \sum_{n=[-M/2]}^{[M/2]} [f_i(n) - \mu_\omega]^2 \tag{5.15}$$

$$\mu_{\omega} = \frac{1}{M} \sum_{n=[-M/2]}^{[M/2]} f_i(n) \tag{5.16}$$

Contraste: Para a extração da característica *contraste* foi utilizada uma matriz de coocorrência do nível de cinza (*gray level co-occurrence matrix - GLCM*) considerando os pixels da borda do nódulo e da região de seu entorno extraídos anteriormente. A GLCM $P_d(i, j, \theta, d)$ reflete a probabilidade da distribuição da transição do nível de cinza *i* para o nível de cinza *j*, considerando a direção θ e uma distância *d* (neste caso, $d = 1$). A medida contraste utilizada neste projeto é dada pela Equação 5.17, na qual *N* é o número de níveis de cinza (256 nas imagens utilizada neste estudo) e *R* é igual ao número total de pares de pixels na região utilizada para o cálculo da característica em uma determinada direção angular. Mais detalhes sobre a extração desta característica podem ser encontrados em (MUDIGONDA; RANGAYYAN; DESAUTELS, 2000).

$$Contraste = \sum_{n=0}^{N-1} n^2 \sum_{i-j=n} \left(\frac{P(i, j, \theta, d)}{R} \right) \tag{5.17}$$

5.2.1.3 Momentos de Hu

Em (HU, 1962) sete momentos invariantes baseados nos momentos centrais normalizados foram introduzidos. Estes momentos podem ser utilizados como descritores de forma. São muito úteis porque são invariantes à rotação, translação e escala da forma. A definição dos momentos de Hu é explicada com detalhes em (HU, 1962) e foram sumarizadas nas equações 5.18 a 5.22, na qual $I(x, y)$ é a intensidade do pixel na posição (x,y) de uma imagem *I* representada por uma matriz bidimensional, (p + q) são chamados momentos de ordem, $\bar{x}$ e $\bar{y}$ são os componentes do centroide, η_{pq} são os momentos centrais, μ_{pq} são os momentos invariantes e H_1a H_7 são os momentos de Hu.

$$M_{ij} = \sum_{x} \sum_{y} x^i y^i I(x, y) \tag{5.18}$$

$$\bar{x} = \frac{M_{10}}{M_{00}}, \bar{y} = \frac{M_{01}}{M_{00}} \tag{5.19}$$

$$\eta_{pq} = \sum_{x}\sum_{y}(x-\bar{x})^p(y-\bar{y})^q I(x,y) \tag{5.20}$$

$$\mu_{pq} = \frac{\eta_{pq}}{\eta_{00}^{\gamma}}, \gamma = \frac{p+q}{2} \tag{5.21}$$

$$\begin{aligned}
H_1 =& \mu_{20} + \mu_{02} \\
H_2 =& (\mu_{20} - \mu_{02})^2 + 4(\mu_{11})^2 \\
H_3 =& (\mu_{30} - 3\mu_{12})^2 + (\mu_{03} - 3\mu_{21})^2 \\
H_4 =& (\mu_{30} + \mu_{12})^2 + (\mu_{03} + \mu_{21})^2 \\
H_5 =& (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})((\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2) + \\
& (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03})(3(\mu_{30} + \mu_{12})^2 - (\mu_{03} + \mu_{21})^2) \\
H_6 =& (\mu_{20} - \mu_{02})((\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2) + \\
& 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \\
H_7 =& (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})((\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2) + \\
& (\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03})(3(\mu_{30} + \mu_{12})^2 - (\mu_{03} + \mu_{21})^2)
\end{aligned} \tag{5.22}$$

5.3 Seleção de características

Selecionar as características mais úteis é importante para redução de ruído nos dados, contribuindo para melhorar a acurácia dos modelos e classificação.

A “importância de Gini” foi utilizada como mecanismo de seleção de características neste projeto de pesquisa dada a facilidade de obtenção desta medida. Para calcular esta medida, um classificador to tipo Random Forest (RF) foi implementado utilizando as imagens dos *datasets* e considerando as classes “Benigno” e “Maligno”, sendo que a importância de Gini é um dos *outputs* do classificador. Tal valor está relacionado com a frequência em que a característica é selecionada no processo de divisão dos nós de cada árvore de decisão com base no quão discriminativa ela é no processo de classificação (MENZE et al., 2009).

5.4 Discretização das características

O primeiro passo para a criação das gramáticas é a discretização das características contínuas extraídas da imagem. Esse passo é necessário para que a característica discretizada seja representada por um rótulo que, por sua vez, será utilizado para compor uma sequência de símbolos que irá representar um nódulo benigno ou maligno. Neste projeto de pesquisa foram utilizados os algoritmos KBinsDiscretizer (PEDREGOSA et al., 2011) e Ômega (RIBEIRO et al., 2008). O KBinsDiscretizer foi utilizado neste projeto devido ao fato de sua fácil implementação, pois está disponível na biblioteca Scikit Learn, e o Ômega dado a flexibilidade de escolher níveis de inconsistência diferentes para os *bins*.

A implementação do algoritmo KBinsDiscretizer utilizada neste projeto tem três parâmetros de entrada: *n_bins* é a quantidade de *bins* que será gerada, *encode* é o método utilizado para codificar o resultado, ou seja, como será gerado o identificador de cada *bin*, e *strategy* é a estratégia utilizada para definir a largura de cada *bin*. Para o parâmetro *encode* foi utilizado o valor “ordinal” que gera o identificador no formato de um inteiro, o valor atribuído ao parâmetro *strategy* foi “uniform” para que os *bins* tivessem todos a mesma largura e foram testados vários valores para o parâmetro *n_bins* para a calibração do processo de discretização. Este algoritmo não leva em consideração a classe de cada instância, pois ele divide as características em *bins* considerando apenas o valor de cada característica.

O algoritmo Ômega utiliza uma medida de inconsistência (proporção de instâncias presentes em um *bin* que não pertence à classe majoritária) para determinar o número final de *bins*. Conforme o número de *bins* diminui, o número de inconsistências aumenta. Desta forma, o objetivo é manter um número de *bins* mínimo com um número mínimo de inconsistências, estabelecendo um balanceamento entre essas duas medidas. Ômega tem um custo de processamento linear considerando N valores ordenados e pode ser dividido em quatro passos. Os primeiros três passos são utilizados para a discretização das características e serão explicados a seguir, enquanto que o quarto passo é relacionado ao processo de seleção de características e não foi utilizado neste projeto (RIBEIRO et al., 2008). Este algoritmo leva em consideração a classe ao qual cada instância pertence, pois utiliza esta informação no cálculo do número mínimo de instâncias em cada *bin* e na medida de inconsistência. Neste projeto de pesquisa, utilizamos as classes BC/BnE, BE, MC/MnE e ME na discretização das características de forma e dos momentos de Hu. Para a discretização

das características de textura/gradiente foram consideradas apenas as classes B e M, pois a informação se a borda do nódulo é circunscrita ou espiculada se relaciona com características de forma e não de textura.

Seja f uma determinada característica e f_i o valor da característica f em uma instância i. Ômega faz uso de uma estrutura que liga cada valor da instância f_i com a classe c_i (da i-ésima instância). Considere uma instância I_i sendo um par (f_i, c_i) e considere também $U_{inf\{k\}}$ e $U_{sup\{k\}}$ como limites inferior e superior de um intervalo T_k. Temos que uma instância $I_i = (f_i, c_i)$ pertence a um intervalo $T_k = [U_{inf\{k\}}, U_{sup\{k\}}]$ se e somente se $U_{inf\{k\}} \leq f_i \leq U_{sup\{k\}}$. O exemplo utilizado na explicação do algoritmo Ômega a seguir foi adaptado de (RIBEIRO et al., 2008).

Passo 1. Inicialmente o algoritmo ordena as instâncias de acordo com os valores contínuos das características e define dois pontos de corte iniciais. O primeiro ponto de corte é colocado antes da instância que possui o menor valor da característica e o segundo ponto de corte é colocado logo após a instância que possui o maior valor da característica. Percorrendo as instâncias esquerda para a direita, sempre que a classe e o valor da característica da instância atual são diferentes da classe e do valor da característica da instância imediatamente anterior, um novo ponto de corte é criado.

Neste primeiro passo, cada intervalo criado possui a menor inconsistência possível (zero), pois todas as características pertencentes a um determinado intervalo possuem a mesma classe. Entretanto, o número de intervalos tende a ser muito grande, e um processo de discretização que produz um grande número de intervalos não é desejado, pois não há um ganho na aplicação do algoritmo (RIBEIRO et al., 2008). A Figura 21 exibe um exemplo dos valores da característica ordenados em ordem crescente e os pontos de corte encontrados (linha entre os valores contínuos) após o passo 1.

Passo 2. Neste passo, os pontos de corte podem ser alterados para que cada intervalo criado possua uma quantidade de instâncias maior ou igual ao número mínimo determinado pelo parâmetro H_{min}. Desta forma, o algoritmo remove os pontos de corte à direita dos intervalos que não satisfazem a restrição de frequência mínima fornecida como parâmetro de entrada H_{min}, sendo que apenas ao último intervalo é permitido não satisfazer esta restrição.

Quanto maior o valor de H_{min}, menor será o número de intervalos obtidos neste passo. Entretanto, quanto maior o H_{min}, maior será o número de inconsistências geradas pelo processo de discretização. Por este motivo, é importante manter esse

Figura 21: Pontos de corte criados após o passo 1 do algoritmo Ômega.

Valores ordenados da característica *f*

0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08
B	M	M	B	B	M	B	B

Rótulo da classe

Pontos de corte gerado no passo 1

0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08
B	M	M	B	B	M	B	B

Fonte: O autor (2023).

valor baixo, mesmo que apenas uma pequena redução de intervalos seja obtida. A Figura 22 exibe um exemplo dos pontos de corte obtidos no passo 1 e que são eliminados no passo 2 quando $H_{min} = 2$.

Figura 22: Pontos de corte eliminados após o passo 2 do algoritmo Ômega utilizando $H_{min} = 2$.

Eliminação de pontos de corte no passo 2

0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08
B	M	M	B	B	M	B	B

Fonte: O autor (2023).

Passo 3. Neste passo, o algoritmo funde intervalos consecutivos, limitando a taxa de inconsistência durante a fusão. Seja M_{T_k} a classe majoritária de um intervalo T_k. A taxa de inconsistência ζ_{T_k} de um intervalo T_k é dada pela Equação 5.23, na qual $|T_k|$ é o número de instâncias de um intervalo T_k e $|M_{T_k}|$ é o número de instâncias da classe majoritária no intervalo T_k. Ou seja, ela mede a proporção de instâncias desse intervalo que não pertencem à classe majoritária. O algoritmo Ômega funde intervalos consecutivos que possuam a mesma classe majoritária que tenham uma taxa de inconsistência abaixo ou igual a um limite fornecido como entrada ζ_{max} $(0 \leqslant \zeta_{max} \leqslant 1)$.

$$\zeta_{T_k} = \frac{|T_k| - |M_{T_k}|}{|T_k|} \tag{5.23}$$

A Figura 23 disponibiliza um exemplo de um ponto de corte encontrado no passo 2 que é eliminado no passo 3 quando é utilizado $\zeta_{max} = 0,35$. A taxa de inconsistência ζ_{T_k} do segundo e do terceiro intervalos exibidos na Figura 23 são $\zeta_{T_2} = 0/2$ e $\zeta_{T_3} = 1/3$.

Como T_2 e T_3 possuem a mesma classe majoritária **A** e $\zeta_{T_2} \leqslant \zeta_{max}$ e $\zeta_{T_3} \leqslant \zeta_{max}$, o segundo e o terceiro intervalos são fundidos.

Os pontos de corte que restarem após a aplicação do passo 3 do algoritmo são os pontos de corte finais.

Figura 23: Ponto de corte eliminado no passo 3 do algoritmo Ômega considerando $\zeta_{max} = 0,35$.

Eliminação de pontos de corte no passo 3

0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08
B	M	M	B	B	M	B	B

Fonte: O autor (2023).

A Tabela 6 exibe um exemplo de discretização dessas características considerando as classes Benigno (B), Maligno (M), Benigno Circunscrito (BC), Benigno Espiculado (BE), Maligno Circunscrito (MC) e Maligno Espiculado (ME). É importante ressaltar que os nódulos com classes BnE ou MnE do *dataset* ACC foram combinados com os nódulos com classes BC ou MC do *dataset* ST, respectivamente, para a realização dos experimentos. Para as características de gradiente/textura, foram consideradas apenas as classes Benigno e Maligno no processo de discretização, pois nos modelos hierárquicos propostos não são utilizados as informações de borda (circunscrita ou espiculada) para características desta categoria. Na Tabela 6 pode-se perceber que um mesmo valor de rótulo para uma característica pode aparecer em mais de uma classe, por exemplo, o valor *ccb* aparece nas classes BC, BE e MC, pois existem nódulos nestas classes que possuem valores de compacidade que foram atribuídos ao *bin b* durante o processo de discretização.

5.5 Classificação dos nódulos utilizando gramáticas

5.5.1 Grafo AND-OR

O modelo hierárquico utilizado neste projeto é o grafo AND-OR (Seção 2.5), pois com este modelo é possível não apenas representar as características do nódulo de forma hierárquica, mas também construir gramáticas livres de contexto a partir de sua estrutura.

A Figura 24 exibe a representação de um nódulo benigno e a Figura 25 mostra a representação de um nódulo maligno, neste caso diferindo apenas nos nós folhas. Nes-

Tabela 6: Exemplo de discretização das características. Nos valores das características discretizadas, as duas primeiras letras representam a característica (cc = compacidade, si = índice de espiculação, co = contraste e ac = acutância) e a terceira letra representa o *bin* ao qual o valor da característica pertence.

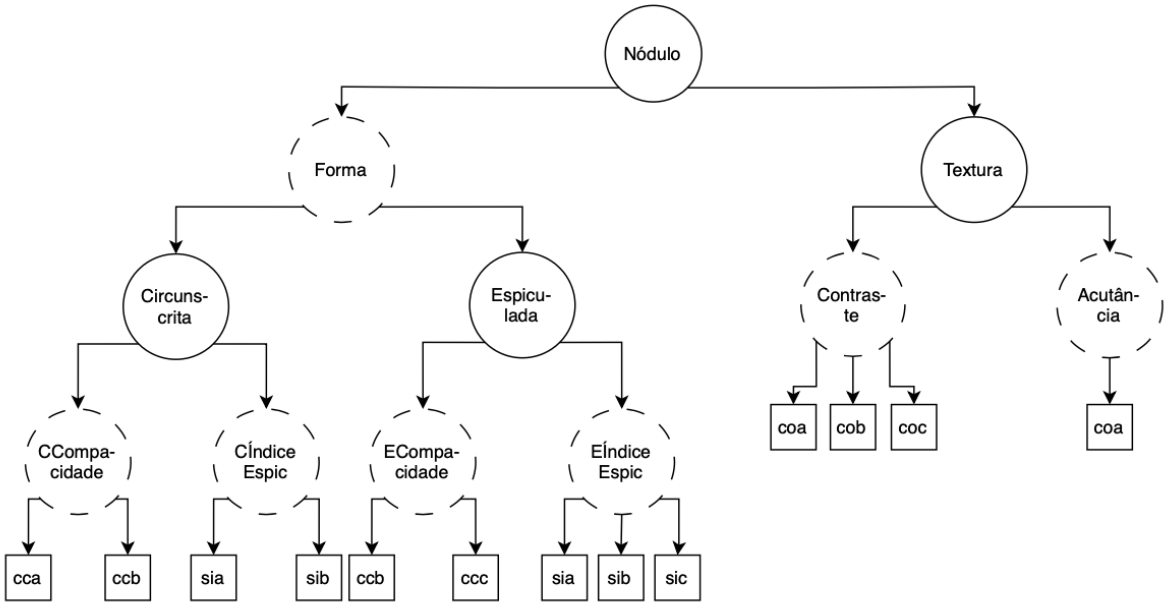
Característica	Classe	Valores
Compacidade	BC	cca, ccb
Compacidade	BE	ccb, ccc
Compacidade	MC	ccb, ccc, ccd
Compacidade	ME	ccb, ccc, ccd
Índice de espiculação	BC	sia, sib
Índice de espiculação	BE	sia, sib, sic
Índice de espiculação	MC	sid, sie, sif
Índice de espiculação	ME	sif
Contraste	B	coa, cob, coc
Acutância	B	aca
Contraste	M	coc, cod
Acutância	M	aca, acb

Fonte: O autor (2023)

sas figuras foram consideradas duas características de forma (*compacidade* e *índice de espiculação*) e duas características de gradiente/textura (*contraste* e *acutância*). A construção do grafo AND-OR é feita após o processo de seleção e de discretização das características. Nos exemplos exibidos nas Figuras 24 e 25 os símbolos apresentados nos nós folhas são os valores das características discretizadas e exibidas na Tabela 6.

A partir dos grafos AND-OR pode-se criar as gramáticas livres de contexto que foram utilizadas para descrever os nódulos benignos e malignos, nas quais os nós *AND* e *OR* do grafo serão os símbolos não terminais da gramática e os nós folhas serão os símbolos terminais. Os nós *AND* e seus filhos podem ser vistos como regras gramaticais da forma $A \rightarrow B.C.D$ (sendo A o nó *AND* e B, C, D seus filhos) e os nós *OR* e seus filhos podem ser vistos como regras gramaticais da forma $A \rightarrow B|C|D$ (sendo A o nó *OR* e B, C, D seus filhos), sendo que os símbolos "." e "|" representam as condições lógicas *E* e *OU*, respectivamente. Desta forma, para a gramática benigna tem-se que $S =$ {Nódulo}, $V_N =$ {Forma, Textura, Circunscrita, Espiculada, CCompacidade, CÍndice Espic, ECompacidade, EÍndice Espic, Contraste, Acutância} e $V_T =$ {cca, ccb, ccc, sia, sib, sic, coa, cob, coc, aca}. Para a gramática maligna, S e V_N são os mesmos da gramática benigna e $V_T =$ {ccb, ccc, ccd, sib, sid, sie, sif, coc, cod, aca, acb}.

Figura 24: Representação de um nódulo benigno utilizando um grafo AND-OR e sua gramática correspondente. Os círculos com as bordas cheias representam os nós *AND*; os círculos com as bordas tracejadas representam os nós *OR*; os nós folhas são representados por quadrados.



Gramática para descrever os nódulos benignos

O símbolo | representa a condição lógica OU e o símbolo . representa a condição lógica E.

Nódulo ⟶ Forma . Textura

Forma ⟶ Circunscrita | Espiculada

Circunscrito ⟶ CCompacidade . CÍndice Espic

Espiculado ⟶ ECompacidade . EÍndice Espic

Textura ⟶ Contraste . Acutância

CCompacidade ⟶ cca | ccb

CÍndice Espic ⟶ sia | sib

ECompacidade ⟶ ccb | ccc

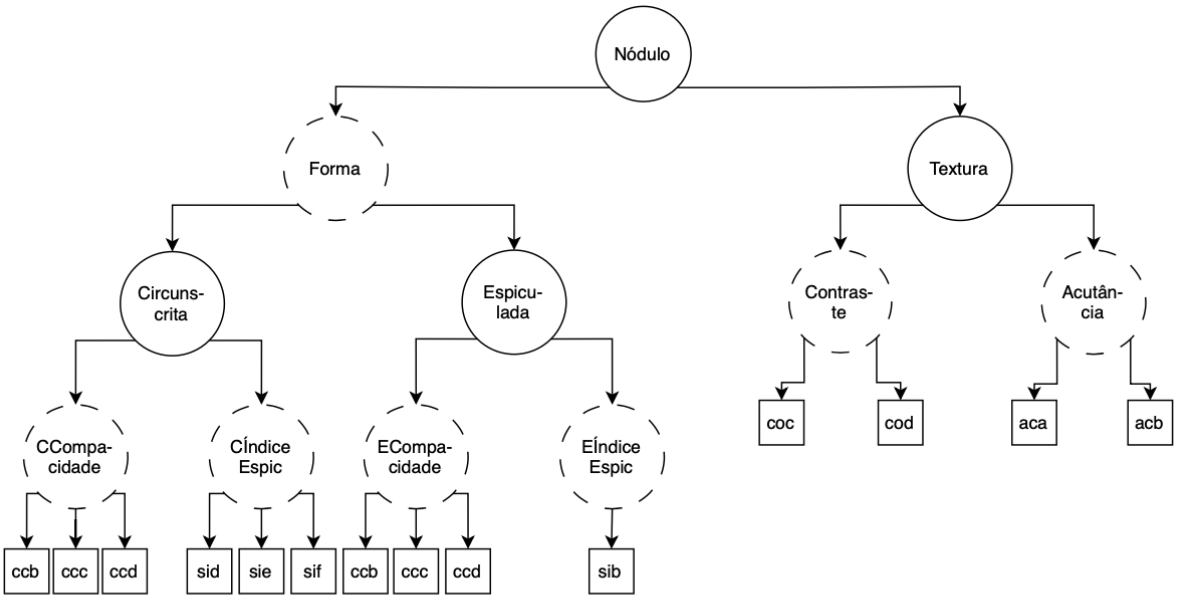
EÍndice Espic ⟶ sia | sib | sic

Contraste ⟶ coa | cob | coc

Acutância ⟶ aca

Fonte: O autor (2023)

Figura 25: Representação de um nódulo maligno utilizando um grafo AND-OR e sua gramática correspondente. Os círculos com as bordas cheias representam os nós *AND*; os círculos com as bordas tracejadas representam os nós *OR*; os nós folhas são representados por quadrados.



Gramática para descrever os nódulos malignos

O símbolo | representa a condição lógica OU e o símbolo . representa a condição lógica E.

Nódulo ⟶ Forma . Textura

Forma ⟶ Circunscrita | Espiculada

Circunscrito ⟶ CCompacidade . CÍndice Espic

Espiculado ⟶ ECompacidade . EÍndice Espic

Textura ⟶ Contraste . Acutância

CCompacidade ⟶ ccb | ccc | ccd

CÍndice Espic ⟶ sid | sid | sif

ECompacidade ⟶ ccb | ccc | ccd

EÍndice Espic ⟶ sib

Contraste ⟶ coc | cod

Acutância ⟶ aca | acb

Fonte: O autor (2023).

5.5.2 Estimação das probabilidades

Para fazer a estimação das probabilidades da gramática inicialmente não estocástica foi utilizado o algoritmo de estimação de passo único proposto por (FU, 1982),

baseado em estimação por máxima verossimilhança. Para que este algoritmo possa ser utilizado é necessário assumir que o conjunto de sequências (cadeias) utilizada no processo de inferência possui a mesma distribuição de probabilidade da linguagem estocástica desconhecida. Este algoritmo foi escolhido por possuir uma base teórica mais simples de ser implementado, pois faz a contagem da utilização das produções gramaticais durante a análise das cadeias de treinamento, considerando também as frequências dessas cadeias na amostra. Também é importante mencionar que o algoritmo não requer que informação estrutural acerca das árvores de derivação esteja presente nas cadeias a serem analisadas.

O primeiro passo desse algoritmo é atribuir um contador, inicialmente igual a zero, para cada produção da gramática. Em seguida, é realizada a análise sintática de cada sequência do conjunto de treinamento considerando a gramática em questão. Sempre que uma regra gramatical (produção) é utilizada, seu contador é incrementado em uma unidade. Após todas as sequências da amostra terem sido analisadas, os contadores são normalizados para que cada produção assuma um valor de probabilidade entre 0 (zero) e 1 (um) e que a somatória das produções que possuam o mesmo lado esquerdo seja igual a 1 (um).

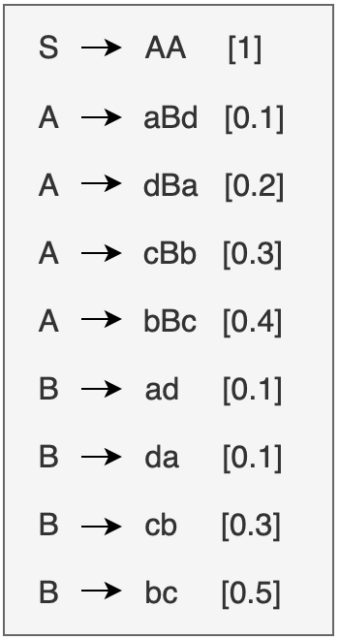
O processo de transformação de uma gramática *G* (Definição 2 da Seção 2.4) em uma gramática estocástica G_s (Definição 11 da Seção 2.4) é feita da seguinte forma: para uma dada amostra de treinamento $S = x_1, x_2..., x_n$ e uma gramática *G* com produções do tipo $A_i \rightarrow \alpha_j$, na qual $A_i \in V_N$ e $\alpha_j \in (V_N \cup V_T)^*$, a probabilidade $\hat{p}_{ij}$ atribuída a cada produção gramatical é calculada pelas Equações 5.24 e 5.25, na qual $n_{ij}(x_k)$ é a quantidade de vezes que a produção $A_i \rightarrow \alpha_j$ foi utilizada durante a análise da sequência x_k.

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_j n_{ij}} \tag{5.24}$$

$$n_{ij} = \sum_{x_k \in S} n_{ij}(x_k) \tag{5.25}$$

Se assumirmos que a amostra de treinamento possui a mesma distribuição da linguagem desconhecida, a probabilidade estimada $\hat{p}_{ij}$ aproxima p_{ij} à medida que o número de sequências *n* que fazem parte da amostra de treinamento tende ao infinito. A Figura 26 ilustra um conjunto de regras de produção de uma gramática estocástica.

Figura 26: Regras de produção de uma gramática estocástica.



Fonte: O autor (2023)

5.5.3 Analisador sintático

Para uma determinada gramática e uma sequência, esta sequência pode ou não ser reconhecida pela gramática. Caso seja reconhecida, haverá uma ou mais árvores sintáticas capazes de descrever suas derivações. No caso das gramáticas estocásticas, ao analisar uma sequência o analisador sintático deve fornecer a(s) árvore(s) de derivação e a probabilidade dessa sequência segundo a gramática. A probabilidade de uma sequência é a soma das probabilidades de suas árvores de derivação (caso exista mais de uma), e a probabilidade de cada árvore sintática é o produto das probabilidades das regras de produções utilizadas para compor a árvore sintática.

O algoritmo de Earley (EARLEY, 1970) é um analisador sintático para gramáticas livres de contexto capaz de prover todas as árvores sintáticas de uma determinada sequência. Para fazer as análises sintáticas neste projeto foi utilizado o *Probabilistic Earley parser*[2], que é uma biblioteca para análise sintática de uma sequência de *tokens*, dada uma gramática estocástica. Entretanto, no caso específico da biblioteca utilizada, ela fornece apenas a árvore sintática com a maior probabilidade, sendo que esta árvore sintática foi utilizada para determinar a probabilidade de uma cadeia x e sua árvore mais provável t_{max} dada uma gramática G, $P(x, t_{max}|G)$.

[2]https://github.com/digitalheir/java-probabilistic-earley-parser

5.5.4 Classificador Bayesiano

O classificador das imagens que foi utilizado neste projeto é um classificador Bayesiano, conforme definido na Seção 2.6.

No presente projeto, considerando duas classes de imagens *c1* e *c2* e que uma imagem x pode pertencer à classe *c1* ou *c2* com alguma probabilidade, temos que *c1* e *c2* podem ser consideradas como valores de uma variável aleatória *c*. A cada classe *ci* está associada uma gramática estocástica G_i, e seu analisador sintático *i* gerado fornece o valor de $P(x, t_{max}|G_i)$. O classificador Bayesiano foi escolhido para ser utilizado neste projeto pois ele é capaz de calcular a *posteriori* $P(G_i|x, t_{max})$, e irá classificar a imagem *x* para a classe *ck* que apresentar maior *posteriori* (Equação 5.26):

$$k = \arg\max_i \ P(G_i|x, t_{max}) \tag{5.26}$$

5.6 Validação e testes

Neste projeto foi utilizada a técnica chamada de validação cruzada estratificada com *k-folds* que consiste em particionar o conjunto de dados em *k* subconjuntos disjuntos de mesmo tamanho mantendo a mesma proporção das classes em cada um dos subconjuntos. Em seguida, *k-1* subconjuntos são utilizados para o treinamento do modelo, o qual é testado no k-ésimo subconjunto restante, sobre o qual são calculadas as medidas de desempenho sendo consideradas (acurácia, por exemplo). Repete-se esse procedimento *k* vezes alternando, de forma circular, o subconjunto utilizado para testes. Ao término de todas as iterações é calculada a média dos k valores de cada medida de desempenho, obtendo dessa forma uma estimativa mais confiável sobre o modelo utilizado.

A validação cruzada foi realizada considerando apenas o *dataset* ST, apenas o *dataset* ACC e quando as imagens desses dois *datasets* foram juntadas compondo um único conjunto de dados. Além da validação cruzada, também foram realizados experimentos nos quais o treinamento do modelo se deu utilizando imagens de um *dataset* específico e os testes ocorreram com imagens de outro *dataset*.

Conforme constatado durante a revisão bibliográfica (Capítulo 3), as principais medidas utilizadas para avaliar o desempenho de um classificador são: AUC, acurácia,

sensibilidade e especificidade. Estas mesmas medidas foram utilizadas nesse trabalho para medir o desempenho do classificador. A definição de cada uma destas medidas podem ser encontradas na Tabela 3.

5.7 Considerações finais

Neste capítulo foram apresentados os passos executados para definição, uso e implementação de gramáticas para classificação de nódulos em diferentes classes, incluindo desde os passos iniciais de extração de características até os algoritmos empregados para definição de gramáticas estocásticas.

No Capítulo 7 são apresentados os resultados experimentais obtidos a partir do uso das implementações aqui apresentadas.

6 GERAÇÃO DE IMAGENS SINTÉTICAS DE NÓDULOS

Como apresentado no Capítulo 4, a segunda parte do presente projeto consiste em definir uma abordagem para gerar imagens sintéticas com nódulos, visando a prover material para testes de sistemas CAD, o qual também poderia ser utilizado como material didático da área de radiologia.

Três possíveis abordagens foram utilizadas para geração dos nódulos benignos e malignos. A primeira abordagem consiste em fundir contornos de nódulos reais para criar um novo nódulo. Na segunda abordagem são realizadas pequenas alterações em um contorno real existente para gerar um novo contorno sintético. A última abordagem consiste em transpor um nódulo real existente para um mamograma diferente do mamograma original. Essas abordagens estão descritas nas próximas seções.

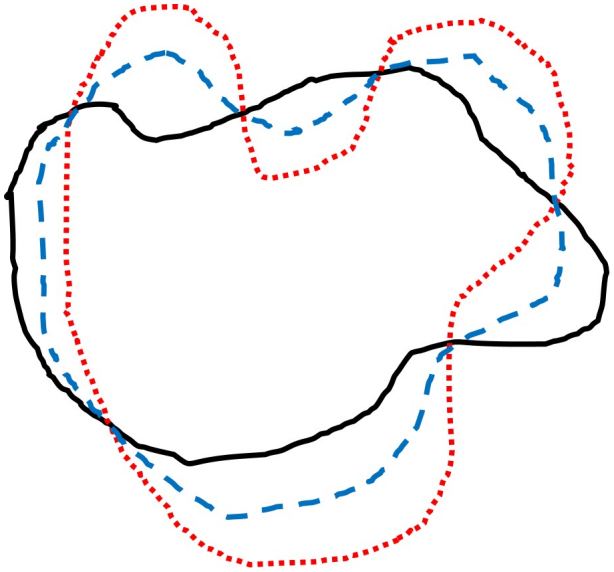
6.1 Fusão de contornos reais

Para a implementação desta abordagem foi escolhido, de forma arbitrária, um contorno de um nódulo real existente. Em seguida, combinou-se o contorno deste nódulo com o contorno de todos os outros nódulos de mesma classe (Benigno ou Maligno) e de forma independente (sempre em pares) para gerar novos contornos. Por exemplo, considere que a base de imagens seja formada por três nódulos benignos (*Na*, *Nb* e *Nc*) e por três nódulos malignos (*Nd*, *Ne* e *Nf*). Considere ainda que o contorno do nódulo *Na* foi escolhido de forma arbitrária para ser combinado com o contorno dos demais nódulos benignos. Desta forma, ao final do processo, as combinações dos contornos do nódulo *Na* com os demais nódulos terá gerado os novos contornos supostamente benignos *NaNb* e *NaNc*. De forma semelhante, supondo que o contorno do nódulo maligno *Nd* seja o contorno escolhido aleatóriamente, ao final do processo também terão sido gerados os novos contornos supostamente malignos *NdNe* e *NdNf*.

A Figura 27 mostra um exemplo da fusão de dois contornos de nódulos para gerar

um terceiro nódulo considerando apenas o seu contorno. Neste exemplo, o nódulo com o contorno preto foi escolhido de forma arbitrária e o seu contorno será mesclado ao contorno do nódulo com contorno vermelho (linha pontilhada). Os dois nódulos foram centralizados de acordo com seu centro de massa e o contorno de um novo nódulo (azul - linha tracejada) foi criado baseando-se nos nódulos selecionados. Neste exemplo, o novo contorno é constituído por pontos equidistantes dos contornos escolhidos previamente.

Figura 27: Representação da construção do contorno de um novo nódulo mediante junção de contornos. Contorno azul (linha tracejada) representa a borda de um nódulo gerado a partir de outros dois contornos.



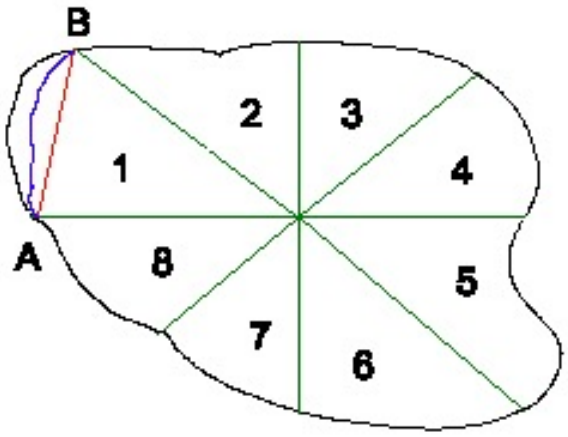
Fonte: O autor (2023).

Após a combinação de todos os contornos existentes na base de dados com o contorno do nódulo selecionado inicialmente, é utilizado o classificador Bayesiano criado anteriormente para classificar os contornos gerados (Seção 5.5). Este classificador irá classificar os novos contornos criados como pertencentes às classes Benigno ou Maligno. Quando o classificador não é capaz de reconhecer um novo contorno (probabilidade igual a zero para as duas classes), este contorno é descartado. Desta forma, ao final do processo obtém-se novos contornos rotulados como pertencentes à classes Benigno ou Maligno.

6.2 Alteração de contornos reais

Esta abordagem consiste em alterar uma pequena parte do contorno de um nódulo existente gerando um novo contorno diferente do contorno do nódulo original. Inicialmente o contorno do nódulo original é fatiado em oito partes de forma radial, semelhantemente a uma pizza fatiada em oito partes, conforme apresentado na Figura 28. Escolheu-se dividir a imagem em octantes para que a parte do contorno alterada fosse relativamente pequena quando comparada com a contorno completo, mas que ainda assim não fosse desprezível. Em seguida, a alteração é realizada na parte do contorno de apenas uma das oito porções que formam o contorno. A escolha de qual porção que será alterada pode ser feita de forma aleatória ou fixa, por exemplo, escolher sempre a primeira porção. Em seguida, é traçado um segmento de reta unindo os pontos *A* e *B* que são os pontos onde se encontram a intersecção do contorno com os segmentos de reta que fatiam o nódulo. Na sequência, é traçada uma nova porção do contorno entre os pontos *A* e *B*. Esta nova porção é formada por pontos equidistantes do contorno original e do segmento de reta, gerando o contorno final. A alteração é exemplificada na Figura 28, na qual o segmento de reta entre os pontos *A* e *B* é apresentado na cor vermelha e o novo contorno é apresentado na cor azul.

Figura 28: Construção do contorno de um novo nódulo mediante alteração do contorno original. Contorno em preto é o contorno original do nódulo. Em verde estão representados os segmentos fatiando o contorno original em oito partes. O segmento de reta (na cor vermelha) entre os pontos *A* e *B* é o guia para a alteração do contorno original, representado em azul e formado por pontos equidistantes do contorno original e do segmento de reta $\overline{AB}$).



Fonte: O autor (2023).

6.3 Transferência de um nódulo real para um novo mamograma

Esta abordagem teve como objetivo verificar a viabilidade de alterar a posição de um nódulo presente em um mamograma. Mais precisamente, foi testada a possibilidade de transferir um nódulo encontrado em um mamograma para um outro mamograma no qual não existisse nenhum nódulo. Esta verificação foi importante para validar o quanto as estruturas internas presentes nos nódulos impactariam ou influenciariam na escolha do novo local em uma outra imagem.

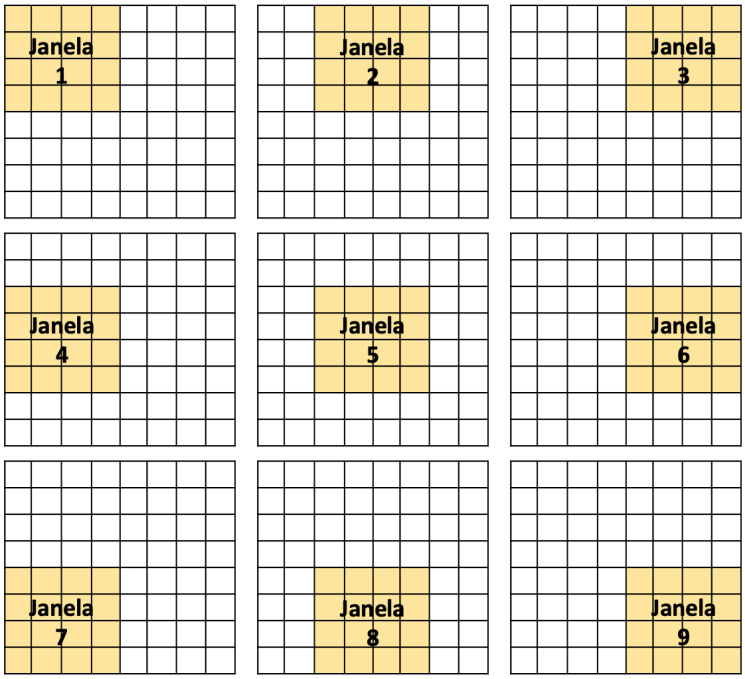
Para isto, esta abordagem procurou encontrar o local ideal para um novo nódulo com base na diferença entre a média de nível de cinza do local escolhido e a média de nível de cinza do nódulo. Inicialmente é calculado o nível de cinza médio do nódulo. Em seguida, utiliza-se uma janela deslizante do tamanho do nódulo e que percorre todo o mamograma calculando o nível de cinza médio de cada região. A janela desliza pelo mamograma inicialmente da esquerda para a direita dando um passo sempre igual ao tamanho da metade do comprimento do nódulo e de cima para baixo com um passo sempre igual ao tamanho da metade da altura do nódulo. Quando o nível médio de cinza de todas as regiões do mamograma tiverem sido calculados, é escolhida a região que apresenta o nível de cinza médio mais próximo ao nível de cinza médio do nódulo. A Figura 29 exibe uma janela deslizante de tamanho 4x4 percorrendo uma imagem de tamanho 8x8 pixels.

Para suavizar os valores dos níveis de cinza dos pixels pertencentes às bordas dos nódulos transpostos, foram realizados testes atribuindo a esses pixels valores considerando o valor médio da vizinhança dos pixels utilizando máscaras do tamanho 3x3, 5x5 e 7x7. A Figura 30 exibe uma máscara de tamanho 3x3 ao redor de um ponto de interesse simulando uma parte do contorno de um nódulo em vermelho.

6.4 Aplicação de textura para os nódulos sintéticos

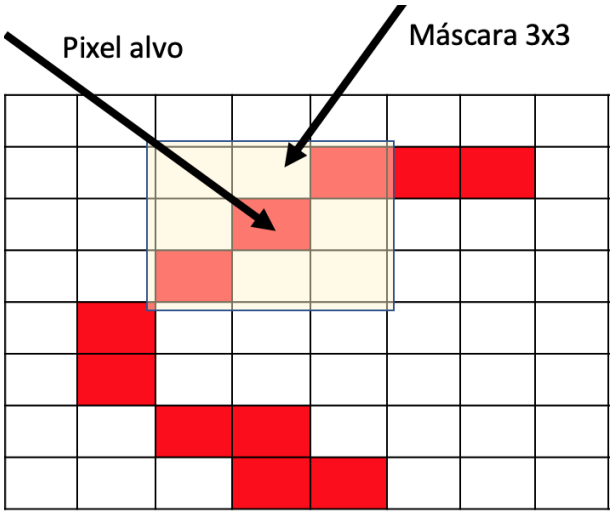
Esta seção descreve os experimentos que foram realizados para a criação da textura para o novo nódulo gerado mediante uso de técnicas de processamento de imagens. A aplicação de textura para os nódulos sintéticos é importante pois um nódulo sintético precisa apresentar uma textura semelhante a outros nódulos para que eles não pareçam artificiais.

Figura 29: Exemplo de janela deslizante 4x4 na cor amarela percorrendo uma imagem representada em uma matriz de tamanho 8x8 pixel.



Fonte: O autor (2023).

Figura 30: Máscara de tamanho 3x3 na cor amarela clara ao redor de um píxel de interesse que simula a borda de um nódulo na cor vermelha.



Fonte: O autor (2023).

Utilização da média dos valores dos pixels de nódulos conhecidos. Esta abordagem consiste em computar a média dos valores dos pixels (nível de cinza) internos a um nódulo (*media_pixels_interno*) e computar a média dos valores dos pixels na região de interesse (ROI) mais próxima externa ao mesmo nódulo (*media_pixels_externo*).

Em seguida, calcula-se a diferença entre as duas médias encontradas e considera-se o seu valor absoluto (*diferenca_pixels*).

De posse de um novo contorno para o nódulo gerado anteriormente e uma nova região de interesse na qual o nódulo será inserido, a nova textura gerada é dada pela Equação 6.1, na qual *pixel_original* é o valor original do pixel na nova região de interesse e *ruido* é um inteiro gerado de forma aleatória em um intervalo de -10 a +10.

$$novo\_valor\_pixel = pixel\_original + diferenca\_pixels + ruido \tag{6.1}$$

É atribuído o valor de 255 ou 0 ao *novo_valor_pixel* caso o valor calculado seja superior a 255 ou inferior a 0, respectivamente. Esta abordagem pode ser utilizada considerando-se apenas um nódulo real existente ou um conjunto de nódulos reais para os cálculos de *media_pixels_interno* e *media_pixels_externo*.

Utilização de regressão linear para inferência do valor do pixel do nódulo. Essa abordagem fez uso de duas características para criação do modelo de regressão linear: i) a distância de um pixel ao centro de massa de um novo nódulo e ii) a menor distância do pixel à borda do nódulo, sendo o valor do pixel utilizado como variável resposta durante o treinamento do modelo. Para a criação do modelo, tanto as características quanto a variável resposta utilizadas no processamento passaram por uma etapa de pré-processamento de tal forma que seus valores foram normalizados conforme pode ser visto na Equação 6.2, na qual z é o valor obtido a partir da normalização, x é o valor da característica, u é o valor médio da característica e s é o desvio padrão calculado. O objetivo dessa abordagem foi validar se apenas com essas duas características seria possível criar a textura para um novo nódulo.

$$z = \frac{(x - u)}{s} \tag{6.2}$$

6.5 Considerações

Neste capítulo foram apresentadas as abordagens desenvolvidas visando a gerar imagens sintéticas de nódulos utilizando as gramáticas previamente definidas no Capítulo 5.

No Capítulo 8 são apresentados os resultados experimentais obtidos a partir da aplicação das abordagens aqui apresentadas.

PARTE III

RESULTADOS E DISCUSSÕES

7 CLASSIFICAÇÃO DE NÓDULOS

Este capítulo apresenta os experimentos e os resultados obtidos no que diz respeito à classificação dos nódulos encontrados nos mamogramas. Uma visão geral dos cenários analisados pode ser visto a seguir:

- **Classificação dos nódulos com características de Calgary:** exibe os resultados obtidos utilizando as imagens do *dataset* ST com as características extraídas e fornecidas pelos pesquisadores da Universidade de Calgary.
  - **Discretização das características:** utilização de dois algoritmos para analisar o impacto da discretização das características na classificação dos nódulos.
  - **Seleção de características:** descreve a utilização da importância de Gini no processo de seleção de características.
  - **Modelos gramaticais:** apresenta os resultados obtidos com três modelos gramaticais distintos na classificação dos nódulos.
  - **Utilização de outros classificadores:** apresenta os resultados obtidos pelos classificadores ANN, KNN, SVM e RF.
  - **Discuções:** analisa o desempenho obtido pelos modelos gramaticais quando comparado com os demais classificadores na tarefa de classificação dos nódulos.
- **Classificação dos nódulos com características próprias:** exibe os resultados obtidos utilizando as imagens dos *datasets* ST e ACC com as características extraídas por algoritmos implementados durante este projeto de pesquisa.
  - **Criação do modelo poligonal:** utilização de dois algoritmos para analisar o impacto do modelo poligonal na classificação dos nódulos. Para esta análise foram utilizadas apenas imagens do *dataset* ST, características de

forma extraídas pelos algoritmos implementados neste projeto e classificadores ANN, SVM e KNN.

  - **Discretização das características:** utilização do algoritmo Ômega para discretizar as características extraídas das imagens dos *datasets* ST e ACC a partir do modelo poligonal RDP.
  - **Seleção das características:** descreve a utilização da importância de Gini no processo de seleção de características.
  - **Modelos gramaticais:** exibe os resultados obtidos pelos modelos gramaticais no processo de classificação dos nódulos encontrados nas imagens dos *datasets* ST e ACC.
  - **Outros classificadores:** apresenta os resultados obtidos pelos classificadores ANN, KNN, SVM, RF e LGBM.
  - **Discuções:** analisa o desempenho dos modelos apresentados.

- **Limitações** e **Vantagens:** apresenta as limitações e as vantagens da utilização de modelos gramaticais para esta tarefa de classificação.

7.1 Classificação dos nódulos com características de Calgary

Esta seção apresenta os experimentos e resultados obtidos pelos classificadores gramaticais quando utilizados com as características dos nódulos extraídas pelos pesquisadores da Universidade de Calgary – Canadá. Os experimentos foram realizados utilizando as imagens do *dataset* ST detalhado na Seção 4.1.1.

Os experimentos foram realizados utilizando a técnica de validação cruzada (*k-fold cross validation*) com $k = 23$ para testar os três modelos gramaticais propostos e suas variações (diferentes números de características). Este valor de k foi escolhido devido à pequena quantidade de imagens disponíveis no *dataset* ST, o que possibilitou que o treinamento do modelo utilizasse 106 imagens e os testes fossem realizados com cinco imagens, aproximadamente, a cada iteração.

7.1.1 Discretização das características

Conforme descrito na Seção 5.4, os algoritmos Ômega e KBinsDiscretizer foram utilizados no processo de discretização das características.

Durante este processo, para o algoritmo Ômega foi empregado um processo de calibração que variou os valores do parâmetro H_{min} (utilizando os valores 2, 3, 4 e 5) que especifica a menor quantidade de elementos que um *bin* pode conter e do parâmetro ζ_{max} (0,35, 0,40 e 0,45) que especifica o nível máximo de inconsistência aceito em cada *bin*.

Para o algoritmo KbinsDiscretizer, foram variados os valores do parâmetro *n_bins* (20, 30 e 40), o valor do parâmetro *encode* foi mantido fixo ('ordinal'), enquanto o valor utilizado do parâmetro *strategy* foi 'uniform' para que todos os *bins* tivessem sempre a mesma largura.

Com os algoritmos e seus parâmetros definidos, foi possível analisar o impacto que a etapa de discretização de características exerce no processo de classificação nos nódulos. Os resultados dos testes são apresentados na Seção 7.1.3.

7.1.2 Seleção de características

A importância de Gini foi utilizada como critério de seleção das características durante o processo de criação dos modelos gramaticais conforme explicado na Seção 5.3. Para a seleção das características mais importantes todas as imagens do *dataset* ST serviram como input para o RF implementado, sendo que as características mais importantes foram utilizadas nos modelos gramaticais. A Tabela 7 exibe as características ordenadas pela importância de Gini.

Nos modelos criados, exceto quando todas as características de forma e de textura/gradiente foram utilizadas, foi limitado o número de cada categoria de característica a 70% da quantidade daquela categoria de característica. Por exemplo, quando a importância de Gini foi utilizada para selecionar as características de forma, o maior número de características utilizada foi cinco (de um total de oito características possíveis). Para textura/gradiente, quando a técnica foi empregada, o maior valor de características utilizada foi dois (de um total de quatro características possíveis). O valor de 70% da quantidade da categoria da característica foi escolhido devido à quantidade limitada de características disponíveis. Um valor superior a 70% implicaria na utiliza-

Tabela 7: Importância de Gini calculada para cada característica de forma, textura e gradiente. A tabela está ordenada pela importância de Gini.

Característica	Categoria	Importância de Gini
Dimensão fractal 1D Ruler	forma	0,16172
Índice de espiculação	forma	0,13561
Concavidade fracionada	forma	0,12935
Fator de Fourier	forma	0,10775
Dimensão fractal 2D Ruler	forma	0,09420
Dimensão fractal 1D Box counting	forma	0,07918
Compacidade	forma	0,06631
Contraste	textura	0,06326
Dimensão fractal 2D Box counting	forma	0,05663
Acutância tradicional	gradiente	0,04050
Acutância	gradiente	0,03417
Coeficiente de variação	gradiente	0,03126

Fonte: O autor (2023).

ção de quase a totalidade de características disponíveis e um valor abaixo dos 70% limitaria em muito a quantidade máxima de características disponíveis para alimentar os modelos.

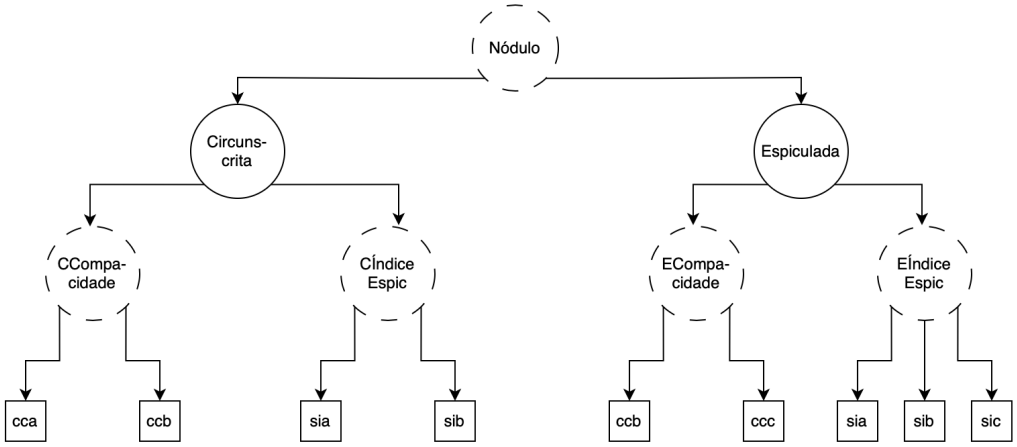
7.1.3 Modelos gramaticais

Foram criados três modelos gramaticais distintos para a realização dos experimentos. O grafo AND-OR do **Modelo 1** faz uso apenas das características de forma e considera o tipo da borda do nódulo (circunscrito ou espiculado). O grafo AND-OR do **Modelo 2** utiliza características de forma e de textura e também considera o tipo da borda do nódulo. O grafo AND-OR do terceiro modelo criado, **Modelo 3**, fez uso de características de forma e textura, mas não utilizou os tipos da borda (circunscrito ou espiculado) na modelagem dos nódulos. As características de forma e de textura utilizadas estão definidas nas Seções 5.2.1.1 e 5.2.1.2. Exemplos dos grafos AND-OR do **Modelo 1**, **Modelo 2** e **Modelo 3** podem ser vistos nas Figuras 31, 32 e 33, respectivamente. Para efeito de simplificação, nos grafos AND-OR (Figuras 31, 32 e 33) apenas um pequeno conjunto de características estão representadas, mas nos experimentos realizados, o **Modelo 1** foi testado com três, quatro, cinco e oito características e os **Modelos 2** e **3** foram testados com cinco, seis, sete e doze características, pois estes últimos também fizeram uso das características de textura/gradiente além das características de forma. As características utilizadas foram selecionadas de acordo

com a importância de Gini apresentada na Tabela 7.

É importante enfatizar que são utilizadas sempre duas gramáticas, uma que representa os nódulos benignos e outra que representa os nódulos malignos, ou seja, o classificador Bayesiano é composto por dois analisadores sintáticos, um criado a partir da gramática que representa os nódulos benignos e outro a partir da gramática que representa os nódulos malignos conforme explicado nas Seções 2.6 e 5.5.4. Para a construção deste modelo, quando o algoritmo Ômega é empregado na discretização são considerados classes Benigno e Maligno e o tipo da borda do nódulo (circunscrito e espiculado). Quando o algoritmo KBinsDiscretizer é utilizado não é necessário considerar nenhuma classe ou tipo da borda do nódulo.

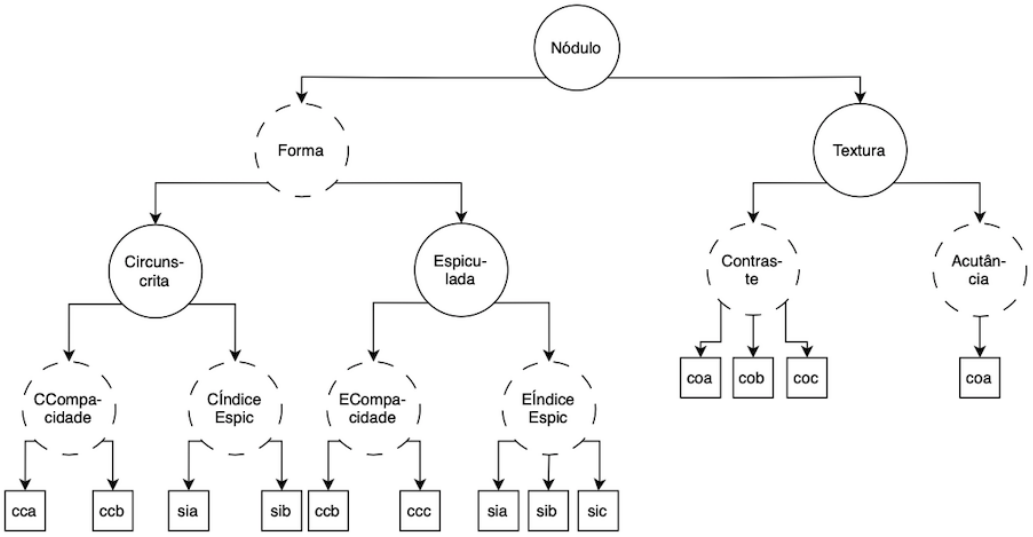
Figura 31: Grafo AND-OR do **Modelo 1** com características de forma e considerando o tipo da borda do nódulo (circunscrito ou espiculado).



Fonte: O autor (2023).

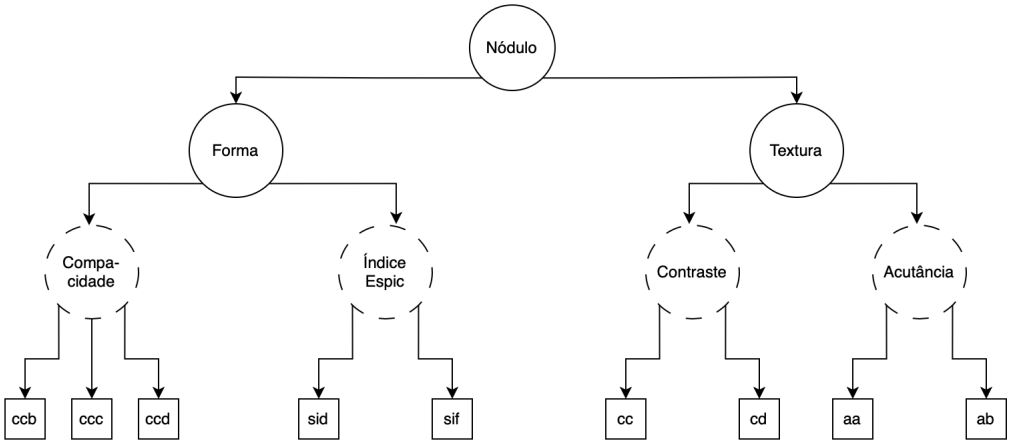
As melhores acurácias, considerando três, quatro, cinco e oito características, alcançadas pelo **Modelo 1** foi 97% tanto utilizando o algoritmo Ômega quanto o algoritmo KBinsDiscretizer para realizar a discretização das características. Para o **Modelo 2** as melhores acurácias atingidas foram de 100% com os dois algoritmos de discretização. Com o **Modelo 3** foi possível alcançar acurácia de 93% utilizando o algoritmo Ômega e 100% utilizando o algoritmo KBinsDiscretizer. As tabelas com os melhores resultados para cada quantidade de características utilizadas (Tabelas 14, 15 e 16) podem ser vistas no Apêndice C. É importante ressaltar que diferentes hiperparâmetros foram testados para os algoritmos de discretização, conforme relatado na Seção 7.1.1, mas as melhores acurácias foram obtidas com $H_{min} = 2$ e $\zeta_{max} = 0,35$ para o Ômega e $n_{bins} = 40$ para o algoritmo KbinsDiscretizer para todos os cenários testados.

Figura 32: Grafo AND-OR do **Modelo 2** com características de forma e textura e considerando o tipo da borda do nódulo (circunscrito ou espiculado).



Fonte: O autor (2023).

Figura 33: Grafo AND-OR do **Modelo 3** com características de forma e textura sem considerar o tipo da borda do nódulo (circunscrito ou espiculado).



Fonte: O autor (2023).

7.1.4 Outros classificadores

De acordo com a revisão sistemática detalhada no Capítulo 3, as técnicas de reconhecimento de padrões mais utilizadas na classificação dos nódulos são ANN, SVM e KNN, nesta ordem. Além dessas técnicas, o algoritmo RF também foi utilizado, pois existem muitos trabalhos que utilizaram RF ou árvores de decisões para lidar com o problema de classificação de nódulos (Capítulo 3).

Para uma classificação comparável aos resultados anteriormente apresentados,

os classificadores foram avaliados utilizando uma validação cruzada na qual cada *fold* possuía exatamente as mesmas imagens utilizadas por todos os classificadores, incluindo os classificadores gramaticais. Antes de executar os testes, todas as características foram normalizadas utilizando o algoritmo *MinMaxScaler* (PEDREGOSA et al., 2011) e, consequentemente, tiveram seus valores pertencentes ao intervalo [0,1]. Esta etapa é importante para evitar que características com escalas diferentes influenciem nos resultados. As mesmas características utilizadas nos modelos gramaticais (características de forma e de textura/gradiente) também foram utilizadas como entrada para os modelos não gramaticais.

A Tabela 8 exibe os hiperparâmetros utilizados na execução de cada classificador. Todas as possíveis combinações das 12 características (oito de forma e 4 de textura/gradiente) foram testadas, dando um total de 4095 subconjuntos diferentes. Esta abordagem de *força bruta* foi empregada para garantir que no final dos testes terá sido obtida a maior acurácia possível para cada classificador considerando os hiperparâmetros escolhidos. Desta forma foi possível verificar como se comparam as melhores acurácias obtidas pelos classificadores ANN, SVM, KNN e RF (nos cenários propostos) com os classificadores gramaticais que utilizaram a importância de Gini como método para seleção de características.

Tabela 8: Hiperparâmetros testados para cada classificador. Para ANN: α é o termo de regularização; $learning\_rating$ é o tamanho do passo utilizado para atualizar os pesos; $n\_neurons$ é o número de neurônios na camada escondida (apenas uma camada escondida foi utilizada); e $f\_activation$ é função de ativação. Para o SVM: $kernel$ é o tipo de kernel utilizado; C é o parâmetro de penalidade do erro; γ é o coeficiente do kernel. Para KNN: k é o número de vizinhos. Para RF: $n\_estimators$ é o número de árvores utilizadas; $max\_features$ é o número de características consideradas ao procurar a melhor divisão.

Classificador	**Hiperparâmetros**
ANN	$\alpha = 0.0001$; $learning\_rating = 0.001, 0.01, 0.1, 1$; $n\_neurons = 2, 3$; $f\_activation$ = sigmoid, hyperbolic tangent, linear
SVM	$kernel$ = linear, polynomial, radial basis function; $C = 0.01, 0.1, 1, 5, 10, 50, 100$; $\gamma = \frac{1}{n\_features}$
KNN	$k = 1, 3, 5, 7, 9$
RF	$n\_estimators = 100$; $max\_features = \sqrt{n\_features}$

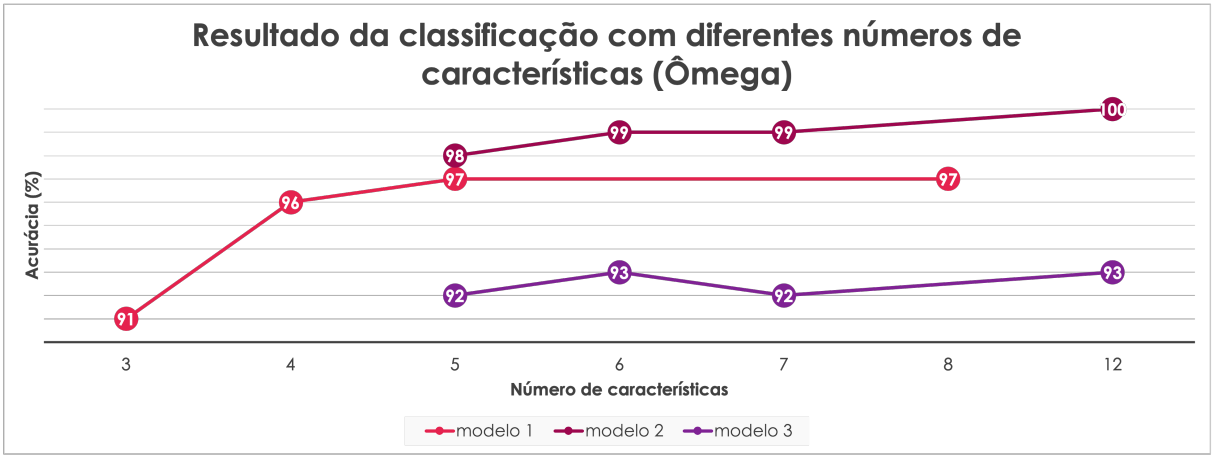
Fonte: O autor (2023).

A maior acurácia (92%) foi obtida pelos classificadores ANN e SVM. Os melhores resultados obtidos podem ser vistos na Tabela 17 no Apêndice C.

7.1.5 Discussões

A Figura 34 exibe as melhores acurácias atingidas por cada modelo gramatical criado quando o número de características mudam e quando o algoritmo Ômega foi utilizado no processo de discretização. Note que o **modelo 1** utiliza menos características que os outros modelos, pois este modelo não faz uso de características de textura/gradiente. Como pode ser visto, o **modelo 1** é o modelo que tem o maior impacto quando o número de características diminui (acurácia vai de 97% com oito características de forma para 91% com três características). Os outros dois modelos gramaticais provaram-se ser mais estáveis com relação à diminuição do número de características. Para os **modelos 2** e **3**, inicialmente foram utilizadas doze características (oito de forma e quatro de textura), seguido de sete características (cinco de forma e duas de textura), em seguida foram utilizadas seis características (quatro de forma e duas de textura) e finalmente cinco características (três de forma e duas de textura). A seleção de características foi feita utilizando a importância de Gini exibida na Tabela 7.

Figura 34: Acurácias obtidas considerando os três modelos gramaticais - Discretização: Ômega. O **modelo 1** utiliza três, quatro, cinco e oito características de forma. Os **modelos 2** e **3** utilizam cinco, seis, sete e doze características.

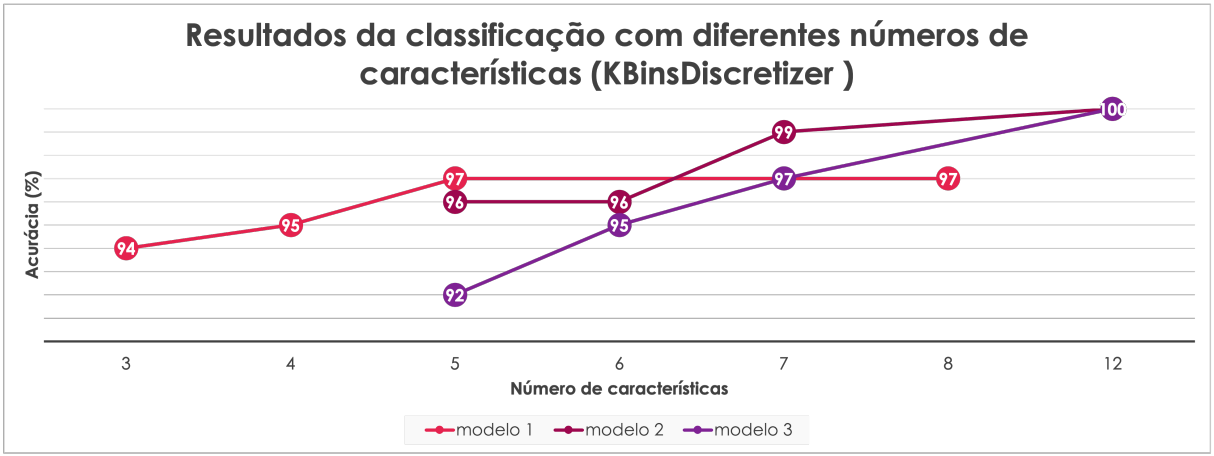


Fonte: O autor (2023).

A Figura 35 mostra as maiores acurácias considerando os três modelos gramaticais propostos quando o número de características utilizada é alterado e o algoritmo KbinsDiscretizer é empregado. De forma geral, pode ser percebido que quando todas

as características são utilizadas as acurácias são maiores que as acurácias obtidas quando o algoritmo Ômega é utilizado. Entretanto, conforme o número de características diminui, a acurácia tende a decrescer de forma mais abrupta. Por esta razão, a discretização com o algoritmo Ômega aparenta ser mais robusto pra esta problema, talvez pelo fato de considerar a informação do tipo da borda (circunscrito ou espiculado) e a classe do nódulo (benigno ou maligno) durante o processamento das características.

Figura 35: Acurácias obtidas considerando os três modelos gramaticais - Discretização: KbinsDiscretizer. O **modelo 1** utiliza três, quatro, cinco e oito características de forma. Os **modelos 2** e **3** utilizam cinco, seis, sete e doze características.



Fonte: O autor (2023).

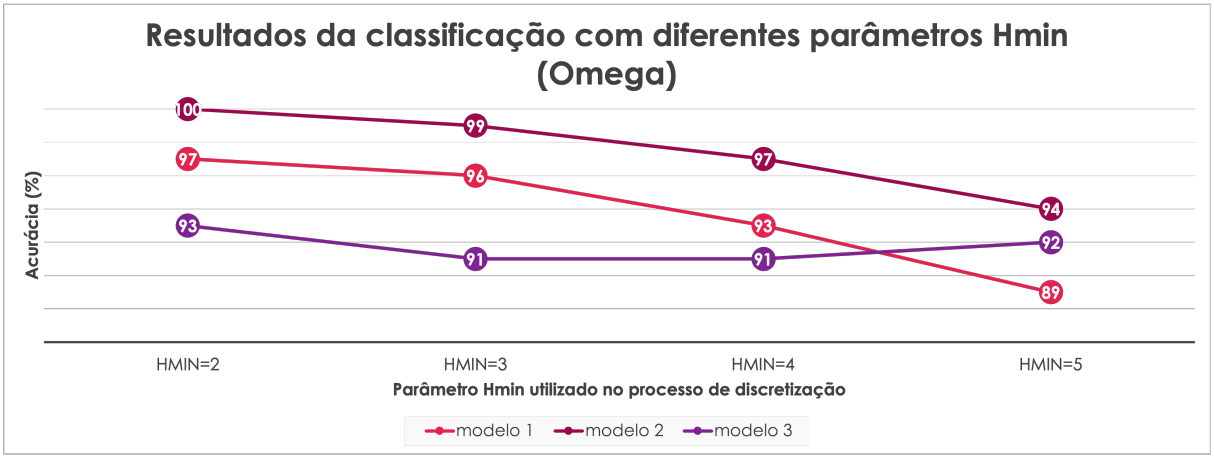
De forma geral, uma possível causa para que o **modelo 1** seja menos robusto que os demais pode ser a ausência de características de textura/gradiente. Geralmente, os nódulos circunscritos tendem a ser benignos enquanto nódulos espiculados tendem a ser malignos. Consequentemente, as características de forma são ideais para serem utilizadas no processo de classificação. Entretanto, alguns nódulos benignos podem ser espiculados e alguns nódulos malignos podem ser circunscritos e, neste cenário, a ausência de características de textura/gradiente e o número limitado de características de forma podem comprometer o desempenho do modelo. Outra possível razão pode ser que com o número muito limitado de características de forma (apenas três em um dos cenários) as gramáticas utilizadas não foram genéricas o suficiente para lidar com nódulos que não estavam presentes no conjunto do treinamento, o que acabou levando a classificações errôneas.

Além disso, na média, o **modelo 3** obteve resultados inferiores quando comparado aos outros dois modelos. Este fato pode ser utilizado para ilustrar a importância de considerar o tipo da borda indicando se os nódulos eram circunscritos ou espiculados

na criação dos modelos gramaticais.

A Figura 36 exibe as maiores acurácias alcançadas por cada modelo gramatical quando todas as características foram utilizadas. Pode-se observar a importância do parâmetro H_{min}, pois quanto maior o valor do parâmetro, menores são as acurácias obtidas. Todos os modelos gramaticais apresentaram um bom desempenho quando o $H_{min} = 2$, mas as acurácias decrescem em aproximadamente 10% quando o $H_{min} = 5$. Foi percebido durante os experimentos que o valor do parâmetro ζ_{max} teve um impacto muito pequeno durante os testes praticamente não alterando os resultados obtidos. Por este motivo os resultados obtidos quando ζ_{max} foi alterado não está ilustrado na Figura 36.

Figura 36: Acurácias obtidas no processo de classificação considerando os três modelos gramaticais e todas as características (oito de forma e quatro de textura/gradiente) e diferentes valores do parâmetro H_{min}.

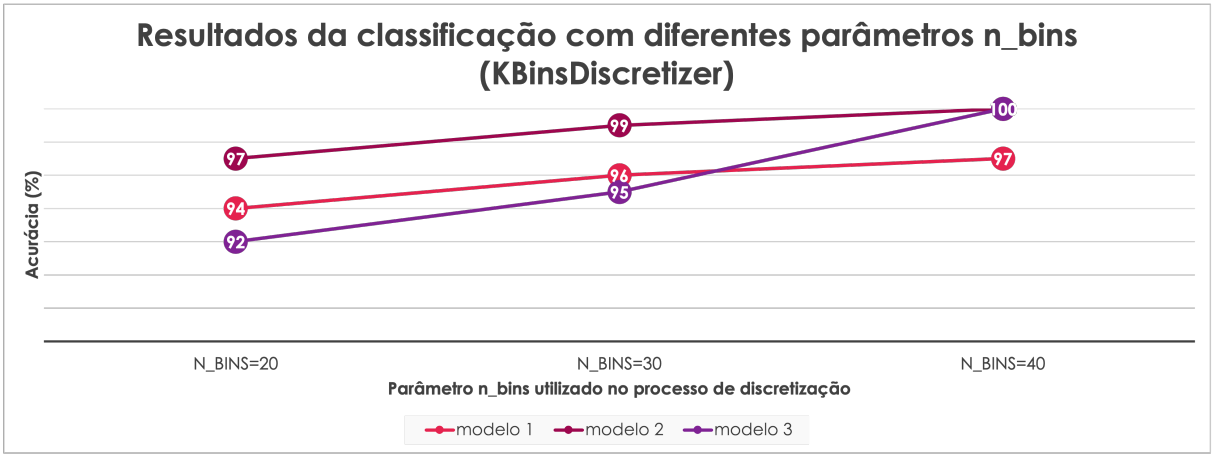


Fonte: O autor (2023).

A Figura 37 exibe as maiores acurácias obtidas quando o algoritmo KbinsDiscretizer foi utilizado no processo de discretização e todas as características foram utilizadas para construir os modelos. Pode-se perceber que conforme o número de *bins* (parâmetro *n_bins*) aumenta a acurácia do modelo também aumenta. Este comportamento também está presente quando o algoritmo Ômega foi utilizado, mas naquele caso o número de *bins* é influenciado pelo parâmetro H_{min}.

O processo de discretização mostrou ser crucial para o desempenho do método proposto. O parâmetro H_{min} é um parâmetro de entrada do algoritmo Ômega que restringe o número mínimo de elementos que cada *bin* deve ter. De forma geral, quando o valor de H_{min} é alto, menos *bins* são obtidos durante o processo de discretização. Consequentemente, quanto maior o parâmetro H_{min}, maiores são as inconsistências

Figura 37: Acurácias obtidas no processo de classificação considerando os três modelos gramaticais e todas as características (oito de forma de quatro de textura/gradiente) e diferentes valores do parâmetro n_bins.



Fonte: O autor (2023).

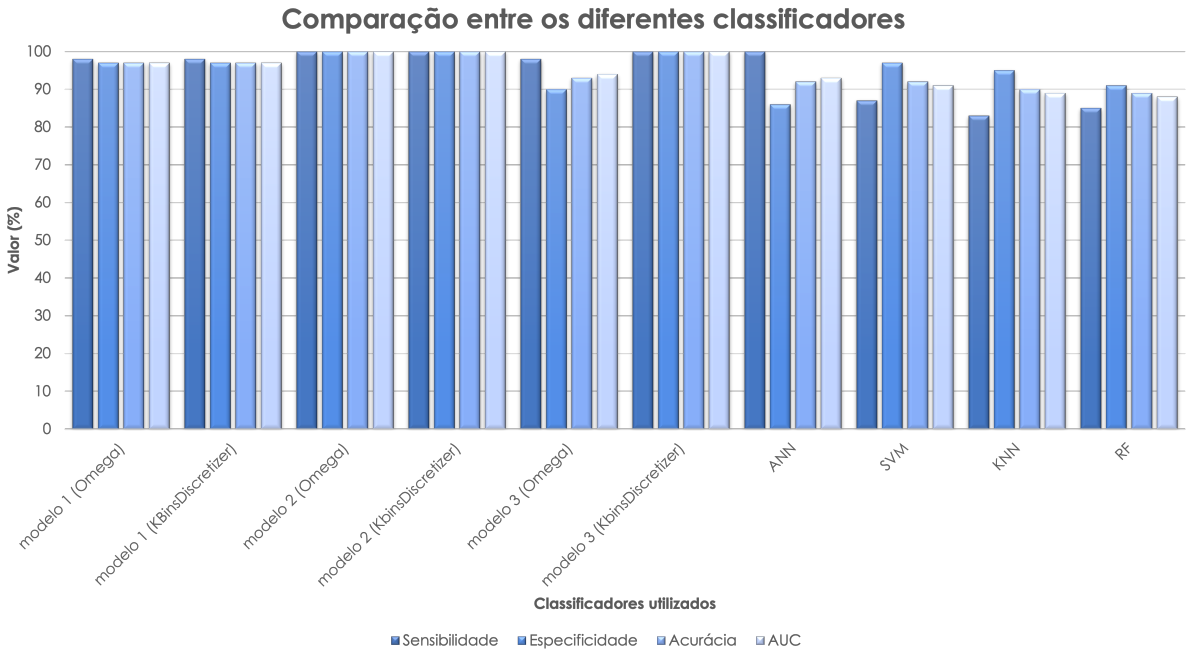
geradas durante o processo. Considerando este fato, é importante manter o valor desse parâmetro o menor possível, mesmo que apenas uma pequena redução no número de *bins* seja atingida (RIBEIRO et al., 2008). O mesmo comportamento também é válido para o algoritmo KbinsDiscretizer, ou seja, é importante escolher um valor adequado para o parâmetro n_bins, pois as acurácias obtidas pelos modelos tendem a ser mais altas para valores mais altos desse parâmetro, conforme ilustrado nas Figuras 36 e 37. Neste trabalho, os algoritmos de discretização foram aplicados com os mesmos parâmetros para todas as características. Entretanto, pode ser que os modelos gramaticais criados utilizando características discretizadas com algoritmos configurados com parâmetros diferentes apresentem medidas de desempenho diferentes. Por exemplo, pode ser que um modelo gramatical formado por duas características obtenha a maior acurácia quando uma característica é discretizada utilizando o parâmetro $n\_bins = 40$ e a outra característica é discretizada utilizando o parâmetro $n\_bins = 30$. Embora esta análise não tenha sido realizada neste projeto, ela pode ser feita em trabalhos futuros.

Analisando as Figuras 34, 35, 36 e 37 podemos observar que os modelos propostos apresentam acurácias maiores quando o número de características e o número de *bins* aumentam. O **modelo 2** obteve desempenho superior aos demais modelos ressaltando a importância das características de textura e da informação do tipo de borda (circunscrito/espiculado) empregadas neste modelo. O **modelo 1** tem desempenho igual ou superior ao **modelo 3** quando seis ou mais características são empregadas, ilustrando a importância da informação do tipo de borda na criação dos modelos. En-

tretanto, quando o número de características diminui, esses dois modelos tendem a apresentar um resultado similar. Neste sentido, embora tanto a informação do tipo de borda, quanto as características de textura sejam importantes, a informação do tipo de borda parece ter uma importância maior que a importância das características de textura para o desempenho dos modelos gramaticais.

Os modelos gramaticais foram comparados com alguns dos modelos mais utilizados encontrado na literatura: ANN, SVM, KNN e RF. As maiores acurácias obtidas pelos modelos gramaticais foram superiores às maiores acurácias obtidas pelos outros modelos em aproximadamente 10% nos experimentos executados. Entretanto, esses resultados não implicam que as abordagens mais tradicionais não devam continuar a ser exploradas para a classificação dos nódulos e sim que os resultados atingidos pelos modelos baseados em gramáticas poderiam ser mais explorados no problema de classificação dos nódulos. A Figura 38 exibe a comparação entre os melhores resultados alcançados pelos classificadores gramaticais e pelos classificadores mais encontrados na literatura.

Figura 38: Maiores acurácias alcançadas no processo de classificação dos nódulos considerando os três modelos gramaticais (modelos 1, 2 e 3) e os modelos mais encontrados na literatura.



Fonte: O autor (2023).

7.2 Classificação dos nódulos com características próprias

Nesta seção são exibidos os resultados dos experimentos realizados quando foram utilizadas as características dos nódulos extraídas por algoritmos implementados durante este projeto de doutorado conforme explicado na Seção 5.2. Foram utilizadas as imagens dos *datasets* ST e ACC detalhados na Seção 4.1. Entretanto, para avalidar o impacto dos modelos poligonais (Seção 7.2.1) no processo de classificação apenas o *dataset* ST foi utilizado, pois o *dataset* ACC ainda não estava disponível.

Apesar da importância das características de textura/gradiente mencionada na Seção 7.1, durante este projeto de pesquisa apenas as características de forma e momentos de Hu foram extraídas, pois as imagens do *dataset* ST não estavam disponíveis. Para esta base em questão, estavam disponíveis apenas as coordenadas dos pixels dos contornos dos nódulos, o que nos possibilitava reconstruir o contorno dos nódulos e extrair as características a partir dali.

7.2.1 O modelo poligonal

Foram implementados os algoritmos Ramer-Douglas-Pecker (RDP) e Peter Borne (PB) na criação dos modelos poligonais para representação dos nódulos, conforme detalhados na Seção 5.1. Também foram implementados algoritmos para extração das características da forma dos nódulos que estão detalhados na Seção 5.2. Apenas as imagens do *dataset* ST foram utilizadas nesta etapa e como apenas os contornos das imagens estavam disponíveis não foi possível fazer a extração de características de textura e gradiente.

O objetivo desta etapa foi verificar a influência que os modelos poligonais exerciam na classificação dos nódulos, para que, após definido o modelo poligonal mais adequado (com base no desempenho de classificadores não gramaticais), fossem realizados testes com os modelos gramaticais e com o *dataset* ACC. O teste t de Student foi executado para cada par de parâmetros utilizados na criação dos modelos utilizando os algoritmos RDP e PB considerando 100 *holdouts*. A hipótese alternativa foi que o modelo poligonal (a escolha do parâmetro) tinha uma influência significativa no desempenho da classificação dos nódulos quando eram utilizadas as características extraídas a partir desse modelo. A hipótese nula foi que o modelo poligonal não tinha uma influência significativa no processo de classificação. Além disso, duas

matrizes $n \times n$ para cada modelo poligonal foram criada, na qual n é o número de modelos poligonais criados. A primeira matriz contém os p-valores considerando a acurácia da classificação (M_{acc}), enquanto que a segunda matriz contém os p-valores considerando a AUC estimada da classificação (M_{AUC}) para cada par (*i,j*) de modelo poligonal.

Ramer-Douglas-Peucker: Para determinar se o modelo poligonal tinha alguma relevância na extração das características de forma, foram extraídas características dos modelos poligonais criados considerando os valores de $\epsilon = 0,1\%, 0,2\%, ..., 2\%$ (incremento de 0,1 a cada modelo criado) do perímetro do contorno do nódulo, totalizando 20 modelos poligonais.

Peter Borne: Para criar os modelos poligonais foram considerados como parâmetros o contorno original do nódulo e um número de pontos alvo (quantidade de pontos do modelo final) com base no número de pontos gerados pelo algoritmo RDP. A quantidade média de pontos para cada modelo criado utilizando o algoritmo PB foi 100.54, 50.33, 34.81, 29.38, 25.50, 22.93, 20.48, 18.35, 16.74, 15.38, 14.26, 13.25, 12.23, 11.59, 10.92, 10.43, 9.96, 9.58, 9.18 e 8.84. Desta forma, os modelos poligonais produzidos pelos dois algoritmos tiveram o mesmo número de pontos e uma comparação mais equânime sobre o processo de classificação pôde ser realizada.

As Tabelas 9 e 10 exibem os melhores resultados obtidos e os parâmetros utilizados na criação dos classificadores não gramaticais e dos modelos poligonais.

Tabela 9: Maiores acurácias obtidas pelos classificadores para os modelos poligonais criados utilizando o algoritmo RDP. $kernel$ é o tipo de kernel utilizado, k é o número de vizinhos, α é o termo de regularização, n_{neuron} é o número de neurônios utilizados na camada escondida (apenas uma camada escondida foi utilizada) e $f_{activation}$ é a função de ativação.

Classificador	**Hiperparâmetros**	ϵ	**Acurácia (%)**
ANN	$\alpha = 0,1$ $n_{neurons} = 2$ $f_{activation} =$ hyperbolic tangent	0,4%	85
SVM	$kernel =$ Linear	0,5%	84
KNN	$k = 6$	0,6%	85

Fonte: O autor (2023).

Na análise executada, o valor ótimo para ϵ fica entre 0.4% e 0.6% do perímetro do contorno para o algoritmo RDP. Para o algoritmo PB, o número de pontos alvo ideal é de aproximadamente 50 pontos para o classificador ANN. Entretanto, quando outros classificadores são utilizados, o número de pontos alvo pode variar.

Tabela 10: Maiores acurácias obtidas pelos classificadores para os modelos poligonais criados utilizando o algoritmo PB. $kernel$ é o tipo de kernel utilizado, k é o número de vizinhos, α é o termo de regularização, n_{neuron} é o número de neurônios utilizados na camada escondida (apenas uma camada escondida foi utilizada) e $f_{activation}$ é a função de ativação.

Classificador	Hiperparâmetros	N. de pontos	Acurácia (%)
ANN	$\alpha = 0,001$ $n_{neurons} = 2$ $f_{activation} =$ identity	50,33	85
SVM	$kernel =$ Linear	29,38	84
KNN	$k = 8$	20,48	84

Fonte: O autor (2023).

Para facilitar a análise e a visualização, as matrizes com os p-valores foram binarizadas: a célula (i, j) é representada em cor escura se o p-valor é menor que $0,05$, indicando a rejeição da hipótese nula, isto é, o resultado da classificação (acurácia ou AUC estimada) utilizando as características extraídas dos modelos poligonais criados são estatisticamente diferentes, caso contrário a célula é representada em branco.

As Figuras 39, 40 e 41 exibem a M_{acc} e a M_{AUC} para cada classificador testado nesta análise. O comportamento dos p-valores para ambas as matrizes é similar para todos os algoritmos de classificação. De forma geral, é possível observar que as matrizes dos modelos poligonais criados utilizando o algoritmo PB contêm mais células escuras, indicando que este algoritmo é mais sensível à escolha do hiperparâmetro.

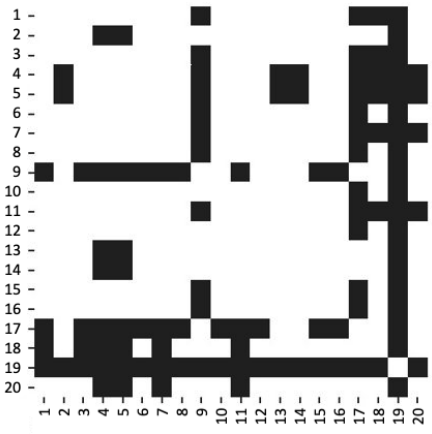
Além de ser mais instável, a utilização dos modelos poligonais utilizando o algoritmo PB não é recompensada com uma maior acurácia na classificação dos nódulos (Tabelas 9 e 10). Por estes motivos, o uso do algoritmo RDP é mais apropriado para a representação poligonal dos nódulos devido à sua estabilidade superior e desempenho igual ou superior no que diz respeito à etapa de classificação e, por esse motivo, foi o algoritmo utilizado nos demais experimentos.

7.2.2 Discretização das características

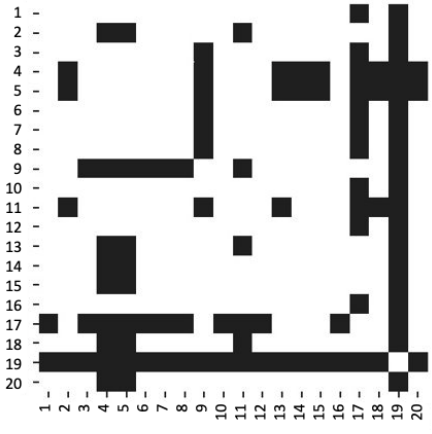
Na Seção 7.2.1 foram apresentados os resultados obtidos quando os algoritmos RDP e PB foram utilizados para representar os contornos dos nódulos. Como foi discutido, o algoritmo RDP se mostrou mais estável e as características dos nódulos utilizadas nos modelos gramaticais foram extraídas a partir dos modelos poligonais criados a partir do algoritmo RDP. Além de algoritmos para extração das características

Figura 39: Comparação dos algoritmos RDP e PB - Classificador ANN. (a) M_{acc} utilizando algoritmo RDP. (b) M_{AUC} utilizando algoritmo RDP. (c) M_{acc} utilizando algoritmo PB. (d) M_{AUC} utilizando algoritmo PB. Céluas em branco representam p-valores maiores que 0,05 e células pretas representam p-valores menores que 0,05.

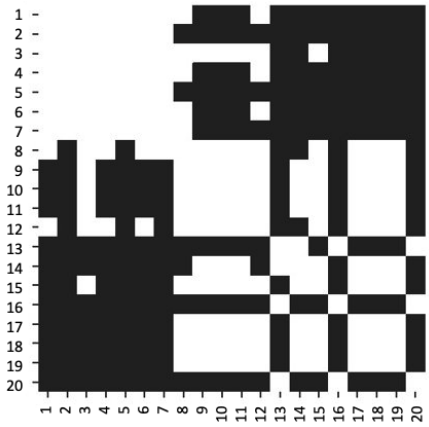
(a) M_{acc} com RDP



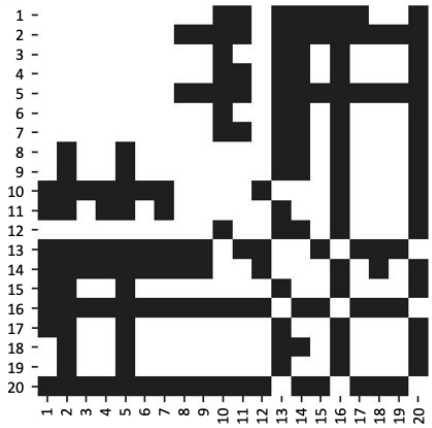
(b) M_{AUC} com RDP



(c) M_{acc} com PB



(d) M_{AUC} com PB



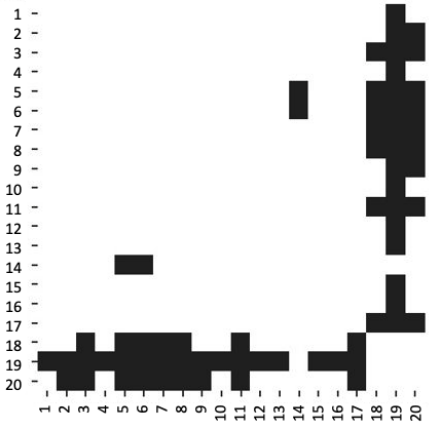
Fonte: (HIRAMA et al., 2020).

de forma, neste projeto de pesquisa também foram implementados algoritmos para extração dos momentos de Hu, explicados nas seções 5.2.1.1 e 5.2.1.3. Entretanto, após as características terem sido extraídas, elas precisam passar pelo processo de discretização explicado na Seção 5.4.

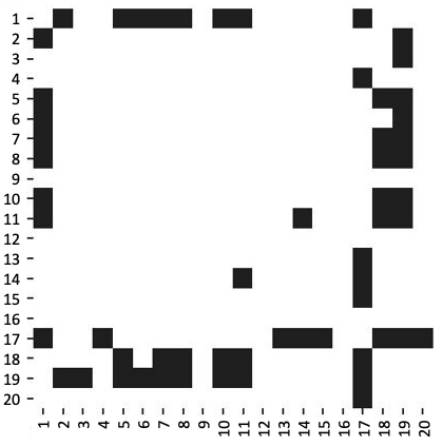
Para a etapa de discretização foi utilizado o algoritmo Ômega, pois ele se mostrou mais robusto que o algoritmo KbinsDiscretizer, conforme discutido na Seção 7.1.5 e exibido nas Figuras 34 e 35. O processo de calibração do algoritmo Ômega encontrou $H_{min} = 2$ e $\zeta_{max} = 0,35$ como melhores valores para os parâmetros de entrada. Ainda assim, para os testes com os modelos gramaticais foram considerados $H_{min} = 2, 4, 6, 8$ e 10, pois este parâmetro mostrou ter um maior impacto no processo

Figura 40: Comparação dos algoritmos RDP e PB - Classificador SVM. (a) M_{acc} utilizando algoritmo RDP. (b) M_{AUC} utilizando algoritmo RDP. (c) M_{acc} utilizando algoritmo PB. (d) M_{AUC} utilizando algoritmo PB. Céluas em branco representam p-valores maiores que 0,05 e células pretas representam p-valores menors que 0,05.

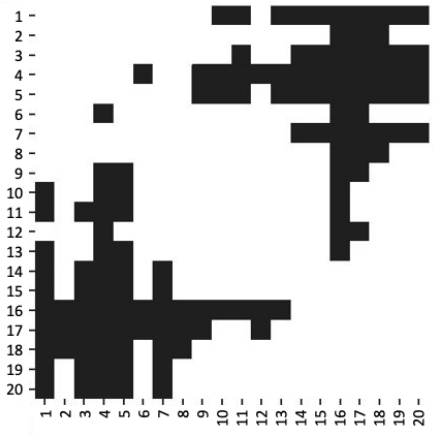
(a) M_{acc} com RDP



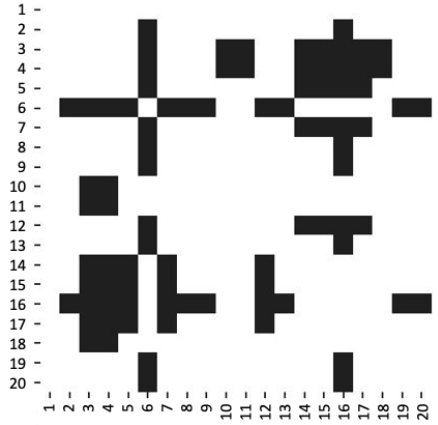
(b) M_{AUC} com RDP



(c) M_{acc} com PB



(d) M_{AUC} com PB



Fonte: (HIRAMA et al., 2020).

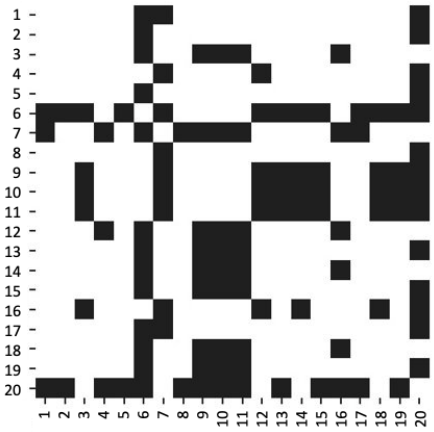
de discretização, enquanto $\zeta_{max} = 0,35$ foi mantido fixo, pois esse parâmetro não apresentou impacto significativo no processo.

7.2.3 Seleção das características

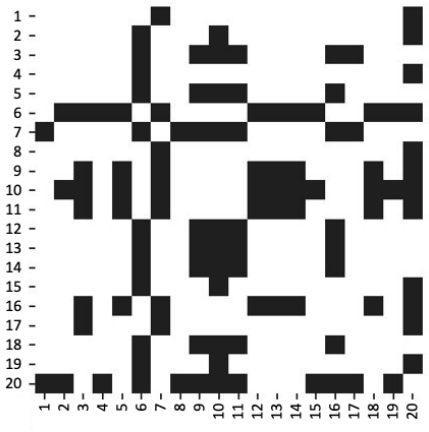
Assim como nos experimentos descritos na Seção 7.1, também foi utilizado o valor da importância de Gini para selecionar as características mais importantes que foram extraídas durante este projeto (Seção 5.3). A medida foi computada utilizando classificadores RF treinado com 100 árvores considerando as classes Benigno e Maligno. Entretanto, para tornar robusto o método utilizado neste projeto de pesquisa, as RF fo-

Figura 41: Comparação dos algoritmos RDP e PB - Classificador KNN. (a) M_{acc} utilizando algoritmo RDP. (b) M_{AUC} utilizando algoritmo RDP. (c) M_{acc} utilizando algoritmo PB. (d) M_{AUC} utilizando algoritmo PB. Céluas em branco representam p-valores maiores que 0,05 e células pretas representam p-valores menors que 0,05.

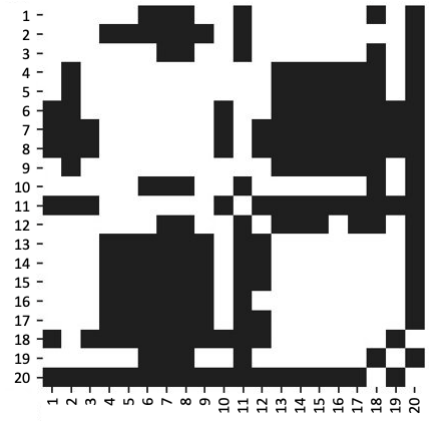
(a) M_{acc} com RDP



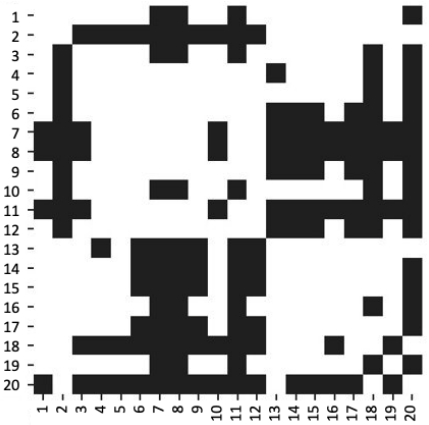
(b) M_{AUC} com RDP



(c) M_{acc} com PB



(d) M_{AUC} com PB



Fonte: (HIRAMA et al., 2020).

ram criadas utilizando apenas as imagens pertencentes aos conjuntos de treinamento, ou seja, quando uma validação cruzada era aplicada, as características mais importantes eram selecionadas considerando apenas as imagens pertencentes aos *folds* utilizados nos treinamentos. De forma semelhante, quando o treinamento acontecia com imagens de um *dataset* e os testes ocorriam com imagens do outro *dataset*, as características mais importantes eram selecionadas considerando apenas o *dataset* de treinamento.

Nos cenários de testes nos quais houve o emprego da combinação de características de forma e momentos de Hu, foram selecionadas as oito características mais importantes de acordo com a importância de Gini. Para os cenários nos quais apenas

as características de forma (oito características) ou apenas os momentos de Hu (sete características) foram utilizadas na classificação dos nódulos, o passo de seleção de características não foi empregado. Esta abordagem foi escolhida para que, nos três cenários, uma quantidade semelhante de características fossem utilizadas possibilitando uma comparação mais justa.

7.2.4 Modelos gramaticais

O modelo gramatical utilizado nesta etapa (Seção 5.5) faz uso apenas das características de forma e dos momentos de Hu para classificar as imagens oriundas dos *datasets* ST e ACC (Seção 9). As características utilizadas foram extraídas das imagens desses dois *datasets* por algoritmos implementados durante este projeto de pesquisa (Seção 5.2). Os grafos AND-OR utilizados para representar as gramáticas dos nódulos benignos e dos nódulos malignos são semelhantes ao grafo AND-OR exibido na Figura 31. Nesta figura, para efeito de simplificação, estão ilustradas apenas as características compacidade e índice de espiculação, mas as demais características seguem o mesmo padrão sendo representadas como nós filhos dos nós “Circunscrita” e “Espiculada”.

Diferentes cenários foram construídos para avaliar os classificadores baseados em gramáticas com relação às suas habilidades de classificar nódulos em imagens de diferentes bases. Quatro experimentos foram executados, dos quais três utilizaram validação cruzada estratificada:

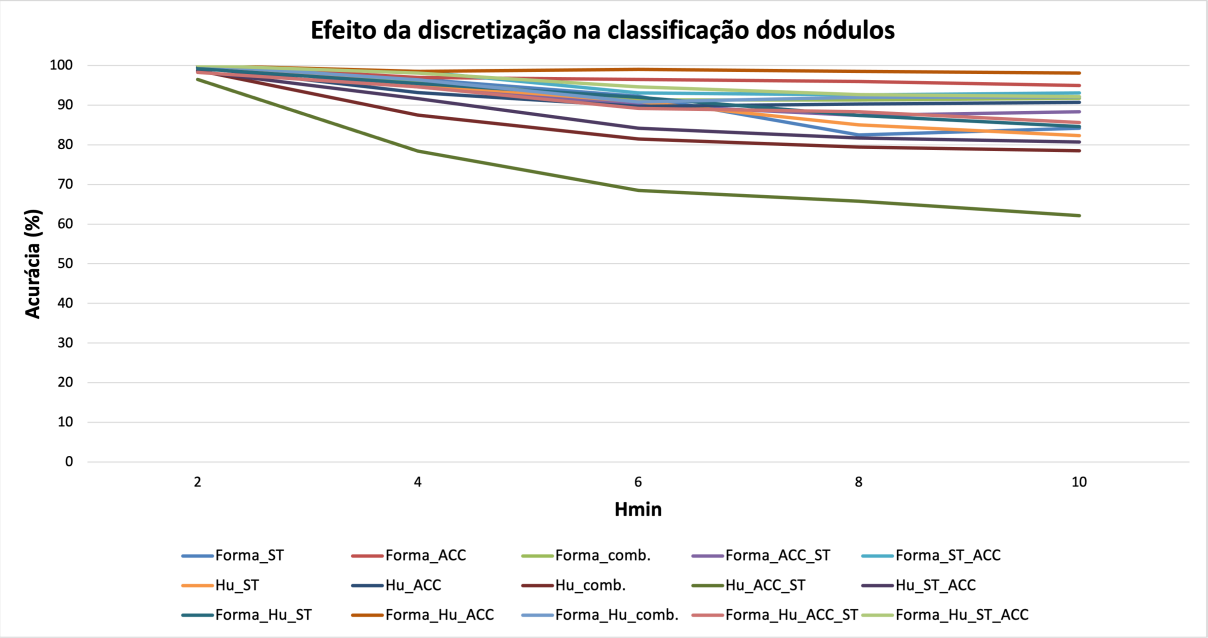
- utilizando apenas o *dataset* ST (k = 23);
- apenas o *dataset* ACC (k = 20);
- utilizando as duas bases combinadas (k = 20);
- realizando o treinamento com uma das bases e os testes com a outra.

Embora a validação cruzada seja geralmente executada utilizando $k = 10$, foram utilizados valores maiores para aumentar o número de imagens no conjunto de treinamento em cada iteração, pois os *datasets* ST e ACC têm um número limitado de imagens. Além disso, os valores diferentes utilizados para k tendem a acomodar melhor a distribuição das imagens em cada iteração do processo. Por exemplo, considerando o *dataset* ST e $k = 23$, o classificador é treinado com aproximadamente 106 imagens e testado com cinco imagens a cada iteração.

Pôde-se perceber que a discretização das características influencia o desempenho dos classificadores, assim como ocorreu nos experimentos reportados na Seção 7.1. A Figura 42 exibe o efeito que o processo de discretização exerce sobre a classificação, pois conforme o valor do parâmetro H_{min} aumenta, a acurácia tende a diminuir. A maior diferença foi notada quando foram utilizados momentos de Hu, o treinamento foi realizado utilizado o *dataset* ACC e os testes foram realizados com as imagens do *dataset* ST (Hu_ACC_ST na Figura 42), na qual a acurácia decaiu de 96% quando $H_{min} = 2$ para 62% quando $H_{min} = 10$.

A menor diferença entre os valores de acurácia obtidos foi mediante a combinação das características de forma e momentos de Hu utilizando o *dataset* ACC (Forma_Hu_ACC na Figura 42). Neste cenário, a acurácia decaiu de 100% ($H_{min} = 2$) para 98% ($H_{min} = 10$).

Figura 42: Efeito da discretização na classificação dos nódulos. Na legenda, **Forma** significa as características de forma; **Hu** significa o uso de momentos de Hu; **comb.** significa os *datasets* combinados; **ACC_ST** significa treinado com *dataset* ACC e testado com o *dataset* ST; **ST_ACC** significa treinado com o *dataset* ST e testado com a *dataset* ACC.



Fonte: O autor (2023).

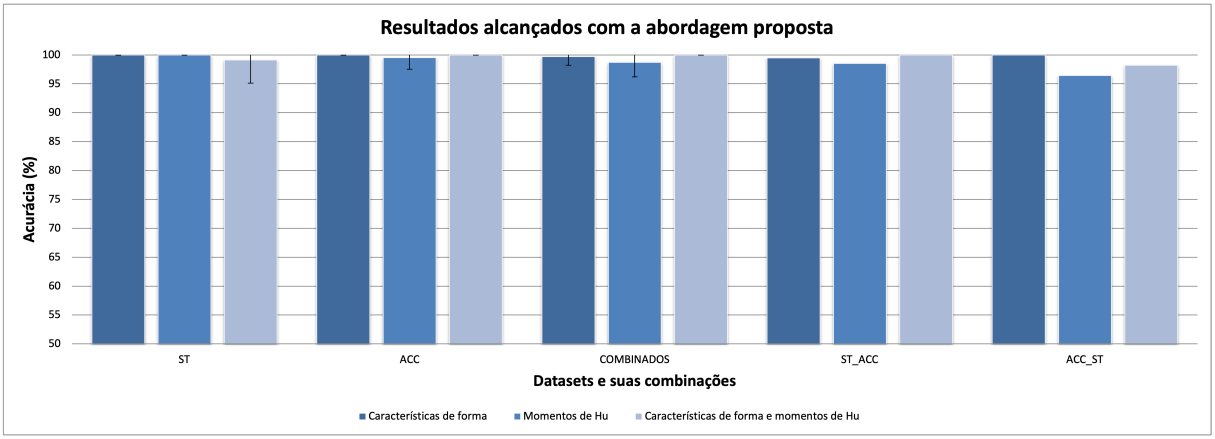
Os próximos resultados apresentados foram obtidos quando o processo de discretização fez uso do parâmetro $H_{min} = 2$, que foi o valor encontrado como o mais adequado para obter as melhores acurácias.

A Figura 43 exibe os resultados da classificação utilizando o modelo gramatical

proposto. As acurácias alcançadas variam entre 98% e 100% para os *datasets* ST, ACC e para estes dois *datasets* combinados.

Para verificar se o processo de discretização utilizando $H_{min} = 2$ estava levando ao *overfiting* nos testes, o modelo foi treinado com imagens de um *dataset* e testado com imagens de outro dataset. As melhores acurácias continuaram variando entre 98% e 100% utilizando características de forma ou a combinação de características de forma com momentos de Hu, mas decaíram quando apenas momentos de Hu foi empregado (96%).

Figura 43: Resultados obtidos com a abordagem gramatical. **COMBINADOS** significa a combinação dos *datasets* ST e ACC; **ACC_ST** significa treinamento com o *dataset* ACC e teste com a *dataset* ST; **ST_ACC** significa treinamento com o *dataset* ST e teste com o *dataset* ACC. A barra de erro representa o desvio padrão encontrado nos cenários nos quais a validação cruzada foi empregada.



Fonte: O autor (2023).

7.2.5 Outros classificadores

Para comparar os resultados obtidos utilizando a abordagem gramatical com outros classificadores, a classificação das imagens dos *datasets* ST e ACC foi realizada utilizando modelos de ANN, SVM, KNN, RF e LGBM. Estes modelos foram treinados considerando os hiperparâmetros apresentados na Tabela 11. Mais precisamente, foram executados experimentos considerando apenas características de forma, apenas momentos de Hu e a combinação dessas características (de forma análoga aos experimentos realizados com os classificadores gramaticais). Além disso, em um cenário de teste foi realizada uma validação cruzada considerando as imagens combinadas dos dois *datasets* e a mesma divisão de *folds* utilizada para o modelos gramaticais, enquanto nos demais cenários o treinamento do modelo foi realizado com imagens de

apenas um *dataset* e os testes foram realizados com imagens do outro dataset. Antes do treinamento dos modelos, todas as características foram normalizadas utilizando o algoritmo *MinMaxScaler* (PEDREGOSA et al., 2011) visando a garantir que seus valores pertencessem ao intervalo [0-1].

Tabela 11: Classificadores criados e seus hiperparâmetros, sendo n_features o número de características e X.var() a variância do conjunto de treinamento.

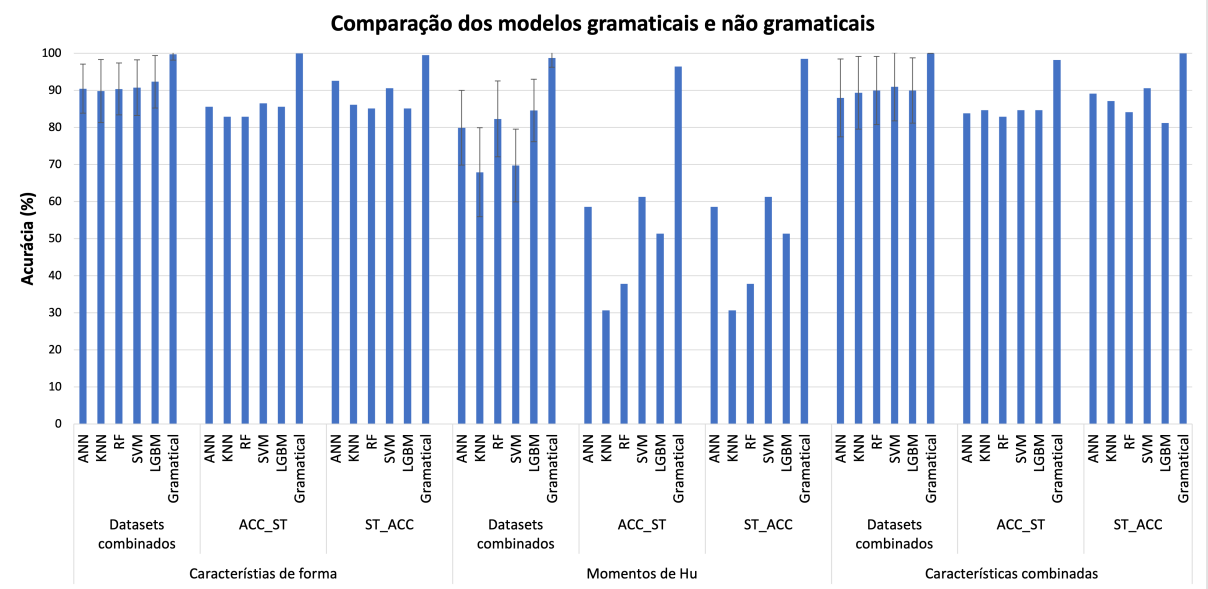
Modelo	**Hiperparâmetros**
ANN	$\alpha = 0.001, 0.0001, 0.00001$; $n\_neurons = 2, 3$; $learning\_rating = 0.001, 0.01, 0.1, 1$; $f\_activation$ = sigmoid, hyperbolic tangent, linear
KNN	$k = 1, 3, 5, 7, 9$
SVM	$\gamma = \frac{1}{n\_features}, \frac{1}{n\_features * X.var()}$ $C = 0.01, 0.1, 1, 5, 10, 50, 100$; $kernel$ = linear, polynomial, radial basis function;
RF	$n\_estimators = 100, 200, 300$; $max\_features = \sqrt{n\_features}, \log_2 n\_features$; $criterion =$ gini, entropy
LGBM	$n\_estimators = 100, 200, 300$; $max\_depth = -1, 5, 10, 20$; $learning\_rate = 1, 0.1, 0.01, 0.001$ $num\_leaves = 10, 20, 30$ $min\_child\_samples = 5, 10, 20$

Fonte: O autor (2023).

A Figura 44 mostra os valores obtidos de acurácias considerando todos os cenários testados. Comparando os resultados obtidos pelo método gramatical proposto com os resultados dos demais classificadores pode-se perceber que os resultados apresentados pela metodologia proposta se mostraram superior. Em especial, a Figura 44 mostra que o modelo proposto foi capaz de classificar os nódulos utilizando momentos de Hu como entrada para o modelo enquanto os demais classificadores mostraram mais dificuldade neste cenário.

Considerando as abordagens não gramaticais, a melhor acurácia obtida foi de 93% com ANN e RF quando características de forma ou a combinação entre características de forma de momentos de Hu foram usados. Quando apenas momentos de Hu foram empregados, os modelos não gramaticais apresentaram seus piores desempenhos, mostrando que esses classificadores não foram capazes de aprender os padrões dos nódulos benignos e malignos com esta categoria de característica.

Figura 44: Comparação dos modelos gramaticais e não gramaticais. Acurácia atingida por cada modelo utilizando características de forma, momentos de Hu e a combinação entre características de forma e momentos de Hu. **datasets combinados**: treinamento = ambos, teste = ambos; **ACC_ST**: treinamento = ACC; teste = ST; **ST_ACC**: training = ST, test= ACC. A barra de erro representa o desvio padrão encontrado nos cenários nos quais a validação cruzada foi empregada.



Fonte: O autor (2023).

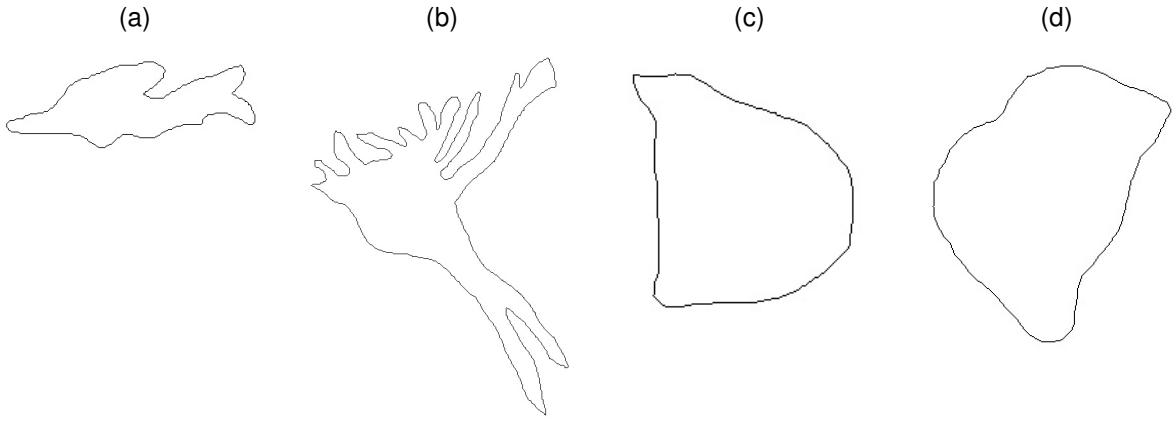
Além da acurácia, as medidas sensibilidade, especificidade, *F1-score* e *Matthews correlation coefficient* (MCC) foram calculadas e apresentadas nas Tabelas 18, 19 e 20 no Apêndice C para facilitar a comparação entre os modelos.

7.2.6 Discussões

De forma geral, as medidas de desempenho apresentaram resultado inferior para o *dataset* ST quando comparado ao *dataset* ACC, especialmente quando o valor do parâmetro H_{min} aumenta (Figura 42). Isto pode ter ocorrido porque o *dataset* ST contém um maior número de imagens com nódulos malignos com o contorno circunscrito e imagens de nódulos benignos com contorno espiculado que as imagens do *dataset* ACC. A Figura 45 exibe exemplos de nódulos benignos com contornos espiculados e nódulos malignos com contornos circunscritos. O contorno é o dado utilizado para o cálculo das características de forma; em geral, espera-se que contornos circunscritos estejam relacionados a nódulos benignos e contornos espiculados estejam relacionados a nódulos malignos. Assim, casos considerados *outliers* em relação a este aspecto podem contribuir para que o aprendizado não seja plenamente suficiente, con-

forme discutido na Seção 2.2.1. Outro fator importante que pode estar influenciando o classificador é que há mais imagens em ACC que em ST (202 e 111, respectivamente) e o classificador pode se tornar mais estável conforme o número de imagens aumenta.

Figura 45: (a) e (b): Exemplos de nódulos benignos com borda espiculada. (c) e (d): Exemplos de nódulos malignos com borda circunscrita.



Fonte: O autor (2023).

Os resultados superiores apresentados pelos métodos gramaticais quando comparados aos demais classificadores testados neste estudo (considerando as imagens disponíveis) podem estar relacionados ao número de imagens presentes nos *datasets*. Os classificadores baseados em gramáticas parecem necessitar de uma quantidade menor de imagens para aprender o padrão dos nódulos benignos e malignos quando comparado aos demais classificadores. Mais experimentos poderiam ser feitos (considerando imagens diferentes *datasets*) para verificar se os métodos gramaticais continuariam apresentando resultados superiores. Em caso afirmativo, os métodos gramaticais poderiam ser mais utilizados neste problema de classificação, em especial quando não há uma grande quantidade de imagens disponíveis para treinamento.

É bastante difícil comparar os resultados obtidos com diferentes estudos nesta área de pesquisa, pois diferentes *datasets*, imagens, métodos de segmentação e medidas de avaliação são utilizadas. Entretanto, pode ser percebido que os resultados obtidos neste projeto de pesquisa são comparáveis aos resultados obtidos por alguns dos trabalhos mais recentes encontrados na literatura em relação a algumas medidas de desempenho. Por exemplo, as melhores acurácias obtida neste estudo (96% a 100%) são semelhantes às acurácias apresentadas em (MOHANTY; RUP; DASH, 2020) (99% a 100%) e (SABER et al., 2021) (98%) e superior à acurácia apresentada em (HEIDARI et al., 2021) (75%).

7.3 Limitações

Apesar dos resultados promissores obtidos pelos classificadores baseados em gramáticas, a abordagem proposta também tem suas limitações.

Uma limitação está relacionada à necessidade de extrair as características dos nódulos que tiveram seus contornos previamente marcados por médicos especialistas ou por um processo de segmentação automático. Esta limitação é bastante significativa, especialmente quando comparada com métodos que utilizam técnicas de *deep learning* nas quais esta etapa não é necessária. As imagens utilizadas neste estudo apresentavam seus contornos anotados por especialistas, uma vez que o desenvolvimento de métodos de segmentação é uma área de pesquisa bastante desafiadora, principalmente para imagens de mamas densas, e não fazia parte do escopo do presente projeto.

Outro fator que pode impactar os resultados obtidos pelos classificadores gramaticais é a etapa de discretização das características. Entretanto, conforme discutido na Seção 7.1.5, esta limitação pode ser superada mediante utilização de algoritmos adequados.

7.4 Vantagens

Uma vantagem da abordagem proposta para classificar nódulos quando comparado com outros classificadores, especialmente com técnicas de *deep learning* é que há indícios de que o padrão dos nódulos benignos e malignos pode ser aprendido a partir de um pequeno conjunto de imagens. Por exemplo, o *dataset* ST tem 111 imagens enquanto o *dataset* ACC tem 202 imagens. Uma possível razão para estes resultados pode estar relacionada ao conjunto de regras escolhidas para criar as gramáticas que representam os nódulos benignos e malignos. Adicionalmente, pode ter contribuído o fato de o classificador Bayesiano ser composto por dois analisadores sintáticos distintos (criados com base nas gramáticas) que calculam as probabilidades da sentença representando o nódulo (sequência de *tokens*/características) pertencer à linguagem de cada gramática. Para uma abordagem que resolva o mesmo problema com técnicas de *deep learning* seria necessário uma quantidade maior de imagens para que o modelo aprendesse todos os pesos necessários para reconhecer os padrões dos nódulos benignos e malignos (KHAN et al., 2019). Para atingir resultados

similares aos apresentados neste projeto, estratégias de *deep learning*, por exemplo, *transfer learning* e/ou *data augmentation* são utilizadas (MALEBARY; HASHMI, 2021) (*dataset* com 1.000 imagens) e (SABER et al., 2021) (*dataset* com 2.576 imagens).

Também pode ser citada como vantagem da abordagem gramatical o fato de os classificadores gerados serem mais estáveis que os classificadores gerados pelas outras abordagens testadas neste projeto, pois as acurácias obtidas pelos classificadores gramaticais tendem a variar menos que as acurácias obtidas pelos demais classificadores (Figura 44).

Como pôde ser visto nas Tabelas 4 e 5, os *datasets* ST e ACC não são perfeitamente balanceados. Este problema de desbalanceamento de classes não é, necessariamente, considerado um problema para os métodos gramaticais, principalmente quando gramáticas estocásticas estão sendo utilizadas. Na verdade, este desbalanceamento será considerado uma informação adicional para o modelo que irá considerar que alguns eventos são mais raros de acontecer e isso será refletido durante a estimação das probabilidades das regras gramaticais.

Além disso, os modelos gramaticais possibilitam que sejam inseridas informações *a priori* na construção dos modelos. Neste trabalho foi adicionada a informação to tipo da borda do nódulo (circunscrita ou espiculada) o que pode ter contribuído para um melhor desempenho deste tipo de classificador.

Outra vantagem do modelo gramatical é a sua explicabilidade inerente, pois usa uma representação uma estrutura hierárquica (Figura 24) ilustrando o que o algoritmo faz, especialmente quando comparado a Redes neurais e outras abordagens mais complexas que se aproximam de “caixas pretas”. Utilizando a Figura 24 como exemplo, seria possível explicar para profissionais da área de saúde que os círculos com as bordas sólidas representam os nós/regras AND, os círculos com as bordas tracejadas representam os nós/regras OR e que os quadrados representam os nós folhas com os valores das características extraídas. Desta forma, acreditamos que mesmo sem conhecimento técnico acerca de métodos gramaticais, os profissionais da área de saúde poderiam entender as regras utilizadas (presentes nas árvores de derivação) para classificar os nódulos nas classes Benigno ou Maligno.

7.5 Considerações

Este capítulo apresentou os resultados de uma nova abordagem sintática para classificar nódulos encontrados em mamogramas como sendo benignos ou malignos. Foram realizados testes utilizando características extraídas por pesquisadores da Universidade de Calgary e características extraídas neste projeto de pesquisa. Dois *datasets* foram utilizados: i) o primeiro contém 111 imagens e foi fornecido por pesquisadores da Universidade de Calgary; ii) o outro *dataset* contém 202 imagens e foi fornecido por pesquisadores do A. C. Camargo Cancer Center.

Os resultados mostraram que a abordagem sintática proposta é robusta, pode aprender o padrão dos nódulos benignos e malignos mesmo com uma pequena quantidade de imagens e apresentou acurácias superiores quando comparado com modelos criados utilizando outras técnicas de aprendizado de máquina, em especial, ANN, SVM, KNN, RF e LGBM. Além disso, os resultados obtidos foram similares aos resultados apresentados em alguns dos estudos mais recentes desta área de pesquisa.

Considerando os testes realizados nos quais os dois *datasets* estavam disponíveis, o modelo gramatical desenvolvido chegou a atingir acurácia de 100%, enquanto as melhores acurácias alcançadas pelos modelos ANN, SVM, KNN, RF e LGBM foram 92%, 90%, 89%, 90% e 92%, respectivamente. Classificadores que utilizaram características de forma também tiveram um desempenho superior na classificação dos nódulos quando comparados aos classificadores que utilizaram apenas momentos de Hu, evidenciando que as características de forma mostraram-se mais adequadas para serem aplicadas neste problema de classificação, considerando os cenários apresentados neste projeto.

Apesar da abordagem proposta depender de uma boa discretização dos dados, os modelos criados provaram não terem sofrido *overfitting*. Os testes mostram acurácias entre 96% e 100% quando o modelo foi treinado com imagens de um *dataset* e os testes foram realizados com imagens de outro *dataset*. Os estudos e experimentos apresentados neste capítulo geraram a publicação de quatro artigos: (PEDRO; MACHADO-LIMA; NUNES, 2019b; HIRAMA et al., 2020; PEDRO; MACHADO-LIMA; NUNES, 2021; PEDRO et al., 2023).

O próximo capítulo apresenta os resultados obtidos com relação à geração de imagens sintéticas de nódulos.

8 GERAÇÃO DE IMAGENS SINTÉTICAS DE NÓDULOS

Neste capítulo são apresentados os experimentos realizados para geração de imagens sintéticas de nódulos mamográficos. Os experimentos são divididos em “Geração de contornos de nódulos benignos e malignos” (Seção 8.1), “Transferência de um nódulo real para um novo mamograma” (Seção 8.2) e “Aplicação de textura para imagens sintéticas de nódulos” (Seção 8.3).

8.1 Geração de contornos de nódulos benignos e malignos

Os contornos dos nódulos sintéticos foram gerados utilizando uma abordagem que faz uso da fusão de contornos de nódulos reais e uma abordagem que altera os contornos de nódulos reais, conforme detalhado nas seções 6.1 e 6.2, respectivamente. Uma vez que os contornos sintéticos tenham sido gerados, as gramáticas criadas para classificação dos nódulos reais foram utilizadas para verificar se os contornos sintéticos seriam classificados dentro da classe para as quais foram gerados. Mais especificamente, foram utilizadas as gramáticas modeladas considerando as oito características de forma, a informação do tipo da borda e as probabilidades das regras gramaticais foram estimadas considerando todas as imagens presentes nos *datasets* ST e ACC. Os momentos de Hu não foram utilizados, pois eles apresentaram um desempenho inferior ao das características de forma. O processo de geração dos contornos e os resultados são descritos nas seções 8.1.1 e 8.1.2.

8.1.1 Fusão de contornos reais

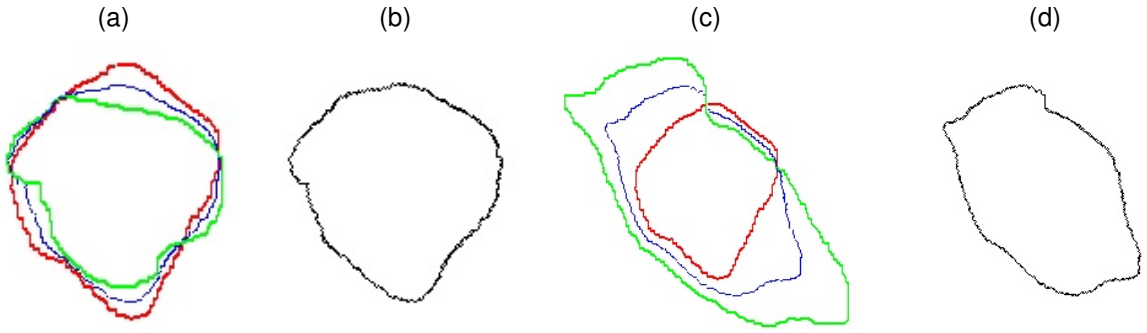
Nesta primeira abordagem, novos contornos foram gerados a partir da fusão de dois outros contornos existentes, conforme detalhado na Seção 6.1. Para gerar um

novo nódulo de uma determinada classe, foi escolhido um contorno de um nódulo daquela classe de forma aleatória, o qual foi combinado de forma independente (sempre em pares) com os demais contornos de nódulos da mesma classe. Assim, por exemplo, um novo contorno de um nódulo benigno é gerado a partir da combinação de dois contornos de nódulos benignos reais. A combinação foi realizada sempre considerando dois contornos de cada vez, tendo sido testados seis algoritmos diferentes, conforme apresentado a seguir. É importante ressaltar que não foram geradas todas as combinações 2 a 2 possíveis, o que resultaria em um total 5.356 contornos gerados a partir dos contornos dos nódulos benignos e 4.465 contornos gerados a partir dos contornos dos nódulos malignos. Neste sentido, apenas um subconjunto de combinações foram geradas, ou seja, escolheu-se o contorno de um nódulo de forma aleatória e fez a fusão deste contorno com o contorno dos demais nódulos, gerando 103 contornos a partir da fusão dos contornos dos nódulos benignos e 94 contornos a partir da fusão dos contornos dos nódulos malignos.

Algoritmo 1: Sem rotacionar o contorno e utilizar o ponto médio entre as bordas.

A Figura 46 mostra dois exemplos da fusão dos contornos de nódulos benignos. Nas Figuras 46a e 46c os contornos em vermelho e verde são os contornos dos dois nódulos benignos reais que serão combinados. O contorno em azul é o contorno resultante intermediário do Algoritmo 1. Nas Figuras 46b e 46d podem ser observados os contornos finais obtidos.

Figura 46: Fusão do contornos de nódulos benignos utilizando o algoritmo 1. (a) e (c) Resultado intermediário; (b) e (d) Resultado final. Contornos vermelhos e verdes são de nódulos reais e contornos em azul são os contornos gerados.

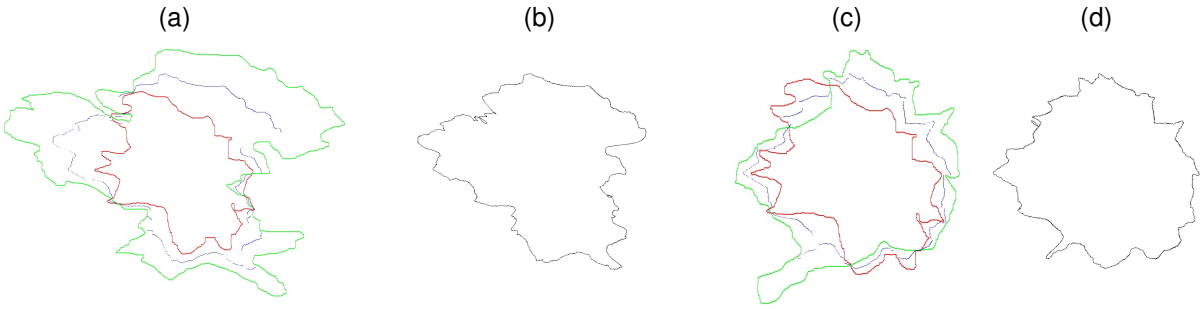


Fonte: O autor (2023).

Dois exemplos da fusão de contornos de nódulos malignos utilizando o Algoritmo 1 podem ser vistos na Figura 47. O resultado intermediário pode ser visto nas Figuras 47a e 47c nas quais os contornos em vermelho e verde são os contornos dos

dois nódulos benignos reais que serão combinados e o contorno em azul é o contorno resultante intermediário. Nas Figuras 47b e 47d pode-se ver os contornos finais.

Figura 47: Fusão do contornos de nódulos malignos utilizando o algoritmo 1. (a) e (c) Resultado intermediário; (b) e (d) Resultado final. Contornos vermelhos e verdes são de nódulos reais e contornos em azul são os contornos gerados.



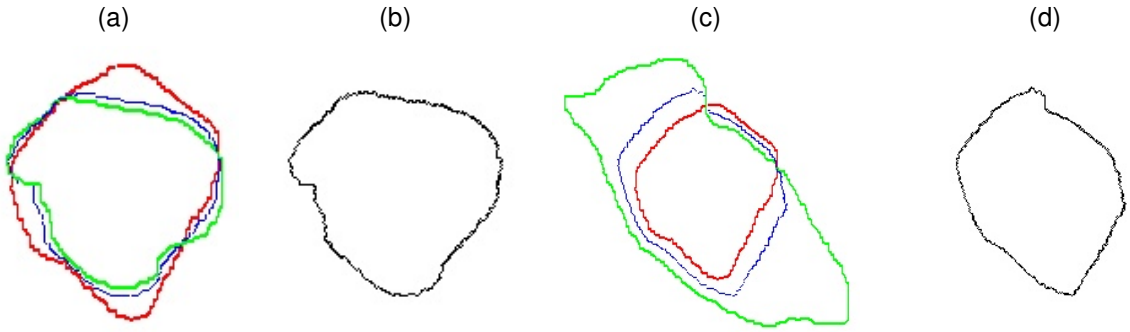
Fonte: O autor (2023).

Algoritmo 2: Sem rotacionar o contorno e utilizar um limiar.

O algoritmo 2 é similar ao algoritmo 1, mas se a distância em um determinado ponto entre os dois contornos for maior que a distância média entre os contornos (limiar), então a distância média será utilizada naquele determinado ponto.

As Figuras 48 e 49 exibem os contornos gerados utilizando o Algoritmo 2 para os nódulos benignos e malignos, respectivamente. A fim de possibilitar a comparação entre os resultados produzidos pelos algoritmos, foram utilizados os mesmos contornos originais apresentados nas Figuras 46 e 47.

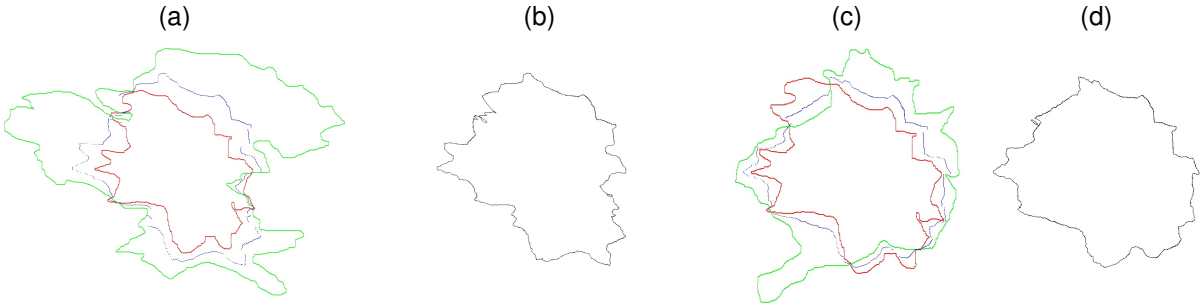
Figura 48: Fusão do contornos de nódulos benignos utilizando o algoritmo 2. (a) e (c) Resultado intermediário; (b) e (d) Resultado final. Contornos vermelhos e verdes são de nódulos reais e contornos em azul são os contornos gerados.



Fonte: O autor (2023).

Algoritmo 3: Rotacionar o contorno, utilizar o contorno resultante com menor distância média entre as bordas (considerando todas as rotações) e utilizar o ponto médio

Figura 49: Fusão do contornos de nódulos malignos utilizando o algoritmo 2. (a) e (c) Resultado intermediário; (b) e (d) Resultado final. Contornos vermelhos e verdes são de nódulos reais e contornos em azul são os contornos gerados.



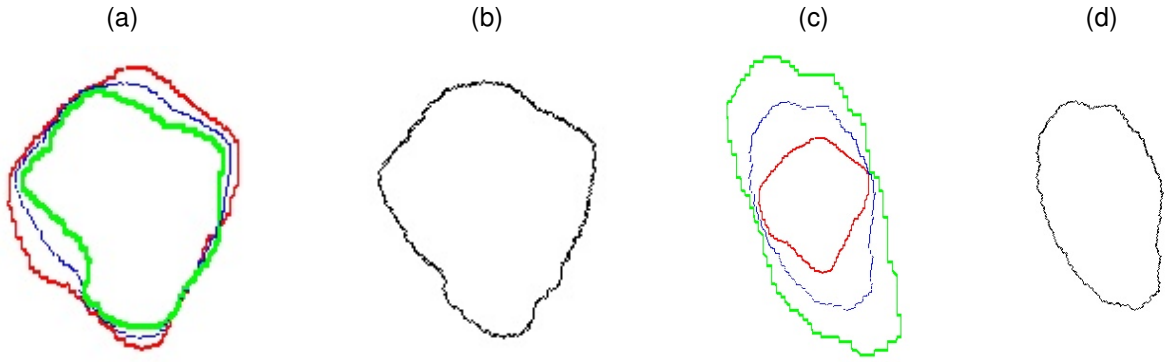
Fonte: O autor (2023).

da distância entre as bordas.

Neste algoritmo um dos contornos é mantido fixo (o contorno do nódulo que foi escolhido para ser combinado com todos os outros contornos de sua classe) e o outro contorno é rotacionado de 30 em 30 graus. Em cada rotação, a distância média é calculada e é escolhido o contorno rotacionado que apresentou a menor distância média entre as bordas dos dois nódulos. Uma vez escolhida a rotação adequada, faz-se o mesmo processo utilizado no Algoritmo 1.

As Figuras 50 e 51 exibem os resultados obtidos utilizando o Algoritmo 3.

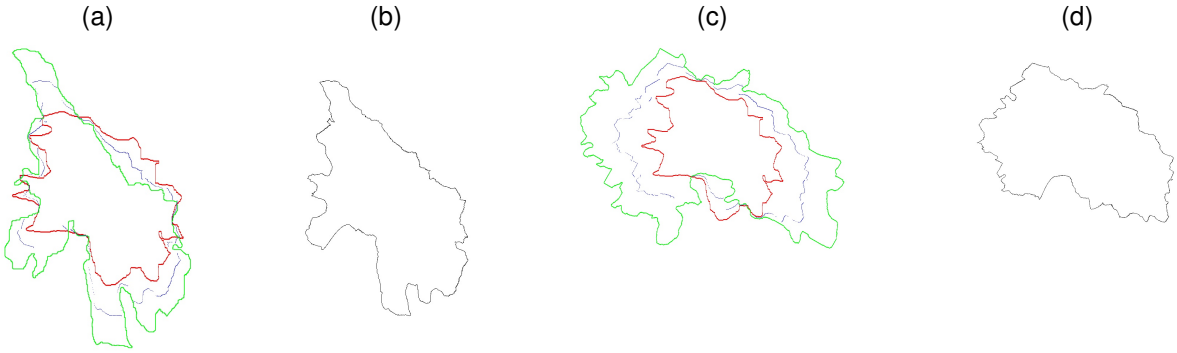
Figura 50: Fusão do contornos de nódulos benignos utilizando o algoritmo 3. (a) e (c) Resultado intermediário; (b) e (d) Resultado final. Contornos vermelhos e verdes são de nódulos reais e contornos em azul são os contornos gerados.



Fonte: O autor (2023).

Algoritmo 4: Rotacionar o contorno, utilizar o contorno resultante com menor variância entre as bordas (considerando todas as rotações) e utilizar o ponto médio da distância entre as bordas. Este algoritmo é bastante similar ao algoritmo 3, ressalvando-se que a escolha do contorno resultante considera a variância em vez da distância

Figura 51: Fusão do contornos de nódulos malignos utilizando o algoritmo 3. (a) e (c) Resultado intermediário; (b) e (d) Resultado final. Contornos vermelhos e verdes são de nódulos reais e contornos em azul são os contornos gerados.

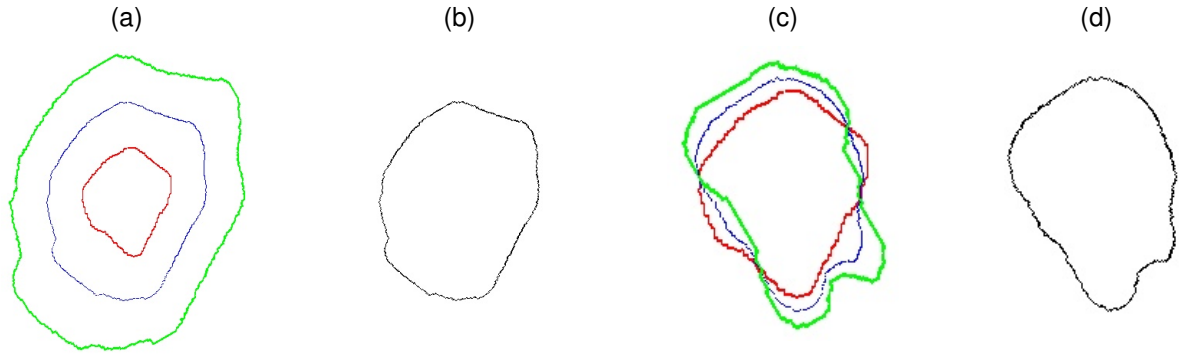


Fonte: O autor (2023).

média.

As Figuras 52 e 53 exibem os resultados obtidos utilizando o Algoritmo 4.

Figura 52: Fusão do contornos de nódulos benignos utilizando o algoritmo 4. (a) e (c) Resultado intermediário; (b) e (d) Resultado final. Contornos vermelhos e verdes são de nódulos reais e contornos em azul são os contornos gerados.



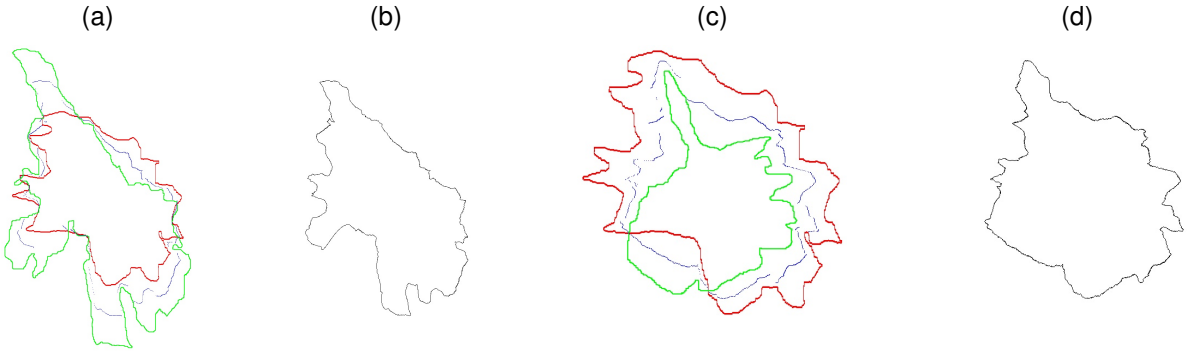
Fonte: O autor (2023).

Algoritmo 5: Rotacionar o contorno, utilizar o contorno resultante com a menor distância média entre as bordas (considerando todas as rotações) e utilizar o limiar. Este algoritmo é similar ao Algoritmo 3 na forma de escolher o contorno com a rotação adequada, mas faz uso do limiar (semelhante ao Algoritmo 2) para traçar o contorno final do nódulo gerado.

As Figuras 54 e 55 exibem os resultados obtidos utilizando o Algoritmo 5.

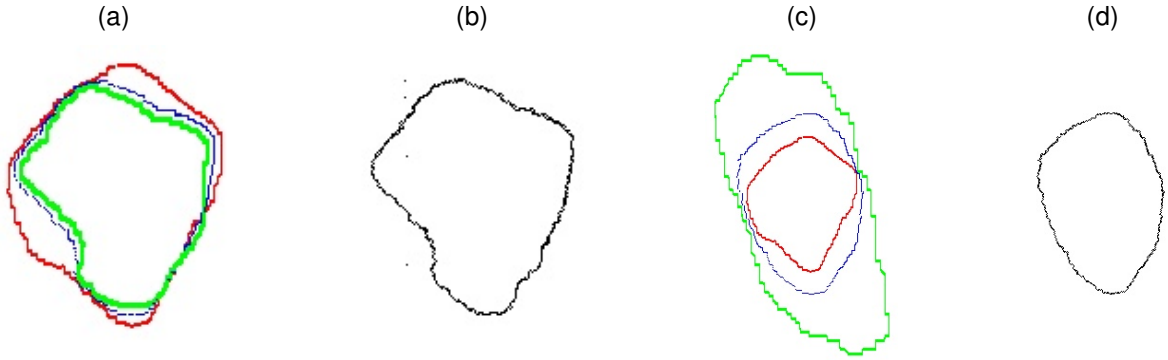
Algoritmo 6: Rotacionar o contorno, utilizar o contorno resultante com menor variância entre as bordas (considerando todas as rotações) e utilizar o limiar. Este algoritmo é bastante similar ao Algoritmo 5, mas escolhe o nódulo rotacionado com base

Figura 53: Fusão do contornos de nódulos malignos utilizando o algoritmo 4. (a) e (c) Resultado intermediário; (b) e (d) Resultado final. Contornos vermelhos e verdes são de nódulos reais e contornos em azul são os contornos gerados.



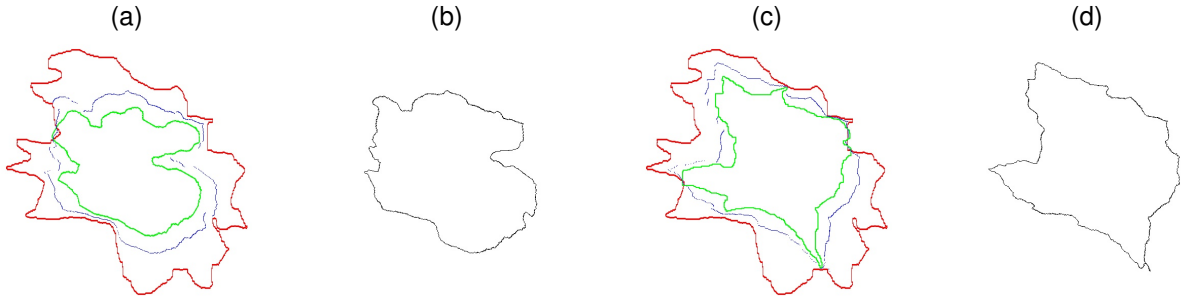
Fonte: O autor (2023).

Figura 54: Fusão do contornos de nódulos benignos utilizando o algoritmo 5. (a) e (c) Resultado intermediário; (b) e (d) Resultado final. Contornos vermelhos e verdes são de nódulos reais e contornos em azul são os contornos gerados.



Fonte: O autor (2023).

Figura 55: Fusão do contornos de nódulos malignos utilizando o algoritmo 5. (a) e (c) Resultado intermediário; (b) e (d) Resultado final. Contornos vermelhos e verdes são de nódulos reais e contornos em azul são os contornos gerados.

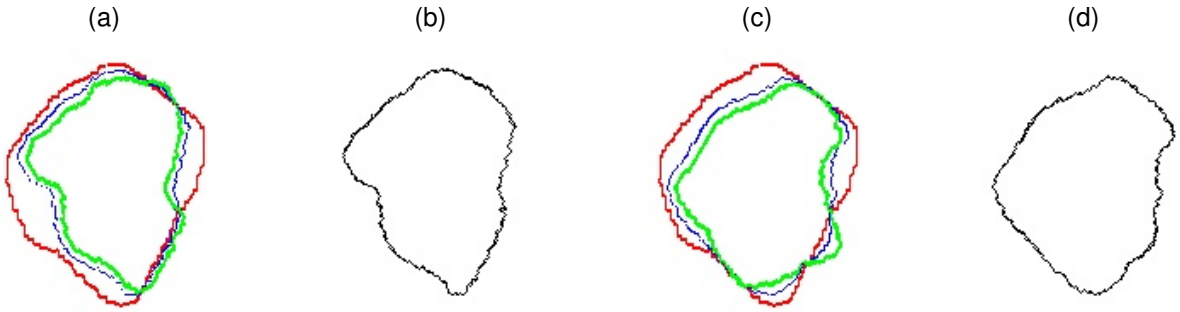


Fonte: O autor (2023).

na menor variância entre as bordas.

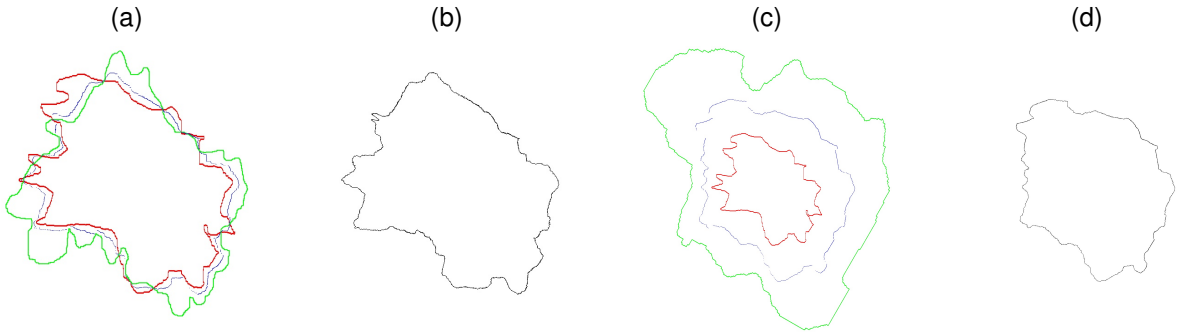
As Figuras 56 e 57 exibem os resultados obtidos utilizando o Algoritmo 6.

Figura 56: Fusão do contornos de nódulos benignos utilizando o algoritmo 6. (a) e (c) Resultado intermediário; (b) e (d) Resultado final. Contornos vermelhos e verdes são de nódulos reais e contornos em azul são os contornos gerados.



Fonte: O autor (2023).

Figura 57: Fusão do contornos de nódulos malignos utilizando o algoritmo 6. (a) e (c) Resultado intermediário; (b) e (d) Resultado final. Contornos vermelhos e verdes são de nódulos reais e contornos em azul são os contornos gerados.



Fonte: O autor (2023).

Os resultados da utilização das gramáticas para classificar os contornos gerados mediante fusão de contornos dos nódulos reais são exibidos na Seção 8.1.3.

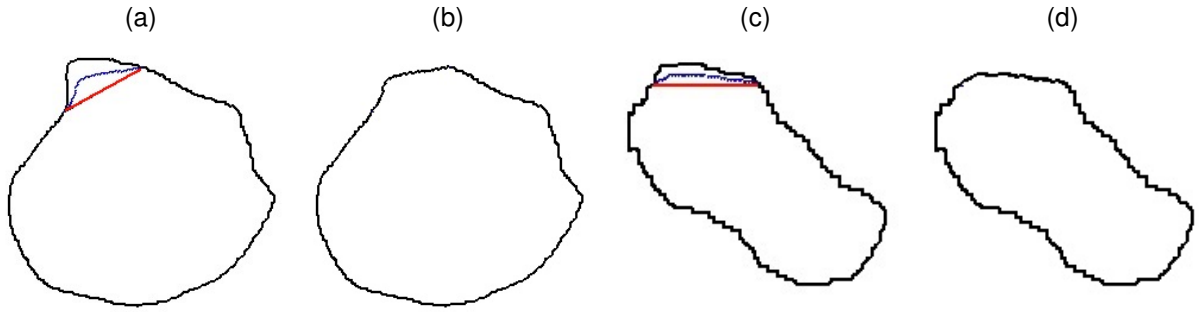
8.1.2 Alteração de contornos reais para geração de novos contornos sintéticos

Outra técnica utilizada para geração dos contornos dos nódulos sintéticos foi alterar apenas 1/8 da borda de um nódulo real existente, conforme explicado na Seção 6.2.Para esta técnica foram testados dois Algoritmos, conforme apresentado a seguir.

Algoritmo 1: Nesta abordagem, cada nódulo foi dividido em 8 partes octantes, sendo que sempre a borda do primeiro octante foi alterada.

Nas Figuras 58a e 58c estão representados em vermelho dois segmentos de reta nos primeiros octantes de cada imagem. As novas bordas (em azul) foram traçadas de forma equidistantes dos segmentos de retas e das bordas originais dos nódulos. Os resultados finais podem ser vistos nas Figuras 58b e 58d.

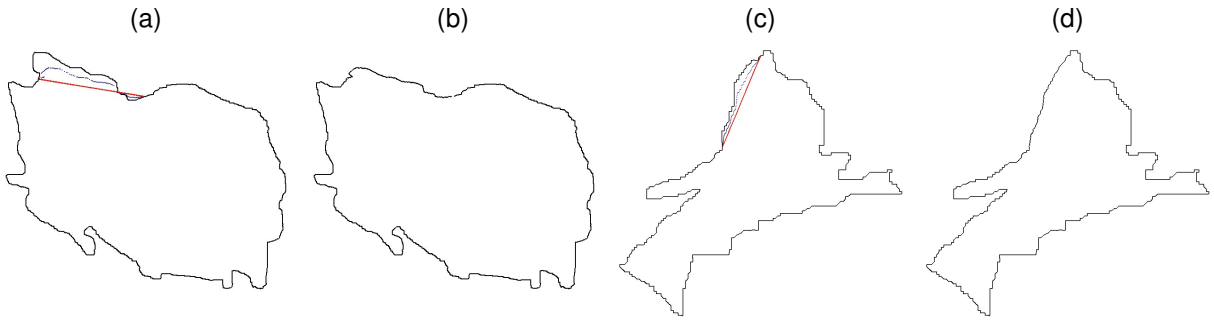
Figura 58: Alteração do contorno de nódulos benignos utilizando o algoritmo 1. (a) e (c) Resultado intermediário. (b) e (d) Resultado final. Segmento de reta suporte para alteração do contorno em vermelho e nova porção do contorno em azul.



Fonte: O autor (2023).

Similar à Figura 58, a Figura 59 exibe esta abordagem quando aplicada a dois contornos de nódulos malignos.

Figura 59: Alteração do contorno de nódulos malignos utilizando o algoritmo 1. (a) e (c) Resultado intermediário. (b) e (d) Resultado final. Segmento de reta suporte para alteração do contorno em vermelho e nova porção do contorno em azul.

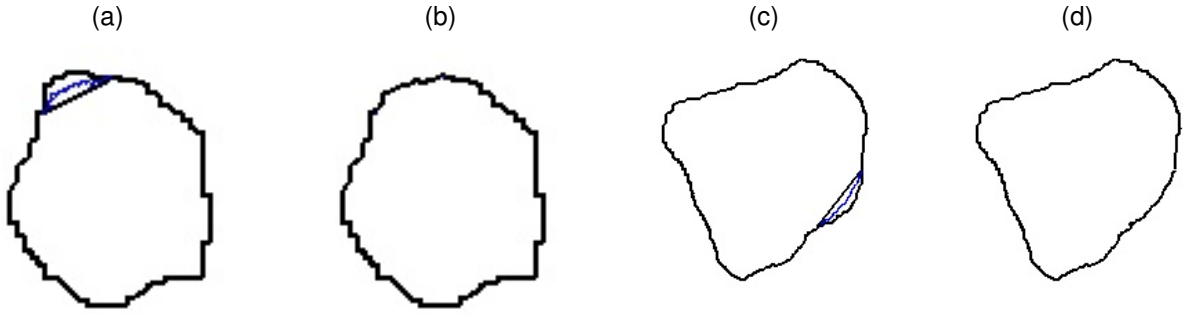


Fonte: O autor (2023).

Algoritmo 2: este algoritmo é similar ao *algoritmo 1*, mas a borda do octante do nódulo a ser alterada foi escolhida de forma aleatória. As Figuras 60 e 61 exibem os resultados obtidos utilizando o Algoritmo 2.

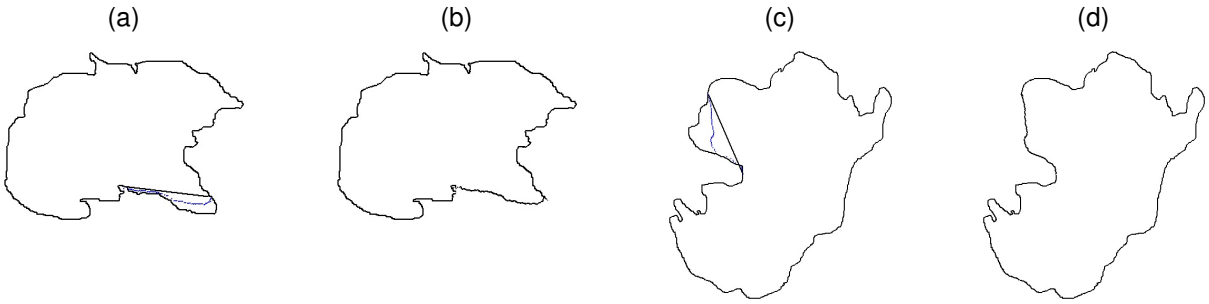
Os resultados da utilização das gramáticas para classificar os contornos gerados

Figura 60: Alteração do contorno de nódulos benignos utilizando o algoritmo 2. (a) e (c) Resultado intermediário. (b) e (d) Resultado final. Segmento de reta suporte para alteração do contorno em preto e nova porção do contorno em azul.



Fonte: O autor (2023).

Figura 61: Alteração do contorno de nódulos malignos utilizando o algoritmo 2. (a) e (c) Resultado intermediário. (b) e (d) Resultado final. Segmento de reta suporte para alteração do contorno em preto e nova porção do contorno em azul.



Fonte: O autor (2023).

por meio da alteração dos contornos dos nódulos são exibidos na Seção 8.1.3.

8.1.3 Resultados da aplicação das gramáticas aos contornos gerados

As Tabelas 12 e 13 exibem, respectivamente, o resultado da classificação das imagens sintéticas de nódulos benignos e malignos que foram criados utilizando as abordagens e algoritmos descritos nas seções 8.1.1 e 8.1.2.

No que diz respeito à geração de contornos sintéticos benignos, pode-se perceber analisando a Tabela 12 que o Algoritmo 2 da abordagem de fusão de contornos obteve a maior quantidade de nódulos benignos classificados corretamente (57%) bem como a menor taxa de contornos inválidos (39%). Para os contornos malignos (Tabela 13), a maior taxa de acerto (49%) e a menor taxa de contornos inválidos gerados (50%) foi

Tabela 12: Resultados da geração de contornos sintéticos utilizando contornos de nódulos benignos. A coluna **Não classificado** contém a quantidade de contornos que não foram reconhecidos nem pela gramática dos nódulos benignos e nem pela gramática dos nódulos malignos. Os melhores resultados são apresentados em negrito.

Abordagem	Algoritmo	Benigno	Maligno	Não classificado
Fusão de contorno	1	55 (53%)	2 (2%)	46 (45%)
Fusão de contorno	**2**	**59 (57%)**	**4 (4%)**	**40** (39%)
Fusão de contorno	3	48 (47%)	2 (2%)	53 (51%)
Fusão de contorno	4	46 (45%)	1 (1%)	56 (54%)
Fusão de contorno	5	52 (50%)	3 (3%)	48 (47%)
Fusão de contorno	6	49 (48%)	2 (2%)	52 (50%)
Alteração de contorno	1	43 (41%)	9 (9%)	52 (50%)
Alteração de contorno	2	41 (39%)	14 (13%)	49 (48%)

Fonte: O autor (2023)

Tabela 13: Resultados da geração de contornos sintéticos utilizando contornos de nódulos malignos. A coluna **Não classificado** contém a quantidade de contornos que não foram reconhecidos nem pela gramática dos nódulos benignos e nem pela gramática dos nódulos malignos. Os melhores resultados são apresentados em negrito.

Abordagem	Algoritmo	Benigno	Maligno	Não classificado
Fusão de contorno	**1**	**1 (1%)**	**46 (49%)**	**47 (50%)**
Fusão de contorno	2	4 (4%)	39 (42%)	51 (54%)
Fusão de contorno	3	5 (5%)	43 (46%)	46 (49%)
Fusão de contorno	4	4 (4)	39 (42%)	51 (54%)
Fusão de contorno	5	1 (1%)	36 (38%)	57 (61%)
Fusão de contorno	6	3 (3%)	39 (42%)	52 (55%)
Alteração de contorno	1	1 (1%)	36 (38%)	58 (61)
Alteração de contorno	2	2 (2%)	39 (41%)	54 (57%)

Fonte: O autor (2023)

obtida pela abordagem de fusão quando o Algoritmo 1 foi empregado.

É importante ressaltar a quantidade de contornos inválidos gerados pelas abordagens e algoritmos propostos, o que pode indicar que é necessário um aprimoramento dos algoritmos empregados. Apesar da quantidade considerável de contornos não reconhecidos, verifica-se que houve uma porcentagem também considerável de nódulos reconhecidos pelas gramáticas, principalmente de contornos benignos. Esta evidência sugere que a combinação entre contornos pode ser uma abordagem inicial para gerar contornos sintéticos de nódulos, podendo ser aprimorada por técnicas mais sofisticadas. Por exemplo, o resultado do reconhecimento, com suas respectivas probabilidades, poderia ser utilizado para realimentar os algoritmos e melhorar os contornos gerados. Enfim, é possível afirmar que as abordagens e algoritmos desen-

volvidos podem constituir o primeiro passo para que imagens sintéticas de nódulos sejam geradas e contribuam tanto para compor *datasets* para teste de sistemas de auxílio ao diagnóstico quando para uso como material didático.

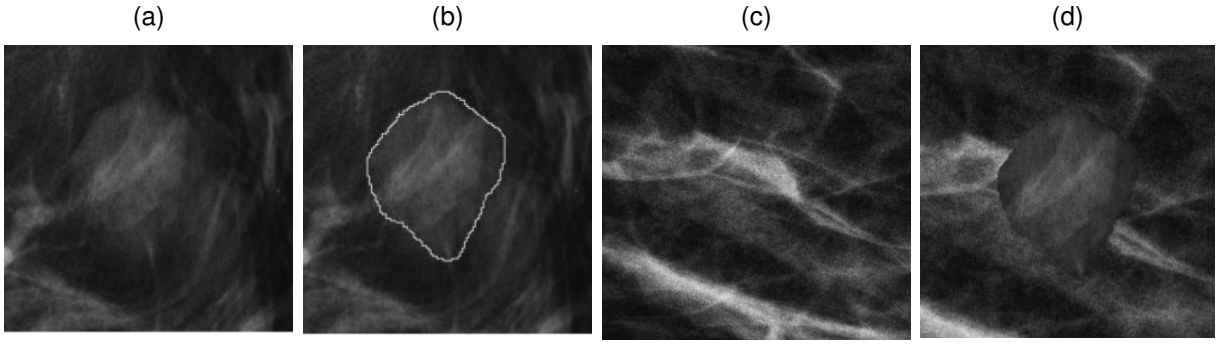
8.2 Transferência de um nódulo real para um novo mamograma

O desenvolvimento desta atividade utilizou o conceito de janelas deslizantes para escolha da nova região para o nódulo com base na diferença entre o valor da média de nível de cinza do local escolhido e o valor da média de nível de cinza do nódulo, conforme explicado na Seção 6.3.

O resultado desta abordagem pode ser conferido nas Figuras 62 e 63. Nas Figuras 62a a 62d são apresentados o nódulo benigno real no local original, o mesmo nódulo demarcado, a nova região de interesse encontrada e o resultado final, respectivamente. Um exemplo similar pode ser visto nas Figuras 63a- 63d, em que se encontra um nódulo maligno real no local original, o mesmo nódulo com o contorno das bordas, a nova região de interesse o resultado final, respectivamente.

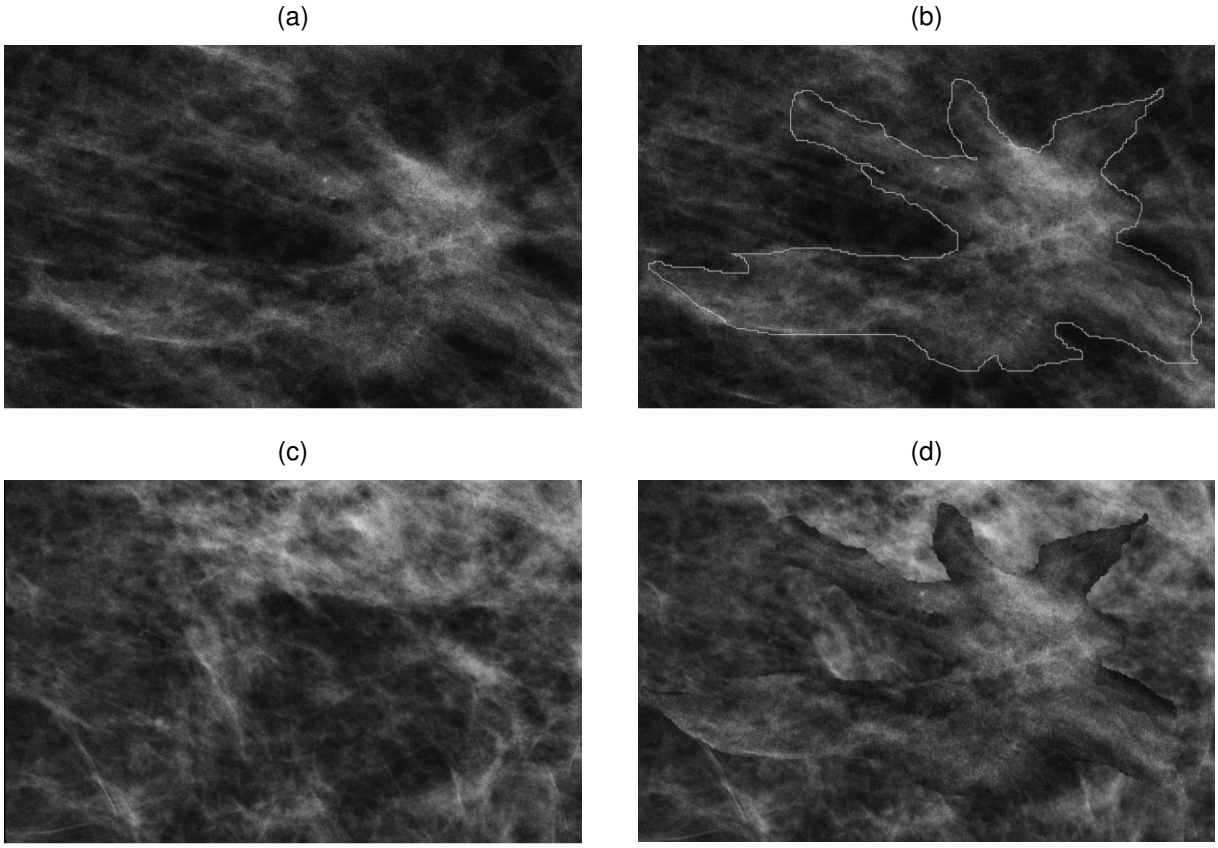
Analisando as Figuras 62d e 63d é relativamente fácil perceber, mesmo para pessoas que não tenham treinamento médico para detecção e classificação de nódulos, que os nódulos foram inseridos de forma artificial naquele local. Isso ocorre, muitas vezes, devido à diferença de estruturas da mama presentes em cada um dos mamogramas, bem como pela diferença da textura do nódulo e da região na qual ele foi inserido.

Figura 62: Transferência de um nódulo benigno real para outro mamograma. (a) Nódulo benigno em sua posição original; (b) contorno do nódulo demarcado. (c) região escolhida para o nódulo. (d) região escolhida após inclusão do nódulo.



Fonte: O autor (2023).

Figura 63: Transferência de um nódulo maligno real para outro mamograma. (a) Nódulo maligno em sua posição original; (b) contorno do nódulo demarcado. (c) região escolhida para o nódulo. (d) região escolhida após inclusão do nódulo.

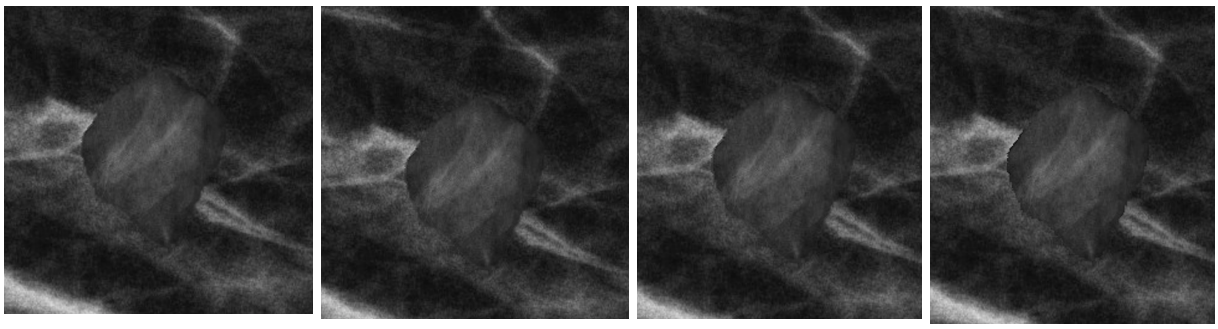


Fonte: O autor (2023).

Para que os nódulos gerados pareçam mais reais após serem inseridos nos mamogramas foi aplicada a técnica de suavização das bordas conforme descrito na Seção 6.3. Exemplos dos resultados alcançados são exibidos na Figura 64. É possível observar que mesmo aplicando a suavização das bordas considerando máscaras de diferentes tamanhos (3x3, 5x5 e 7x7), as imagens resultantes são visualmente similares à imagem que foi utilizada como entrada para o processo (imagem com o nódulo transferido - Figura 62d), sendo que as diferenças nas bordas são observáveis apenas mediante ampliação (*zoom*) das mesmas. A fim de tentar minimizar a artificialidade do resultado, foi conduzida uma investigação sobre o uso de texturas, conforme apresentado a seguir.

Figura 64: Suavização das bordas de um nódulo benigno transferido para outro mamograma. (a) valor médio dos pixels de uma máscara 3x3; (b) valor médio dos pixels de uma máscara 5x5. (c) valor médio dos pixels de uma máscara 7x7. (d) valor da mediana dos pixels de uma máscara 7x7.

(a) (b) (c) (d)



Fonte: O autor (2023).

8.3 Aplicação de textura para as imagens sintéticas de nódulos

Esta seção descreve os resultados obtidos com relação à geração de textura para o contorno de um nódulo sintético (ou o contorno de um nódulo real) conforme descrito na Seção 6.4.

8.3.1 Utilização da média dos valores dos pixels de nódulos conhecidos

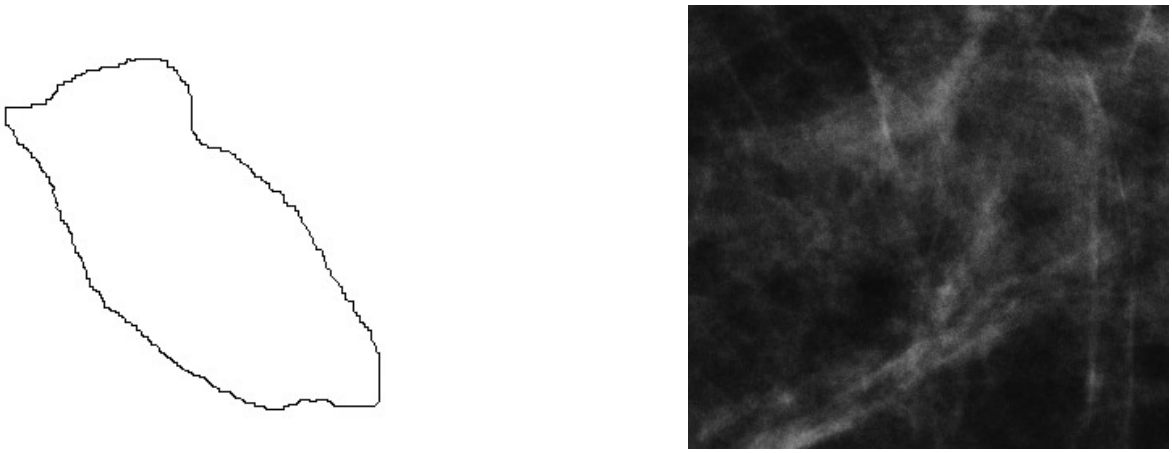
Esta abordagem fez uso dos valores médios dos pixels (níveis de cinza) dos nódulos reais (valor médio dos pixels internos ao nódulo e valor médio dos pixels externos ao nódulo (ROI)) para calcular os valores dos pixels de um novo nódulo, conforme detalhado na Seção 6.4 e na fórmula 6.1. Para os testes com esta abordagem foram utilizados o contorno exibido na Figura 65a e uma região de interesse obtida de um mamograma na qual não existisse um nódulo (Figura 65b).

No primeiro exemplo, a Figura 66a foi utilizada para fazer o cálculo das médias dos valores dos pixels (interna e externa ao nódulo) e o resultado da abordagem pode ser visto na Figura 66b.

Utilizando a mesma abordagem, agora com uma imagem de um outro nódulo escolhida de forma aleatória para o cálculo dos valores médios dos pixels (Figura 67a), mas utilizando o mesmo contorno (Figura 65a) e a mesma região de interesse (Fi-

Figura 65: Contorno de nódulo e nova região de interesse. (a) Contorno de nódulo utilizado; (b) Região de interesse utilizada.

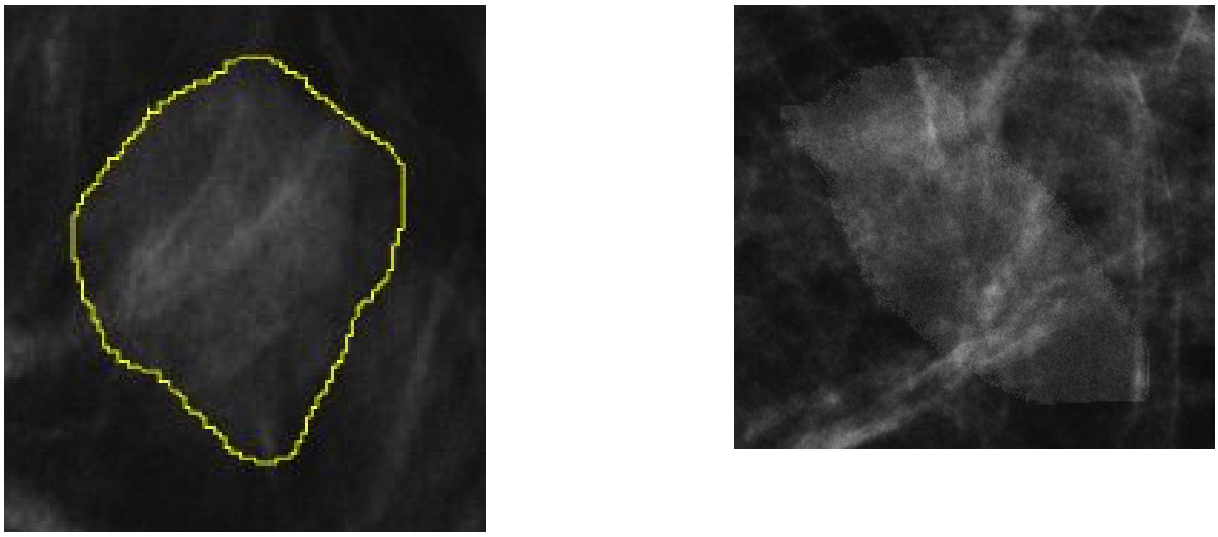
(a) (b)



Fonte: O autor (2023).

Figura 66: Nódulo utilizado para cálculo do nível de cinza e resultado final (Exemplo 1). (a) Nódulo utilizado para o cálculo dos valores médios dos pixels (nível de cinza) internos e externos; (b) Resultado obtido.

(a) (b)



Fonte: O autor (2023).

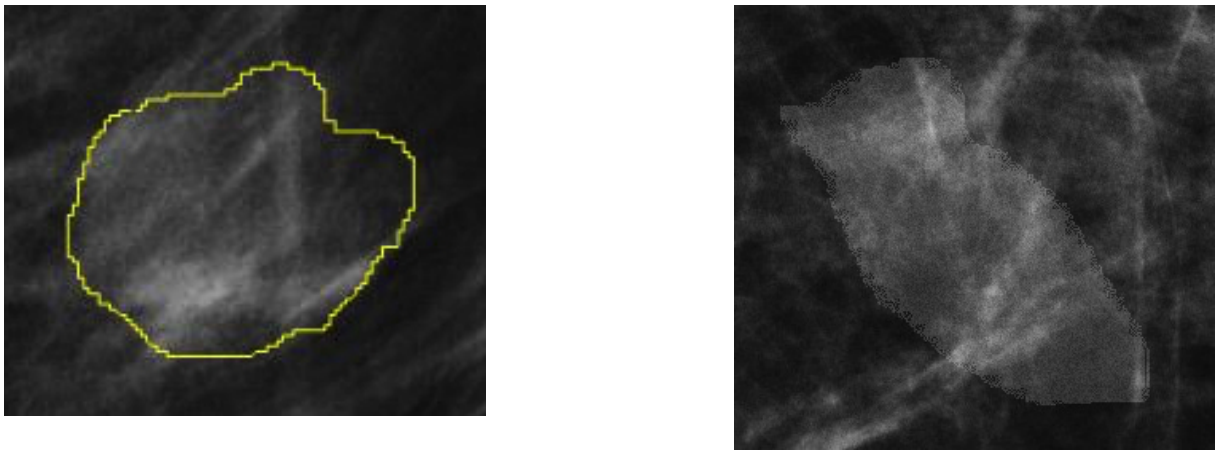
gura 65b) foi obtido o resultado que pode ser conferido na Figura 67b.

É importante mencionar que os nódulos estão demarcados com os contornos na cor amarela apenas para facilitar a visualização de suas bordas (Figuras 66a e 67a), mas para o cálculo das médias foram utilizados os valores reais dos pixels. Nos resultados obtidos optou-se por não demarcar as bordas dos contorno com a cor amarela

para ilustrar de forma mais real as imagens sintéticas geradas (Figuras 66b e 67b).

Figura 67: Nódulo utilizado para cálculo do nível de cinza e resultado final (Exemplo 2). (a) Nódulo utilizado para o cálculo dos valores médios dos pixels (nível de cinza) internos e externos; (b) Resultado obtido.

(a) (b)



Fonte: O autor (2023).

Assim como na abordagem de transferência de um nódulo real para um mamograma, e por motivos semelhantes, a abordagem apresentada nesta seção demonstrou limitações no que diz respeito a gerar uma textura para um nódulo a ser inserido em um local onde inicialmente não existe nódulo. Assim, uma nova abordagem foi empregada utilizando regressão linear, conforme descrito a seguir.

8.3.2 Utilização de regressão linear para inferência do valor do pixel do nódulo

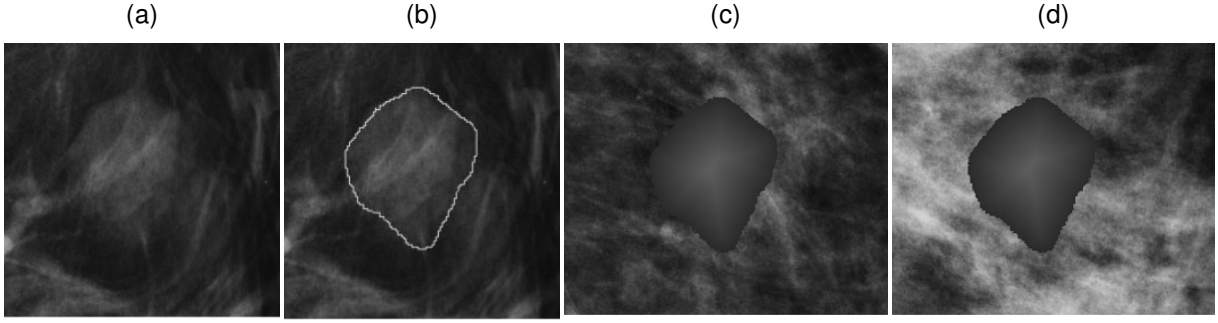
Nesta seção é apresentada uma abordagem que faz uso de regressão linear para inferir o nível de cinza de um pixel que compõe o nódulo. Como entrada para o modelo de regressão foram utilizadas uma característica representando a distância do pixel ao centro de massa de um novo nódulo e uma característica representando a menor distância do pixel à borda do nódulo, sendo o valor do pixel utilizado como alvo durante o treinamento do modelo.

Nos testes com esta abordagem, foram geradas texturas novas para contornos de nódulo já existentes (as texturas originais foram retiradas) e para contornos gerados mediante utilização das abordagens descritas na Seção 8.1.1. Como exemplo, o treinamento do modelo foi realizado utilizando o nódulo exibido na Figura 68a; em seguida utilizou-se o mesmo contorno do nódulo (sem a textura) e procedeu-se o seu

reposicionamento em uma parte do mamograma na qual não havia um nódulo existente.

A Figura 68 exibe o resultado desta abordagem utilizada para gerar a textura de um nódulo benigno. Nas Figuras 68a a 68d estão o nódulo original sem a demarcação do contorno, o nódulo original com o contorno demarcado, e dois resultados gerados, respectivamente.

Figura 68: Resultados da geração de textura utilizando regressão linear. (a) Nódulo original; (b) Nódulo original demarcado; (c) Resultado do experimento 1; (d) Resultado do experimento 2.



Fonte: O autor (2023).

Assim como na abordagem de transferência de um nódulo real e na abordagem de utilização da média dos valores dos pixels de nódulos, a abordagem descrita nesta seção também não foi capaz de gerar uma textura para o nódulo semelhante à textura de um nódulo real, evidenciando a necessidade da utilização de técnicas mais robustas de processamento de imagens ou de aprendizado de máquina.

8.4 Discussões

Este capítulo apresentou uma abordagem inicial para geração de imagens sintéticas de nódulos encontrados em mamogramas. De uma forma geral, para a geração de imagens sintéticas são utilizadas técnicas de processamento de imagens e, neste projeto, gostaríamos de verificar se a utilização de gramáticas, principalmente na seleção de contornos sintéticos, poderia ser útil nesta tarefa, pois os nódulos gerados poderiam ser utilizados tanto para fins didáticos quanto para auxiliar nos testes de sistemas CAD. Conforme pôde ser visto, a geração de nódulos neste trabalho foi dividida em três etapas: i) geração dos contornos dos nódulos; ii) inserção de um nódulo sintético em um mamograma real e iii) geração da textura dos nódulos.

A principal dificuldade no que diz respeito à geração dos contornos dos nódulos está relacionada à criação de um contorno que se pareça com os contornos dos nódulos reais. Neste trabalho foram utilizadas oito abordagens para a geração dos contornos (seis algoritmos de fusão de contornos e dois algoritmos de alteração de contornos). Para selecionar os contornos mais semelhantes aos contornos reais foram utilizadas as gramáticas criadas para a classificação dos nódulos reais em benignos ou malignos. O melhor resultado em termos de aproveitamento dos contornos benignos gerados foi de 57% (Algoritmo 2). Ainda assim, a gramática utilizada no reconhecimento das imagens sintéticas de nódulos não foi capaz de classificar 39% dos contornos gerados (Tabela 12). Quando são avaliados os nódulos malignos gerados, o melhor aproveitamento dos contornos foi obtido pela Algoritmo 1 (49%), mas nesse cenário 50% dos contornos gerados foram descartados (Tabela 13). Esses resultados evidenciam a dificuldade em gerar contornos de nódulos utilizando as duas abordagens propostas que sejam reconhecidos corretamente pelas gramáticas. Quando comparada a taxa de classificação correta entre os nódulos benignos e malignos gerados (Tabelas 12 e 13), pode-se perceber que as gramáticas tem uma taxa de acerto maior na classificação dos nódulos benignos. Este fato pode estar ocorrendo devido à complexidade dos contornos dos nódulos malignos, pois eles tendem a ser mais espiculados que os benignos. Apesar da quantidade considerável de contornos sintéticos não reconhecidos pelas gramáticas, a ideia de usar gramáticas para esta finalidade é inovadora e merece exploração mais aprofundada. Em especial, pode ser interessante investigar se os próprios resultados da classificação podem ser utilizados para retroalimentar os algoritmos e propor alterações nos contornos.

A segunda dificuldade encontrada está relacionada à geração de textura dos nódulos, que também está relacionada à dificuldade em inserir um nódulo sintético em um mamograma real. Esta dificuldade ocorre devido à complexidade das texturas dos nódulos e das imagens mamográficas como um todo. Para gerar a textura de um nódulo sintético não se pode levar em consideração apenas o nódulo em si, mas também o local no qual este nódulo será inserido e o seu entorno, pois a sua localização irá influenciar de forma substancial em sua textura. Na parte interna do nódulo em um mamograma é possível detectar estruturas presentes na mama, como tecido fibroglandular, gordura e ductos lactíferos, sendo que gerar texturas que simulem essas estruturas não é uma tarefa trivial. Como pôde ser visto, a dificuldade para gerar texturas semelhantes às texturas reais não pôde ser superada utilizando técnicas mais básicas de processamento de imagens em conjunto com gramáticas. No entanto, os

resultados apresentados neste capítulo indicam que essa junção de técnicas pode constituir um embrião para o desenvolvimento de técnicas mais avançadas, as quais estavam fora do escopo deste projeto.

8.5 Considerações

Idealmente as imagens geradas de forma sintética deveriam ser fornecidas a um corpo médico para avaliação. Esperava-se que os profissionais da área de saúde fizessem a classificação de um conjunto de imagens com nódulos reais e sintéticos. Esta classificação deveria inicialmente dizer se o nódulo que estava sendo avaliado era real ou sintético e se era benigno ou maligno. Entretanto, com base nos resultados obtidos, pode-se verificar que os métodos utilizados para geração de imagens sintéticas de nódulos não foram suficientes para a produção de nódulos semelhantes aos nódulos reais. Por isso, o aprimoramento das técnicas e a avaliação das imagens geradas por profissionais da saúde são trabalhos futuros que merecem exploração.

No próximo capítulo são apresentadas as considerações finais deste projeto de pesquisa.

PARTE IV

CONCLUSÕES

9 CONCLUSÃO

Este trabalho teve como objetivo principal a utilização de métodos sintáticos para fazer a classificação de imagens de nódulos reais encontrados em mamogramas e para validar a geração de nódulos sintéticos.

Por meio das revisões de literatura conduzidas, foi possível identificar que métodos sintáticos são raramente empregados para resolver os dois problemas mencionados (classificação dos nódulos e validação de nódulos sintéticos). Por este motivo, a proposta apresentada neste trabalho constituiu uma contribuição inédita para esta área de pesquisa.

A primeira hipótese investigada foi de que era possível fazer uso de métodos sintáticos para realizar a classificação de nódulos considerando as classes Benigno e Maligno obtendo resultados similares aos do estado da arte. Esta hipótese foi confirmada, conforme detalhado no Capítulo 7. Neste sentido, foi possível verificar que o método proposto atingiu acurácias variando de 96% a 100% nos melhores casos dependendo dos modelos utilizados, sendo esses resultados semelhantes aos resultados alcançados por outros modelos de aprendizado de máquina que representam o estado da arte nesta área de pesquisa, por exemplo, métodos de *deep learning*.

A segunda hipótese estudada foi de que era possível utilizar métodos sintáticos para validar imagens sintéticas de nódulos. Esta hipótese foi parcialmente confirmada. As gramáticas utilizadas para classificar nódulos foram empregadas com sucesso para classificar os contornos de nódulos nas imagens sintéticas gerados a partir de técnicas de processamento de imagens (fusão e alteração das bordas de nódulos reais). No entanto, as técnicas de processamento de imagens utilizadas não se mostraram suficientes para a criação dos nódulos sintéticos similares a nódulos reais, conforme detalhado no Capítulo 8. Adicionalmente, não foi possível utilizar as gramáticas para contribuir para o aprimoramento das técnicas de processamentos de imagens. Neste sentido, foram encontrados vários desafios que não puderam ser transpassados para a criação de uma base de dados composta com imagens sintéticas de nódulos. En-

tretanto, a criação desta base é algo valioso para a comunidade, pois poderia ser utilizada por outros pesquisadores da área tanto em projetos de detecção, classificação e segmentação de nódulos, quanto para fins didáticos. Com relação à área de saúde, a principal contribuição residiria na utilização destas imagens para treinamento de novos médicos em tarefas de detecção e classificação de nódulos.

As principais contribuições deste trabalho para a área de Engenharia de Computação estão relacionadas à construção de um modelo gramatical robusto capaz de classificar os nódulos encontrados nos mamogramas mesmo utilizando um conjunto limitado de imagens com desempenho semelhante ou superior ao estado da arte. Além disso, foi criado um novo *dataset* com imagens fornecidas por pesquisadores do A. C. Camargo Cancer Center que pode ser utilizado em novas pesquisas relacionadas tanto à detecção, classificação, segmentação e geração de imagens.

Este trabalho também tem potencial de impacto social e econômico, pois o câncer de mama é uma doença responsável pela morte de várias pessoas todos os anos. Quanto mais precocemente for feito o diagnóstico, maiores são as chances de tratamento e de cura. Além disso, quanto mais preciso for o diagnóstico menos biópsias desnecessárias serão realizadas, reduzindo não apenas os gastos financeiros do procedimento, mas também o estresse e a dor física ao quais as mulheres são submetidas.

REFERÊNCIAS

AHO, A. V.; ULLMAN, J. D. The theory of parsing, translation, and compiling. Prentice-Hall, Inc., Englewood Cliffs, NJ, v. 1, 1972.

AL-NAJDAWI, N.; BILTAWI, M.; TEDMORI, S. Mammogram image visual enhancement, mass segmentation and classification. **Applied Soft Computing**, v. 35, p. 175–185, 2015.

AROQUIARAJ, I. L.; THANGAVEL, K. Mass classification method in mammogram using fuzzy k-nearest neighbour equality. **Computing Research Repository**, abs/1406.4770, 2014.

ARORA, R.; RAI, P. K.; RAMAN, B. Deep feature–based automatic classification of mammograms. **Medical & Biological Engineering & Computing**, v. 58, n. 6, p. 1199–1211, 2020. Disponível em: <https://doi.org/10.1007/s11517-020-02150-8>.

AZOUR, F.; BOUKERCHE, A. Design guidelines for mammogram-based computer-aided systems using deep learning techniques. **IEEE Access**, v. 10, p. 21701–21726, 2022.

BASSETT, L. W. Standardized reporting for mammography: BI-RADS^{TM}. **The Breast Journal**, v. 3, n. 5, p. 207–210, 1997. ISSN 1524-4741. Disponível em: <http://dx.doi.org/10.1111/j.1524-4741.1997.tb00172.x>.

BEHESHTI, S. M. A. et al. An efficient fractal method for detection and diagnosis of breast masses in mammograms. **Journal of Digital Imaging**, v. 27, n. 5, p. 661–669, 2014.

BENNDORF, M. et al. Development of an online, publicly accessible naive bayesian decision support tool for mammographic mass lesions based on the american college of radiology (ACR) BI-RADS lexicon. **European Radiology**, v. 25, n. 6, p. 1768–1775, 2015. ISSN 1432-1084. Disponível em: <http://dx.doi.org/10.1007/s00330-014-3570-6>.

BLIZNAKOVA, K. S.; PALLIKARAKIS, N. E. Quantitative evaluation of a mammographic software phantom generator. In: **2009 International Conference on Information Technology and Applications in Biomedicine, 9th.** [S.l.: s.n.], 2009. p. 1–4. ISSN 2168-2194.

BORNE, P. **Polygon simplification**. 2014. Disponível em: <https://www.mathworks.com/matlabcentral/fileexchange/45342-polygon-simplification>.

BRADSKI, G. The OpenCV Library. **Dr. Dobb's Journal of Software Tools**, 2000.

BROWNLEE, J. **Deep Learning for Computer Vision: Image Classification, Object Detection, and Face Recognition in Python**. Machine Learning Mastery, 2019. Disponível em: <https://books.google.com.br/books?id=DOamDwAAQBAJ>.

BRUCE, L. M.; KALLERGI, M.; MENDOZA, A. Wavelet scalar-energy features for recognition of mammographic mass shapes. In: Szu, H. H. (Ed.). **Wavelet Applications VI**. [S.l.: s.n.], 1999. v. 3723, p. 156–162.

CAULKIN, S.; ASTLEY, S. Generating realistic mass lesions in digital mammograms using statistical models. In: **British Machine Vision Conference. Proceedings**. [S.l.]: BMVA Press, 1999. p. 29.1–29.10. ISBN 1-901725-09-X. Doi:10.5244/C.13.29.

CHOI, J. S. et al. Comparison between two-dimensional synthetic mammography reconstructed from digital breast tomosynthesis and full-field digital mammography for the detection of T1 breast cancer. **European Radiology**, v. 26, n. 8, p. 2538–2546, Aug 2016. ISSN 1432-1084.

CHOI, J. S. et al. Comparison of synthetic and digital mammography with digital breast tomosynthesis or alone for the detection and classification of microcalcifications. **European Radiology**, Jun 2018. ISSN 1432-1084.

CHOMSKY, N. On certain formal properties of grammars. **Information and Control**, v. 2, n. 2, p. 137 – 167, 1959. ISSN 0019-9958. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0019995859903626>.

DELOGU, P. et al. Characterization of mammographic masses using a gradient-based segmentation algorithm and a neural classifier. **Computers in Biology and Medicine**, v. 37, n. 10, p. 1479–1491, 2007.

DOMÍNGUEZ, A. R.; NANDI, A. K. Toward breast cancer diagnosis based on automated segmentation of masses in mammograms. **Pattern Recognition**, Elsevier Science Inc., New York, NY, USA, v. 42, n. 6, p. 1138–1148, jun. 2009. ISSN 0031-3203.

DONG, M. et al. An efficient approach for automated mass segmentation and classification in mammograms. **Journal of Digital Imaging**, v. 28, n. 5, p. 613–625, 2015.

EARLEY, J. An efficient context-free parsing algorithm. **Communications of the ACM**, ACM, New York, NY, USA, v. 13, n. 2, p. 94–102, fev. 1970. ISSN 0001-0782.

ELTOUKHY, M. M.; FAYE, I.; SAMIR, B. B. A comparison of wavelet and curvelet for breast cancer diagnosis in digital mammogram. **Computers in Biology and Medicine**, v. 40, n. 4, p. 384–391, 2010.

FERREIRA, M.; SANTOS, C.; MONTEIRO, J. Texture cue based tracking system using wavelet transform and a fuzzy grammar. In: . [S.l.]: IEEE International Conference on Industrial Informatics, 2007. v. 1, p. 393–398.

FU, K. **Syntactic Pattern Recognition and Applications**. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1982.

GALARRETA-VALVERDE, M. A. et al. Three-dimensional synthetic blood vessel generation using stochastic L-systems. In: **Proceedings of SPIE**. [S.l.: s.n.], 2013. v. 8669.

GALARRETA-VALVERDE, M. A. et al. Characterization of vascular tree architecture using the tokunaga taxonomy. In: **Proceedings of SPIE**. [S.l.: s.n.], 2015. v. 9414, p. 9414 – 9414 – 6.

GEORGIOU, H. V. et al. Multi-scaled morphological features for the characterization of mammographic masses using statistical classification schemes. **Artificial Intelligence in Medicine**, v. 41, n. 1, p. 39–55, 2007.

GIGER, M. L.; CHAN, H.; BOONE, J. History and status of cad and quantitative image analysis: The role of medical physics and aapm. **Medical Physics**, v. 35, n. 12, p. 5799–5820, dec. 2008. ISSN 10.1118/1.3013555.

HADJIISKI, L. et al. Analysis of temporal changes of mammographic features: computer-aided classification of malignant and benign breast masses. **Medical Physics**, v. 28, n. 11, p. 2309 – 2317, 2001.

HAMDI, S.; ABDALLAH, A. B.; BEDOUI, M. Grammar-based image segmentation and automatic area estimation. In: **Electrotechnical Conference (MELECON), 2012 IEEE Mediterranean, 16th**. [S.l.: s.n.], 2012. p. 356 –359. ISSN 2158-8473.

HAMDI, S. et al. A new method of cardiographic image segmentation based on grammar. In: . [S.l.: s.n.], 2011. v. 8285, p. 8285 – 8285 – 7.

HAPFELMEIER, A.; HORSCH, A. Image feature evaluation in two new mammography cad prototypes. **International Journal of Computer Assisted Radiology and Surgery**, v. 6, n. 6, p. 721–735, 2011.

HEATH, M. et al. The digital database for screening mammography. In: **International Workshop on Digital Mammography. Proceedings of the 5th**. [S.l.: s.n.], 2000. (Proceedings of the 5th International Workshop on Digital Mammography).

HEIDARI, M. et al. Applying a random projection algorithm to optimize machine learning model for breast lesion classification. **IEEE Transactions on Biomedical Engineering**, v. 68, n. 9, p. 2764–2775, 2021.

HEIDARI, M. et al. A new case-based CAD scheme using a hierarchical SSIM feature extraction method to classify between malignant and benign cases. In: CHEN, P.-H.; DESERNO, T. M. (Ed.). **Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications**. SPIE, 2020. v. 11318, p. 309 – 315. Disponível em: <https://doi.org/10.1117/12.2549130>.

HIRAMA, R. et al. Evaluating the impact of polygonal representations on mass classification. In: **2020 International Conference on Systems, Signals and Image Processing (IWSSIP)**. [S.l.: s.n.], 2020. p. 75–80.

HOPCROFT, J. E.; ULLMAN, J. D. Introduction to automata theory, languages, and computation. Addison-Wesley Publishing Company, 1979.

HU, M.-K. Visual pattern recognition by moment invariants. **IRE Transactions on Information Theory**, v. 8, n. 2, p. 179–187, 1962.

HUO, Z. et al. Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness. **Academic Radiology**, v. 7, n. 12, p. 1077 – 1084, 2000. ISSN 1076-6332.

INCA. **Instituto Nacional do Câncer**. 2023. <http://www2.inca.gov.br/wps/wcm/connect/inca/portal/home>. Acessado: 18-02-2023.

JAVADI, S. M. T.; FAEZ, K. Finding suspicious masses of breast cancer in mammography images using particle swarm algorithm and its classification using fuzzy methods. In: **2012 International Conference on Computer Communication and Informatics**. [S.l.: s.n.], 2012. p. 1–5.

KANADAM, K. P.; CHEREDDY, S. R. Mammogram classification using sparse-ROI: A novel representation to arbitrary shaped masses. **Expert Systems with Applications**, v. 57, p. 204–213, 2016.

KANUNGO, T.; MAO, S. Stochastic language models for style-directed layout analysis of document images. **IEEE Transactions on Image Processing**, v. 12, n. 5, p. 583 – 596, may 2003. ISSN 1057-7149.

KELEs, A.; KELEs, A.; YAVUZ, U. Extracting fuzzy rules for the diagnosis of breast cancer. **Turkish Journal of Electrical Engineering and Computer Sciences**, v. 21, n. 5, p. 1495 – 1503, 2013.

KHAN, H. N. et al. Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. **IEEE Access**, v. 7, p. 165724–165733, 2019.

KHAN, S. et al. Optimized gabor features for mass classification in mammography. **Applied Soft Computing Journal**, v. 44, p. 267–280, 2016.

KHAN, S. et al. A comparison of different gabor feature extraction approaches for mass classification in mammography. **Multimedia Tools and Applications**, v. 76, n. 1, p. 33–57, 2017.

KIM, S. T.; KIM, D. H.; RO, Y. M. Generation of conspicuity-improved synthetic image from digital breast tomosynthesis. In: **2014 International Conference on Digital Signal Processing, 19th**. [S.l.: s.n.], 2014. p. 395–399. ISSN 1546-1874.

LARASATI, R. Explainable AI for breast cancer diagnosis: Application and user's understandability perception. In: **2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)**. [S.l.: s.n.], 2022. p. 1–6.

LI, H. et al. Breast masses in mammography classification with local contour features. **BioMedical Engineering OnLine**, BioMed Central, London, v. 16, p. 44, 2017.

LI, Z. et al. Prediction of cardiac arrhythmia using deterministic probabilistic finite-state automata. **Biomedical Signal Processing and Control**, v. 63, p. 102200, 2021. ISSN 1746-8094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1746809420303347>.

LIMA, S. M. L.; FILHO, A. G. S.; SANTOS, W. P. Detection and classification of masses in mammographic images in a multi-kernel approach. **Computer Methods and Programs in Biomedicine**, v. 134, p. 11–29, 2016.

LIU, Z. et al. Design and implementation of a line simplification algorithm for network measurement system. In: **2011 IEEE International Conference on Broadband Network and Multimedia Technology, 4th**. [S.l.: s.n.], 2011. p. 412–416. ISSN null.

LUO, P. et al. Hierarchical 3D perception from a single image. In: **2009 IEEE International Conference on Image Processing (ICIP), 16th**. [S.l.: s.n.], 2009. p. 4265–4268.

MALEBARY, S. J.; HASHMI, A. Automated breast mass classification system using deep learning and ensemble learning in digital mammogram. **IEEE Access**, v. 9, p. 55312–55328, 2021.

MARISCOTTI, G. et al. Comparison of synthetic mammography, reconstructed from digital breast tomosynthesis, and digital mammography: evaluation of lesion conspicuity and BI-RADS assessment categories. **Breast Cancer Research and Treatment**, v. 166, n. 3, p. 765–773, Dec 2017. ISSN 1573-7217.

MCLEOD, P.; VERMA, B. Variable hidden neuron ensemble for mass classification in digital mammograms [application notes]. **IEEE Computational Intelligence Magazine**, v. 8, n. 1, p. 68–76, 2013.

MENEZES, P. B. Linguagens formais e automatos. Bookman: Institudo de Informática da UFRGS, 5. ed., 2008.

MENZE, B. et al. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. **BMC Bioinformatics**, v. 10, n. 1, p. 213, Jul 2009. ISSN 1471-2105. Disponível em: <https://doi.org/10.1186/1471-2105-10-213>.

MISHRA, S.; RANGANATHAN, H. Multi layer architecture for breast cancer diagnosis. **Indian Journal of Computer Science and Engineering**, v. 5, n. 1, p. 18–25, 2014.

MITCHELL-JR, G. W.; BASSETT, L. W. **Mastologia Prática**. [S.l.]: Revinter, 1988.

MOHANTY, A. K. et al. Texture-based features for classification of mammograms using decision tree. **Neural Computing and Applications**, v. 23, n. 3, p. 1011–1017, 2013. ISSN 1433-3058.

MOHANTY, F.; RUP, S.; DASH, B. Automated diagnosis of breast cancer using parameter optimized kernel extreme learning machine. **Biomedical Signal Processing and Control**, v. 62, p. 102108, 2020. ISSN 1746-8094.

MUDIGONDA, N. R.; RANGAYYAN, R. M.; DESAUTELS, J. E. L. Concavity and convexity analysis of mammographic masses via an iterative boundary segmentation algorithm. In: **Engineering Solutions for the Next Millennium. 1999 IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No.99TH8411)**. [S.l.: s.n.], 1999. v. 3, p. 1489–1494 vol.3. ISSN 0840-7789.

MUDIGONDA, N. R.; RANGAYYAN, R. M.; DESAUTELS, J. E. L. Gradient and texture analysis for the classification of mammography masses. **IEEE Transactions on Medical Imaging**, v. 19, n. 10, p. 1032–1043, 2000.

NANDI, R. J. et al. Classification of breast masses in mammograms using genetic programming and feature selection. **Medical & Biological Engineering & Computing**, v. 44, n. 8, p. 683–694, 2006.

NAPPI, J. et al. Algorithmic 3D simulation of breast calcifications for digital mammography. **Computer Methods and Programs in Biomedicine**, v. 66, n. 1, p. 115 – 124, 2001. ISSN 0169-2607.

NISHIKAWA, R. M. **Computer-aided Detection and Diagnosis**. London: Butterworth-Heinemann: Springer Berlin Heidelberg, 2010. 85-106 p.

NIU, J. et al. Multi-scale attention-based convolutional neural network for classification of breast masses in mammograms. **Medical Physics**, v. 48, n. 7, p. 3878–3892, 2021. Disponível em: <https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.14942>.

OGIELA, L.; TADEUSIEWICZ, R.; OGIELA, M. Cognitive modeling in medical pattern semantic understanding. In: **International Conference on Multimedia and Ubiquitous Engineering, 2008. MUE 2008.** Busan, South Korea: [s.n.], 2008. p. 15 –18.

OMS. **Organização Mundial da Saúde**. 2023. <http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/>. Acessado: 18-02-2023.

PANCHAL, R.; VERMA, B. Neural classification of mass abnormalities with different types of features in digital mammography. **International Journal of Computational Intelligence and Applications**, v. 6, n. 1, p. 61–75, 2006.

PARAG, T. et al. A grammar for hierarchical object descriptions in logic programs. In: **IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.: s.n.], 2012. p. 33 –38. ISSN 2160-7508.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PEDRO, R. W. D. et al. **A Stochastic Grammar Approach to Mass Classification in Mammograms**. 2023. Preprint on webpage at <https://www.computer.org/csdl/journal/tb/5555/01/10049515/1KYocDkC1QQ>.

PEDRO, R. W. D.; MACHADO-LIMA, A.; NUNES, F. L. S. Is mass classification in mammograms a solved problem? - a critical review over the last 20 years. **Expert Systems with Applications**, v. 119, p. 90 – 103, 2019. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417418306821>.

PEDRO, R. W. D.; MACHADO-LIMA, A.; NUNES, F. L. S. A new syntactic approach for masses classification in digital mammograms. In: **IEEE CBMS International Symposium on Computer-Based Medical Systems, 32th**. [S.l.: s.n.], 2019.

PEDRO, R. W. D.; MACHADO-LIMA, A.; NUNES, F. L. S. Towards an approach using grammars for automatic classification of masses in mammograms. **Computational Intelligence**, v. 37, n. 4, p. 1515–1544, 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/coin.12320>.

PEDRO, R. W. D.; NUNES, F. L. S.; MACHADO-LIMA, A. Using grammars for pattern recognition in images: A systematic review. **ACM Computing Surveys**, ACM, New York, NY, USA, v. 46, n. 2, p. 26:1–26:34, nov. 2013. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/2543581.2543593>.

PEZESHKI, H. et al. Mass classification of mammograms using fractal dimensions and statistical features. **Multidimensional Systems and Signal Processing**, v. 32, n. 2, p. 573–605, 2021. Disponível em: <https://doi.org/10.1007/s11045-020-00749-6>.

PRUSINKIEWICZ, P.; LINDENMAYER, A.; HANAN, J. Development models of herbaceous plants for computer imagery purposes. **SIGGRAPH Comput. Graph.**, ACM, New York, NY, USA, v. 22, n. 4, p. 141–150, jun. 1988. ISSN 0097-8930. Disponível em: <http://doi.acm.org/10.1145/378456.378503>.

PYTHON-PACKAGE, L. **LightGBM Python-package**. 2023. Disponível em: <https://lightgbm.readthedocs.io/en/latest/Python-Intro.html>.

QI, S. et al. A generalized earley parser for human activity parsing and prediction. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 43, n. 8, p. 2538–2554, 2021.

RAMER, U. An iterative procedure for the polygonal approximation of plane curves. **Computer graphics and image processing**, v. 1, n. 3, p. 244–256, 1972.

RAMOS, M. V. M.; NETO, J. J.; VEGA, I. S. **Linguagens Formais - Teoria, Modelagem e Implementação**. 1st. ed. [S.l.]: Bookman Companhia Ed, 2009. ISBN 8577804534.

RANGAYYAN, R. M.; MUDIGONDA, N. R.; DESAUTELS, J. E. L. Boundary modelling and shape analysis methods for classification of mammographic masses. **Medical and Biological Engineering and Computing**, v. 38, n. 5, p. 487–496, 2000. ISSN 1741-0444. Disponível em: <http://dx.doi.org/10.1007/BF02345742>.

RANGAYYAN, R. M.; NGUYEN, T. M. Fractal analysis of contours of breast masses in mammograms. **Journal of Digital Imaging**, v. 20, n. 3, p. 223–237, 2007.

RANGAYYAN, R. M. et al. Effect of pixel resolution on texture features of breast masses in mammograms. **Journal of Digital Imaging**, v. 23, n. 5, p. 547–553, 2010.

RIBEIRO, M. X. et al. Data pre-processing: A new algorithm for feature selection and data discretization. In: **International Conference on Soft Computing As Transdisciplinary Science and Technology, Proceedings of the 5th**. New York, NY, USA: ACM, 2008. (CSTST '08), p. 252–257. ISBN 978-1-60558-046-3. Disponível em: <http://doi.acm.org/10.1145/1456223.1456277>.

RODRíGUEZ-ESPARZA, E. et al. Automatic detection and classification of abnormal tissues on digital mammograms based on a bag-of-visual-words

approach. In: HAHN, H. K.; MAZUROWSKI, M. A. (Ed.). **Medical Imaging 2020: Computer-Aided Diagnosis**. SPIE, 2020. v. 11314, p. 500 – 507. Disponível em: <https://doi.org/10.1117/12.2549899>.

SABER, A. et al. A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique. **IEEE Access**, v. 9, p. 71194–71209, 2021.

SAHINER, B. et al. Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. **Medical Physics**, v. 25, n. 4, p. 516 – 526, 1998.

SALOMAA, A. Formal languages. Academic Press, 1973.

SANTOS, V. T. **Segmentação de imagens mamográficas para detecção de nódulos em mamas densas**. Dissertação (Mestrado) — Universidade de São Paulo, 11 2002.

SAUNDERS, R. et al. Simulation of mammographic lesions. **Academic Radiology**, Elsevier, v. 13, n. 7, p. 860–870, 2021/10/18 2006. Disponível em: <https://doi.org/10.1016/j.acra.2006.03.015>.

SCHIABEL, H. **Esquemas CAD: uma análise dos seus aspectos e aplicações como ferramenta de auxílio ao diagnóstico em mamografia**. [S.l.]: Canal 6, 2014.

SCHIE, G. van et al. Generating synthetic mammograms from reconstructed tomosynthesis volumes. **IEEE Transactions on Medical Imaging**, v. 32, n. 12, p. 2322–2331, Dec 2013. ISSN 0278-0062.

SHEN, T. et al. Mass image synthesis in mammogram with contextual information based on gans. **Computer Methods and Programs in Biomedicine**, v. 202, p. 106019, 2021. ISSN 0169-2607. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169260721000948>.

SOLTANPOUR, S.; HOSSEIN, E. Learning novel object parts model for object categorization. In: . [S.l.]: International Symposium on Telecommunications (IST 2010), 5th, 2010. p. 796–800.

SONG, R.; LI, T.; WANG, Y. Mammographic classification based on xgboost and dcnn with multi features. **IEEE Access**, v. 8, p. 75011–75021, 2020.

SOUSA, M. A. Z. **Desenvolvimento de um objeto simulador de mama: investigações da percepção visual da imagem e do desempenho de esquemas Cadx**. Tese (Doutorado) — Universidade de São Paulo, São Carlos, SP, Brasil, 2017.

SUCKLING, J. et al. The mammographic image analysis society digital mammogram database. In: **International workshop on digital mammography. Proceedings of the 2nd**. [S.l.: s.n.], 1994. (Proceedings of the 2nd international workshop on digital mammography), p. 375–378.

SUN, R. et al. Image-based lightweight tree modeling. In: **International Conference on Virtual Reality Continuum and its Applications in Industry. Proceedings of the 8th**. New York, NY, USA: ACM, 2009. (VRCAI '09), p. 17–22. ISBN 978-1-60558-912-1. Disponível em: <http://doi.acm.org/10.1145/1670252.1670258>.

TAHMASBI, A.; SAKI, F.; SHOKOUHI, S. B. Cwla: A novel cognitive classifier for breast mass diagnosis. In: **2011 Iranian Conference of Biomedical Engineering (ICBME), 18th**. [S.l.: s.n.], 2011. p. 255–259.

TAN, M.; PU, J.; ZHENG, B. A new and fast image feature selection method for developing an optimal mammographic mass detection scheme. **Medical Physics**, v. 41, n. 8, 2014.

TAN, M.; PU, J.; ZHENG, B. Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support vector machine (SVM) model. **International Journal of Computer Assisted Radiology and Surgery**, v. 9, n. 6, p. 1005–1020, 2014.

TAYLOR, P.; OWENS, R.; INGRAM, D. Simulated mammography using synthetic 3D breasts. In: ____. **Digital Mammography: Nijmegen, 1998**. Dordrecht: Springer Netherlands, 1998. p. 283–290. ISBN 978-94-011-5318-8.

TODD, C. A.; NAGHDY, G. Method for breast cancer classification based solely on morphological descriptors. In: . [S.l.: s.n.], 2004. v. 5370, p. 857–867.

TRALIC, D.; BOZEK, J.; GRGIC, S. Shape analysis and classification of masses in mammographic images using neural networks. In: **2011 International Conference on Systems, Signals and Image Processing, 18th**. [S.l.: s.n.], 2011. p. 1–5. ISSN 2157-8672.

TRZUPEK, M.; OGIELA, M.; TADEUSIEWICZ, R. Intelligent image content description and analysis for 3D visualizations of coronary vessels. In: NGUYEN, N.; KIM, C.-G.; JANIAK, A. (Ed.). **Intelligent Information and Database Systems**. [S.l.]: Springer Berlin Heidelberg, 2011, (Lecture Notes in Computer Science, v. 6592). p. 193–202. ISBN 978-3-642-20041-0.

TRZUPEK, M.; OGIELA, M. R. Linguistic approach to modeling of coronary arteries in semantic techniques of image retrieval. In: **2014 Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing**. [S.l.: s.n.], 2014. p. 295–299.

VELTHUIZEN, R. P.; GANGADHARAN, D. Mammographic mass classification: initial results. In: . [S.l.: s.n.], 2000. v. 3979, p. 68–76.

VERMA, B. Novel network architecture and learning algorithm for the classification of mass abnormalities in digitized mammograms. **Artificial Intelligence in Medicine**, v. 42, n. 1, p. 67–79, 2008.

WANG, Q.; JIANG, Z. A grammatical framework for building rooftop extraction. In: **2009 IEEE International Geoscience and Remote Sensing Symposium**. [S.l.: s.n.], 2009. v. 3, p. III–334–III–337.

WANG, Y.; BAHRAMI, S.; ZHU, S.-C. Perceptual scale space and its applications. In: . [S.l.]: IEEE International Conference on Computer Vision, 10th, 2005. v. 1, p. 58–65.

WU, Y. et al. A comprehensive methodology for determining the most informative mammographic features. **Journal of Digital Imaging**, v. 26, n. 5, p. 941–947, 2013.

XIONG, C. et al. Robot learning with a spatial, temporal, and causal and-or graph. In: **2016 IEEE International Conference on Robotics and Automation (ICRA)**. [S.l.: s.n.], 2016. p. 2144–2151.

YANG, S.-C. et al. A computer-aided system for mass detection and classification in digitized mammograms. **Biomedical Engineering: Applications, Basis and Communications**, v. 17, n. 05, p. 215–228, 2005.

ZHU, S. C.; MUMFORD, D. A stochastic grammar of images. Foundations and Trends in Computer Graphics and Vision, v. 2, n. 4, p. 259 – 362, 2006.

PARTE V

APÊNDICES E ANEXOS

APÊNDICE A – TRABALHOS FUTUROS

Uma possível evolução deste trabalho de pesquisa, no que diz respeito a classificação de nódulos, poderia estar relacionada à utilização de métodos de segmentação automática dos nódulos. A segmentação automática seria importante porque minimizaria a necessidade de segmentação manual por radiologistas, evitando a realização de um trabalho bastante cansativo e sujeito a erro.

As gramáticas utilizadas foram criadas de forma direta a partir de uma estrutura hierárquica (grafos AND-OR) utilizada para representar todos os nódulos benignos e malignos. Entretanto, poderiam ser conduzidos experimentos para avaliar a possibilidade do aprendizado das regras gramaticais a partir de métodos de inferência gramatical. Neste sentido, cada nódulo poderia ser representado por uma estrutura hierárquica específica e as regras gramaticais seriam inferidas a partir dessas várias estruturas hierárquicas.

Pôde-se perceber neste trabalho que os modelos gramaticais apresentaram um desempenho adequado mesmo com a utilização de um número limitado de imagens. Entretanto, poderia ser analisada a quantidade mínima de imagens necessária para que os modelos gramaticais fossem capazes de aprender os padrões dos nódulos benignos e malignos.

No que diz respeito à geração de imagens sintéticas de nódulos, devem ser realizados mais estudos acerca da capacidade generativa das gramáticas para validar se esta abordagem pode ser utilizada para a criação deste tipo de imagem. Além disso, deve-se investigar outras técnicas de geração de imagens, por exemplo as redes neurais generativas e analisar a aplicabilidade deste tipo de abordagem para atacar o problema de geração de imagens sintéticas de nódulos. Uma forma de validar se as imagens sintéticas são semelhantes às imagens reais, além de uma avaliação realizada por médicos especialistas, seria treinar os modelos gramaticais utilizando as imagens sintéticas e fazer a classificação das imagens reais.

APÊNDICE B – TRABALHOS PUBLICADOS

As contribuições em forma de artigos publicados gerados a partir desta tese estão elencadas a seguir.

- PEDRO, R. W. D.; MACHADO-LIMA, A.; NUNES, F. L. S. Is mass classification in mammograms a solved problem? - a critical review over the last 20 years. Expert Systems with Applications, v. 119, p. 90 – 103, 2019. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417418306821>.
  - Este artigo apresentou a revisão bibliográfica sistemática realizada, na qual foram encontrados o estado da arte e as possíveis lacunas desta área de pesquisa.
- PEDRO, R. W. D.; MACHADO-LIMA, A.; NUNES, F. L. S. A new syntactic approach for masses classification in digital mammograms. In: 32th IEEE CBMS International Symposium on Computer-Based Medical Systems (CBMS), 2019. Disponível em: <https://ieeexplore.ieee.org/document/8787498>.
  - Neste artigo foram apresentados os resultados preliminares da utilização de modelos gramaticais para a classificação dos nódulos, empregando o *dataset* fornecido pelos pesquisadores da Universidade de Calgary e características de forma, textura e gradiente.
- HIRAMA, R. et al. Evaluating the impact of polygonal representations on mass classification. In: International Conference on Systems, Signals and Image Processing (IWSSIP), 2020. p. 75–80. Disponível em: <https://ieeexplore.ieee.org/document/9145038>.
  - Trabalho que avaliou o impacto que as representações poligonais dos nódulos exercem no processo de classificação dos nódulos considerando as classes Benigno e Maligno.

- PEDRO, R. W. D.; MACHADO-LIMA, A.; NUNES, F. L. S. Towards an approach using grammars for automatic classification of masses in mammograms. Computational Intelligence, v.37, p. 01-30, 2020.Disponível em: <https://doi.org/10.1111/coin.12320>.

  – Artigo que apresentou a utilização e calibração de diferentes algoritmos para discretização das características, bem como a utilização da importância de Gini para selecionar as características mais importantes para a classificação dos nódulos. Além disso, este artigo também apresenta uma comparação dos resultados obtidos pelos modelos gramaticais com os resultados obtidos por outros modelos de aprendizado de máquina.

- PEDRO, R. W. D. et al. A stochastic grammar approach to mass classification in mammograms. IEEE/ACM Transactions on Computational Biology and Bioinformatics, PrePrints pp. 1-12, 2023. DOI Bookmark: 10.1109/TCBB.2023.3247144.

  – Trabalho que teve como um dos objetivos verificar a robustez dos modelos gramaticais criados. Para isso, apresentou o novo *dataset* construído com imagens fornecidas pelos pesquisadores do A. C. Camargo Cancer Center, introduziu a utilização de momentos de Hu como características dos nódulos para servir os modelos modelos gramaticais e demonstrou os resultados das classificações quando o treinamento do modelo era realizado com um *dataset* e o teste era realizado com outro *dataset*.

APÊNDICE C – TABELAS RESULTADOS

Tabela 14: Melhores resultados obtidos pelo **Modelo 1** para cada número de características selecionadas utilizando apenas as características de forma e o tipo da borda do nódulo (circunscrito/espiculado) - características extraídas por pesquisadores da Universidade de Calgary. Sen. = Sensibilidade; Esp. = Especificidade; Acc. = Acurácia; AUC = área sob a curva ROC.

Algoritmo Ômega					
N. características	**Hiperparâmetros**	**Sen. (%)**	**Esp. (%)**	**Acc. (%)**	**AUC**
3	$H_{min} = 2; \zeta_{max} = 0.35$	89	93	91	0,91
4	$H_{min} = 2; \zeta_{max} = 0.35$	96	97	96	0,96
5	$H_{min} = 2; \zeta_{max} = 0.35$	98	97	97	0,97
8	$H_{min} = 2; \zeta_{max} = 0.35$	98	97	97	0,97
Algoritmo KbinsDiscretizer					
N. características	**Hiperparâmetros**	**Sen.**	**Esp.**	**Acc.**	**AUC**
3	$n\_bins = 40$	96	93	94	0,94
4	$n\_bins = 40$	96	94	95	0,95
5	$n\_bins = 40$	98	97	97	0,97
8	$n\_bins = 40$	98	97	97	0,97

Fonte: O autor (2023)

Tabela 15: Melhores resultados obtidos pelo **Modelo 2** para cada número de características selecionadas utilizando características de forma e de textura e considerando o tipo da borda do nódulo (circunscrito/espiculado) - características extraídas por pesquisadores da Universidade de Calgary. Sen. = Sensibilidade; Esp. = Especificidade; Acc. = Acurácia; AUC = área sob a curva ROC.

Algoritmo Ômega					
N. características	**Hiperparâmetros**	**Sen. (%)**	**Esp. (%)**	**Acc. (%)**	**AUC**
3	$H_{min} = 2; \zeta_{max} = 0.35$	96	100	98	0,98
4	$H_{min} = 2; \zeta_{max} = 0.35$	98	100	99	0,99
5	$H_{min} = 2; \zeta_{max} = 0.35$	98	100	99	0,99
8	$H_{min} = 2; \zeta_{max} = 0.35$	100	100	100	1
Algoritmo KbinsDiscretizer					
N. características	**Hiperparâmetros**	**Sen.**	**Esp.**	**Acc.**	**AUC**
3	$n\_bins = 40$	98	96	96	0,97
4	$n\_bins = 40$	98	96	96	0,97
5	$n\_bins = 40$	98	100	99	0,99
8	$n\_bins = 40$	100	100	100	1

Fonte: O autor (2023)

Tabela 16: Melhores resultados obtidos pelo **Modelo 3** para cada número de características selecionadas utilizando características de forma e de textura sem considerar o tipo de borda do nódulo (circunscrito/espiculado) - características extraídas por pesquisadores da Universidade de Calgary. Sen. = Sensibilidade; Esp. = Especificidade; Acc. = Acurácia; AUC = área sob a curva ROC.

Algoritmo Ômega					
N. características	**Hiperparâmetros**	**Sen. (%)**	**Esp. (%)**	**Acc. (%)**	**AUC**
3	$H_{min} = 2; \zeta_{max} = 0.35$	96	90	92	0,93
4	$H_{min} = 2; \zeta_{max} = 0.35$	98	90	93	0,94
5	$H_{min} = 2; \zeta_{max} = 0.35$	96	90	92	0,93
8	$H_{min} = 2; \zeta_{max} = 0.35$	98	90	93	0,94
Algoritmo KbinsDiscretizer					
N. características	**Hiperparâmetros**	**Sen.**	**Esp.**	**Acc.**	**AUC**
3	$n\_bins = 40$	91	93	92	92
4	$n\_bins = 40$	98	93	95	96
5	$n\_bins = 40$	98	96	97	97
8	$n\_bins = 40$	100	100	100	1

Fonte: O autor (2023)

Tabela 17: Melhores resultados alcançados por cada classificador (características extraídas por pesquisadores da Universidade de Calgary). Características: compacidade (CP), índice de espiculação (IE), fração de concavidade (FC), fator de Fourier (FF), dimensão fractal 2D box counting (2B), dimensão fractal 1D box counting (1B), dimensão fractal 2D ruler (2R), dimensão fractal 1D ruler (1R), contraste (CO), acutância (AC), acutância tradicional (AT) e coeficiente de variação (CV).

Classificador	**Características**	**Sensibilidade (%)**	**Especificidade (%)**	**Acurácia (%)**	**AUC (%)**
ANN	CC, FC, 2B, 1B, 2R, AC	100	86	92	93
SVM	SI, FC, FF, CO, CV	87	97	92	91
KNN	SI, 2R CO, CV	83	95	90	89
RF	2R, 1R, CO, AC	85	91	89	88

Fonte: O autor (2023)

Tabela 18: Medidas de desempenho alcançadas por cada modelo utilizando características de forma (características extraídas neste projeto de pesquisa): sensibilidade, especificidade, acurácia, F-score e Matthews correlation coefficient (MCC). *Datasets* combinados: treinamento = ambos, teste = ambos; *datasets* ACC→ST: treinamento = ACC; teste = ST; *datasets* ST→ACC: treinamento = ST, teste = ACC. Os melhores resultados estão em negrito.

***Datasets* combinados**						
Medida	**ANN**	**KNN**	**LGBM**	**RF**	**SVM**	**M. Gramatical**
Sensibilidade	91%	90%	92%	92%	88%	**99**%
Especificidade	89%	89%	91%	88%	92%	**100**%
Acurácia	90%	89%	92%	90%	90%	**99**%
F-score	92%	88%	92%	92%	89%	**99**%
MCC	0,81	0,79	0,85	0,81	0,81	**0,99**
***Datasets* ACC→ST**						
Medida	**ANN**	**KNN**	**LGBM**	**RF**	**SVM**	**M. Gramatical**
Sensibilidade	86%	86%	89%	86%	86%	**100**%
Especificidade	84%	80%	83%	80%	86%	**100**%
Acurácia	85%	82%	85%	82%	86%	**100**%
F-score	83%	80%	83%	80%	84%	**100**%
MCC	0,70	0,66	0,71	0,66	0,72	**1**
***Datasets* ST→ACC**						
Medida	**ANN**	**KNN**	**LGBM**	**RF**	**SVM**	**M. Gramatical**
Sensibilidade	91%	94%	81%	93%	93%	**98**%
Especificidade	93%	77%	88%	76%	87%	**100**%
Acurácia	92%	86%	85%	85%	90%	**99**%
F-score	92%	86%	84%	85%	90%	**99**%
MCC	0.85	0,73	0,70	0,71	0,81	**0,99**

Fonte: O autor (2023)

Tabela 19: Medidas de desempenho alcançadas por cada modelo utilizando momentos de Hu (características extraídas neste projeto de pesquisa): sensibilidade, especificidade, acurácia, F-score e Matthews correlation coefficient (MCC). *Datasets* combinados: treinamento = ambos, teste = ambos; *datasets* ACC→ST: treinamento = ACC; teste = ST; *datasets* ST→ACC: treinamento = ST, teste = ACC. Os melhores resultados estão em negrito.

***Datasets* combinados**						
Medida	**ANN**	**KNN**	**LGBM**	**RF**	**SVM**	**M. Gramatical**
Sensibilidade	66%	60%	82%	80%	42%	**99**%
Especificidade	91%	74%	86%	83%	92%	**98**%
Acurácia	79%	67%	84%	82%	69%	**98**%
F-score	73%	63%	82%	80%	54%	**98**%
MCC	0,61	0,35	0,70	0,66	0,41	**0.97**
***Datasets* ACC→ST**						
Medida	**ANN**	**KNN**	**LGBM**	**RF**	**SVM**	**M. Gramatical**
Sensibilidade	0%	39%	84%	84%	41%	**97**%
Especificidade	100%	24%	27%	4%	75%	**95**%
Acurácia	58%	30%	51%	37%	61%	**96**%
F-score	-	31%	59%	53%	46%	**95**%
MCC	-	-0,36	0,14	-0,18	0,17	**0,92**
***Datasets* ST→ACC**						
Medida	**ANN**	**KNN**	**LGBM**	**RF**	**SVM**	**M. Gramatical**
Sensibilidade	0%	39%	84%	84%	41%	**96**%
Especificidade	100%	24%	27%	4%	75%	**100**%
Acurácia	58%	30%	51%	37%	61%	**98**%
F-score	- %	31%	59%	53%	46%	**98**%
MCC	-	-0,36	0,14	-0,18	0,17	**0,97**

Fonte: O autor (2023)

Tabela 20: Medidas de desempenho alcançadas por cada modelo utilizando as características combinadas (características extraídas neste projeto de pesquisa: sensibilidade, especificidade, acurácia, F-score e Matthews correlation coefficient (MCC). *Datasets* combinados: treinamento = ambos, teste = ambos; *datasets* ACC→ST: treinamento = ACC; teste = ST; *datasets* ST→ACC: treinamento = ST, teste = ACC. Os melhores resultados estão em negrito.

***Datasets* combinados**						
Medida	**ANN**	**KNN**	**LGBM**	**RF**	**SVM**	**M. Gramatical**
Sensibilidade	96%	92%	93%	94%	93%	**100**%
Especificidade	80%	82%	86%	86%	88%	**100**%
Acurácia	88%	89%	89%	89%	90%	**100**%
F-score	88%	89%	89%	89%	90%	**100**%
MCC	0,78	0,8	0,81	0,81	0,82	**1**
***Datasets* ACC→ST**						
Medida	**ANN**	**KNN**	**LGBM**	**RF**	**SVM**	**M. Gramatical**
Sensibilidade	93%	89%	86%	86%	86%	**97**%
Especificidade	76%	81%	80%	80%	83%	**98**%
Acurácia	83%	84%	82%	82%	84%	**98**%
F-score	82%	82%	80%	80%	82%	**97**%
MCC	0,69	0,69	0,66	0,66	0,69	**0,96**
***Datasets* ST→ACC**						
Medida	**ANN**	**KNN**	**LGBM**	**RF**	**SVM**	**M. Gramatical**
Sensibilidade	87%	93%	100%	89%	89%	**100**%
Especificidade	90%	80%	65%	78%	91%	**100**%
Acurácia	89%	87%	82%	84%	90%	**100**%
F-score	88%	87%	84%	84%	90%	**100**%
MCC	0,78	0,75	0,69	0,68	0,81	**1**

Fonte: O autor (2023)