UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE DO CAMPUS ARARANGUÁ
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

Thaynara Tessaline Mitie Sei Soares

SISTEMA PARA ANÁLISE DE IMAGENS DE IMUNO-HISTOQUÍMICA PARA TECIDOS DE CARCINOMAS MAMÁRIOS UTILIZANDO TÉCNICAS DE VISÃO COMPUTACIONAL E RECONHECIMENTO DE PADRÕES

Araranguá
2020

Thaynara Tessaline Mitie Sei Soares

SISTEMA PARA ANÁLISE DE IMAGENS DE IMUNO-HISTOQUÍMICA PARA TECIDOS DE CARCINOMAS MAMÁRIOS UTILIZANDO TÉCNICAS DE VISÃO COMPUTACIONAL E RECONHECIMENTO DE PADRÕES

Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia de Computação do Centro de Ciências, Tecnologias e Saúde do Campus Araranguá da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Engenharia de Computação.
Orientador: Prof. Antonio Carlos Sobieranski, Dr.
Coorientador: Prof. Marcelo Daniel Berejuck, Dr.

Araranguá
2020

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Soares, Thaynara Tessaline Mitie Sei
Sistema para análise de imagens de imuno-histoquímica para tecidos de carcinomas mamários utilizando técnicas de visão computacional e reconhecimento de padrões / Thaynara Tessaline Mitie Sei Soares ; orientador, Antonio Carlos Sobieranski, coorientador, Marcelo Daniel Berejuck, 2020.
25 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá, Graduação em Engenharia de Computação, Araranguá, 2020.

Inclui referências.

1. Engenharia de Computação. 2. carcinoma mamário. 3. reconhecimento de padrões. 4. visão computacional. 5. imuno-histoquímica. I. Sobieranski, Antonio Carlos. II. Berejuck, Marcelo Daniel. III. Universidade Federal de Santa Catarina. Graduação em Engenharia de Computação. IV. Título.

Thaynara Tessaline Mitie Sei Soares

SISTEMA PARA ANÁLISE DE IMAGENS DE IMUNO-HISTOQUÍMICA PARA TECIDOS DE CARCINOMAS MAMÁRIOS UTILIZANDO TÉCNICAS DE VISÃO COMPUTACIONAL E RECONHECIMENTO DE PADRÕES

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel em Engenharia de Computaçãoe aprovado em sua forma final pelo Curso de Graduação em Engenharia de Computação.

Araranguá, 11 de dezembro de 2020.

Documento assinado digitalmente
Fabricio de Oliveira Ourique
Data: 16/12/2020 13:56:31-0300
CPF: 916.167.860-00

Prof. Fabrício De Oliveira Ourique, Dr.
Coordenador do Curso

Banca Examinadora:

Documento assinado digitalmente
Antonio Carlos Sobieranski
Data: 16/12/2020 18:52:50-0300
CPF: 005.305.809-77

Prof. Antonio Carlos Sobieranski, Dr.
Orientador

Documento assinado digitalmente
Marcelo Daniel Berejuck
Data: 16/12/2020 14:53:00-0300
CPF: 640.174.519-34

Prof. Marcelo Daniel Berejuck, Dr.
Coorientador
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Anderson Luiz Fernandes Perez
Data: 16/12/2020 10:10:38-0300
CPF: 020.803.459-58

Prof. Anderson Luiz Fernandes Perez, Dr.
Avaliador
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Rodrigo Vinicius Mendonca Pereira
Data: 16/12/2020 17:14:27-0300
CPF: 028.663.779-07

Prof. Rodrigo Vinicius Mendonça Pereira, Dr.
Avaliador
Universidade Federal de Santa Catarina

Sistema para análise de imagens de imuno-histoquímica para tecidos de carcinomas mamários utilizando técnicas de visão computacional e reconhecimento de padrões

Thaynara Tessaline Mitie Sei Soares* Marcelo Daniel Berejuck†
Antonio Carlos Sobieranski‡

2020, dezembro

Resumo

A doença com maior incidência em mulheres no mundo e sua principal causa de óbito é o câncer de mama. As chances de sobrevivência podem ser aumentadas com o diagnóstico precoce, obtido através da análise de imagens do tecido canceroso por um profissional da área de patologia. Embora existam sistemas computacionais que atuem no auxílio ao diagnóstico, muitas vezes o esforço manual é preferível para detecção e contagem celular na amostra a fim de se obter métricas quantificáveis da expressão da doença no tecido. Nesse contexto, o trabalho em questão apresenta a modelagem de um sistema computacional, cujo objetivo é a detecção e contagem automatizada de células em amostras de câncer mamário, pigmentados por imuno-histoquímica utilizando técnicas de visão computacional e reconhecimento de padrões. Além do uso de métodos computacionais já difundidos, criou-se aqui um método iterativo para segmentação de núcleos celulares sobrepostos. Resultados preliminares demonstraram 94.66% de precisão do algoritmo na detecção nuclear.

Palavras-chaves: Carcinoma Mamário. Reconhecimento de Padrões. Visão Computacional. Imuno-Histoquímica.

*thaynara.mitie@grad.ufsc.br
†berejuck@ieee.org
‡a.sobieranski@ufsc.br

Immunohistochemistry image analysis system for breast cancer tissue using computer vision and pattern recognition techniques

Thaynara Tessaline Mitie Sei Soares* Marcelo Daniel Berejuck†
Antonio Carlos Sobieranski‡

2020, dezembro

Abstract

The disease with the highest worldwide incidence in women and their main cause of death is breast cancer. The chances of survival can be increased with early diagnosis, obtained through the analysis of images of cancerous tissue by a professional in the field of pathology. Although computational systems exist that can assist in the diagnosis, manual effort is often preferable for detecting and counting cells in the sample in order to obtain quantifiable metrics of the expression of the disease in the tissue. In this context, the work in question presents the modeling of a computer system whose objective is the detection and automated counting of cells in breast cancer samples, pigmented by immunohistochemistry using computer vision and pattern recognition techniques. In addition to the use of already-widespread computational methods, an iterative method was created here for the segmentation of overlapping cell nuclei. Preliminary results have demonstrated a $94,66\%$ accuracy of the algorithm in nuclear detection.

Key-words: Breast Cancer. Patterns Recognition. Computer Vision. Immunohistochemistry.

1 Introdução

Consoante às informações do relatório em (BRAY et al., 2018), em 2018 foram detectados 2.088.849 novos casos de câncer de mama no mundo, sendo em sua totalidade exclusivo em mulheres. Dentre os casos detectados, 30% vieram a óbito pela doença. Estima-se que para o intervalo compreendido entre 2018 a 2023 aproximadamente 7 milhões de

*thaynara.mitie@grad.ufsc.br
†berejuck@ieee.org
‡a.sobieranski@ufsc.br

mulheres morrerão devido a esse tipo de câncer. O diagnóstico precoce aumenta as chances de sobrevivência do paciente, uma vez que o tratamento adequado pode ser aplicado, permitindo que haja até 95% de chances de remissão total da doença (AMARANTE, 2020). Uma das ferramentas de diagnósticos disponíveis é a biópsia de mama (exame anatomopatológico), procedimento este que consiste em coletar uma pequena quantidade de tecido para ser analisado por um patologista (ONCOGUIA, 2020). A amostra de tecido é preservada por uma técnica conhecida como FFPE – *formalin fixed paraffin embedded*[1] e analisada através de um exame de Imuno-histoquímica, onde o especialista irá avalia-lá quanto à presença e intensidade dos receptores, ditos como 'portas de entrada', aptos a receberem estímulos para as células malignas crescerem.

A Imuno-histoquímica (IHC) pode ser descrita como um método para constatar e quantificar a manifestação de determinada doença (COELHO, 2018). Os exemplares de tecidos de células coletadas são pigmentados com um reagente, ou marcador, de maneira a detectar antígenos específicos (proteínas) na amostra e fornecer não somente dados sobre a presença do câncer nas células, mas também prover métricas precisas e confiáveis quanto a quantidade relativa ou real da proteína no tecido (DABBS, 2017).

A identificação e contagem celular faz-se necessária para a técnica de IHC, visto que em sistemas de pontuações como Allred, que avalia os receptores de estrogênio e progesterona na amostra, conta-se a proporção de células positivas para câncer e sua intensidade, para que um tratamento de câncer adequado ao problema apresentado seja proferido.

No que diz respeito a identificação, ou contagem celular, é de domínio público programas que fornecem ferramentas para tal, assim como em *IMAGEJ®* e suas diferentes distribuições (RUEDEN et al., 2017), bem como em algoritmos e programas em estágios de produção que serão abordados na Seção 2. Por mais que a área de análise de imagens digitais em patologias mamárias esteja em crescente ascensão (CHANG; MRKONJIC, 2020), o emprego do esforço manual ou puramente visual ainda é utilizado para marcação de núcleos celulares de câncer. Tal tratativa, por inferência manual, pode gerar interpretação intrinsecamente subjetiva na hora de proferir um diagnóstico visto que a quantificação da expressão de receptores na amostra pode variar de acordo com o observador (LEHR et al., 1997). Em (ELMORE et al., 2015) há o estudo que demonstra a discordância entre patologistas em relação a quantificação de células nos diagnósticos de biópsias de câncer de mama. Observa-se no estudo que há grande nível de concordância entre os especialistas para amostras de câncer invasivo (94%), níveis mais baixos para casos atípicos (48%) e senso comum estimado em $75,3\%$ na interpretação geral das amostras. Os casos atípicos requerem maior atenção durante a análise pois pequenas variações no percentual de detecção de células receptoras podem ser decisivas na definição do tratamento (BATTIFORA et al., 1993). Dessa forma, o auxílio computacional para o laudo da doença faz-se valioso para quantificar de modo satisfatório a manifestação de câncer neste tipo de amostra.

Sistemas de visão computacional podem ser verificados na literatura para automatização do processo de quantificação da expressão de Imuno-histoquímica. Os sistemas podem ser moldados de maneira a utilizar abordagem clássica, de modo a encontrar parâmetros para o particionamento de um espaço de cores ou ainda para a detecção de contornos, como ocorre em técnicas de *Thresholding* e *Canny*, respectivamente. Ainda podem seguir abordagens conexionistas com o uso de inteligência artificial e suas várias ramificações para

[1] FFPE: Fixada em Formalina e Incluída em Parafina. Método de conservação e preparação para o corte em lâminas da amostra de biópsia.

identificação de padrões e segmentação de regiões de interesse. Observa-se em (CHANG; MRKONJIC, 2020) que embora abordagens mais recentes em análise de imagens digitais de patologias mamárias possam ter obtido alto nível de acurácia na execução proposta, tanto na área clássica quanto conexionista, estes utilizam-se de imagens de *slides* inteiros (*Whole Slides Image* - WSI), que são digitalizações de alta resolução da lâmina (DPA, 2012), independentes de fatores externos, como a falta de equalização na iluminação ou carência de foco na imagem. Além disso, necessitam da digitalização da amostra, o que pode ser um processo custoso já que uma WSI pode possuir até 6 GB a depender da ampliação utilizada (LAURO et al., 2013).

Neste contexto, este documento apresenta os resultados obtidos com a implementação de um algoritmo desenvolvido para a detecção de células cancerosas em imagens obtidas por técnica de Imuno-histoquímica a partir tecidos de carcinoma mamário. O algoritmo foi desenvolvido utilizando técnicas de visão computacional e reconhecimento de padrões, com o objetivo de ser uma solução viável para o problema aqui apresentado. O *dataset* (conjunto de dados) utilizado provém de câmeras de *smartphone* e foi fornecido pelo médico patologista Arthur Conelian Gentili, atualmente integrante do CEPON (Centro de Pesquisas Oncológicas) e colaborador deste trabalho (LATTES, 2020). Resultados experimentais demonstram significativo grau de acurácia do algoritmo para um determinado padrão de imagens que será abordado na Seção 4 (abordagem proposta).

O presente documento está organizado da seguinte forma: A Seção 2 contém os trabalhos correlatos pertinentes ao problema de identificação de células apresentado. Na Seção 3 são apresentadas a fundamentação teórica acerca dos métodos de Processamento Digital de Imagens e Reconhecimento de Padrões para solução do problema de pesquisa. Na Seção 4 há a discussão sobre a abordagem proposta. Na Seção 5 são apresentados os resultados experimentais. Por fim, na Seção 6 são apresentadas as conclusões, discussões assim como os trabalhos futuros.

2 Trabalhos Relacionados

A revisão de literatura realizada possibilitou identificar duas principais vertentes em abordagens computacionais para a solução do problema de detecção automatizada de núcleos celulares. Uma das técnicas é fundamentada no uso de abordagem clássica com métodos de morfologia matemática e processamento de imagens, enquanto a outra abordagem é baseada em técnicas conexionistas, com o uso de inteligência artificial – sendo esta a área de maior incidência em artigos atuais.

Em (SILVA, 2015) há a utilização de técnicas de morfologia matemática para realizar a contagem de células na amostra. Os resultados demonstraram 89% de sucesso para estimar um valor médio de núcleos na amostra. Neste trabalho o tipo de células identificado é único, ou seja, pertencem ao mesmo conjunto de cores, e a imagem recebida pelo software deve ser em formato TIFF (*Tagged Image File Format*), característica por possuir elevada definição real de cores. O conjunto de dados para validação foi fornecido pelo *INMETRO* e o software foi escrito em *MATLAB*®.

No software comercial denominado *QuickCount*® (TIONG et al., 2018), escrito em C++, é proposto a contagem automatizada e em grande escala de células, utilizando métodos computacionais não conexionistas. Embora o software permita a contagem de células de linhagens diferentes ao identificar o canal de cores que será extraído com base no nome do arquivo das amostras na imagem, o mesmo apresenta somente um tipo de célula

em análise, logo, pertencem ao mesmo padrão de cores, monocromáticas. O conjunto de dados utilizado para testes é próprio, em formato PNG (*Portable Network Graphics*) e capturadas por um microscópio *Olympus IX71* com ampliação de 10x.

Em (MOUELHI et al., 2018) há um algoritmo capaz de realizar automaticamente a classificação *Allred* (*score* para classificação de proporção e intensidade de células positivas para receptores de estrogênio na amostra), cujos resultados registraram mais de 98% de precisão na detecção dos núcleos e classificação da pontuação de câncer *Allred* em um banco de dados de 84 imagens. O método de segmentação apresentado baseia-se em técnicas como limiares locais adaptativos, operações morfológicas, filtro Laplaciano modificado e um aprimoramento do método de segmentação da imagem denominado *Watershed*.

Seguindo a mesma linha, há os softwares de distribuição gratuita *IMAGEJ*® e *CellProfiler*® com recursos para contagem de células automatizada, além de diversos outros métodos, como descritos em (CHOUDHRY, 2016), (Al-Kofahi et al., 2010) (RICCIO et al., 2018) e (PANAGIOTAKIS; ARGYROS, 2018).

Em contraste aos enfoques anteriores, na literatura várias abordagens computacionais podem ser verificadas, descrevendo métodos para detecção, contagem e classificação de células, específicas de carcinoma mamário ou não, utilizando aprendizagem profunda e redes neurais. Por exemplo, em (HOSSEINI; CHEN; JABLONSKI, 2020) utiliza-se a técnica de aprendizado profundo, também conhecido como *Deep Learning*, modelo Mask R-CNN, para detecção e contagem de células de retina em grandes coleções de dados, a qual se obteve precisão superior a 80%. As imagens utilizadas para treinamento da rede convolucional eram de alta resolução (2048 $\times$ 2048 pixels), porém a técnica também trabalha com resoluções inferiores (512 $\times$ 512 pixels), e estas imagens foram fornecidas pelo *Royal College of Surgeons* (RCS) capturadas de retinas de ratos e obtidas por microscopia eletrônica de Transmissão (TEM). Para a realização do experimento, foi utilizado uma máquina CPU Intel® Core™ i7–8700K com seis núcleos e 3,7 GHz de frequência, 64 GB de memória DDR2 e uma GPU NVIDIA™ GeForce® RTX 2060.

Tendo em vista os estudos apresentados acima, percebe-se que as técnicas baseadas em processamento de imagens e reconhecimento de padrões são mais específicas, logo apresentam maior porcentagem de acurácia, pois o fluxo computacional foi construído especificamente para tratar um tipo de problema. No entanto, abordagens ditas clássicas como anteriormente mencionadas tendem a apresentar baixa generalidade, alta dependência dos parâmetros de entrada, assim como o mesmo tipo de imagem de entrada. Técnicas conexionistas, que se utilizam de inteligência artificial são mais genéricas, porém apresentam em geral menor percentual de acurácia sendo estas muito dependentes do conjunto de treinamento, que requerem grandes bases de dados, além de necessitarem de hardwares de alto desempenho. Para o presente trabalho optou-se pelo desenvolvimento de um fluxo computacional clássico, visto que o problema em questão é moderadamente bem comportado, com relativamente baixa variabilidade entre as amostras.

3 Fundamentação Teórica

3.1 Princípios Gerais

Desde a criação da primeira câmera digital em 1969, a captura de imagens tornou-se algo vital na sociedade, com funções que vão desde registro histórico a diagnósticos completos por imagem. Sendo assim, muito se desenvolveu em relação à algoritmos para fazer

tratativas na imagem a fim de detectar alguma característica em especial. São inúmeras as áreas que estudam e implementam a utilização de imagens para fins específicos, como a detecção de placas no sistema de trânsito, identificação facial presente em diversos aplicativos bancários ou governamentais, ou ainda como método complementar ao diagnóstico médico como na área de CAD (*Computer-Aided Diagnosis*), em português, diagnóstico auxiliado por computador.

Dentro da área de CAD, especificamente voltado a estudos relativos ao câncer de mama, existem múltiplos softwares e algoritmos disponíveis para incremento do diagnóstico. Em (RANGAYYAN; AYRES; DESAUTELS, 2007) estimou-se em 30% a 70% a redução da taxa de mortalidade ao utilizar programa de computador para interpretação de imagens de mamografia, ou ainda em (MOHEBIAN et al., 2017) é proposto uma técnica para cálculo de reincidência de câncer de mama num período de 5 anos em pacientes, onde foi obtido o índice de 19.3%. De maneira mais ousada, existem ainda programas de computador os quais conseguem automatizar todo um processo ou proferir diagnósticos sem intervenção humana, como acontece em (TIONG et al., 2018), ou em (MOUELHI et al., 2018), respectivamente – ambos descritos na Seção 2.

É perceptível na literatura a reincidência, ou a tendência, a se utilizar determinados métodos para solução de um problema pontual, mais especificamente a detecção e contagem automatizada de células cujo tema é alvo deste trabalho. Nesta seção serão abordados alguns dos princípios chave para que o presente trabalho se tornasse possível. Adiante serão discorridos tópicos relativos à técnicas de suavização de imagens, operações de pré-processamento, segmentação de regiões de interesse e detecção de contornos e formatos geométricos.

3.2 Suavização

No contexto de processamento de imagens, a suavização é o ato de desfocar uma determinada imagem para que haja a remoção de ruídos ou a atenuação das bordas de contornos de objetos. Matematicamente, aplica-se uma operação de convolução entre a imagem e uma pequena matriz $(n \times m)$, denominada *kernel* com filtro passa-baixa que percorrerá toda a imagem removendo regiões de alta frequência.

A convolução é uma operação sobre a imagem, na qual uma máscara (matriz $n \times m$) que contém uma função é rotacionada em 180° e caminha linearmente sobre a imagem realizando operações de soma de produtos a cada deslocamento (GONZALEZ; WOODS, 2009) gerando então uma imagem final. A função contida pela máscara é fator determinante para o tipo de suavização resultante. Popularmente destacam-se 3 funções utilizadas para tal, das quais tem-se a suavização por média aritmética espacial, suavização por mediana espacial e a suavização Gaussiana.

Abstendo-se do princípio matemático por trás de cada conceito, de maneira geral a filtragem por média aritmética espacial tem como finalidade uma filtragem mais genérica, pois mantém os objetos de interesse destacados, porém levemente borrados, e incorpora pequenos ruídos ao fundo da imagem, fazendo com que desapareçam em muitos casos. A suavização por mediana, assim como sugerido por sua nomenclatura e demonstrado em (GONZALEZ; WOODS, 2009): "substitui o valor de um pixel pela mediana dos valores de intensidade na vizinhança desse pixel". É muito utilizado para a remoção de ruídos aleatórios e ruídos do tipo *salt and pepper*, que consistem na presença de pequenos ruídos pretos e brancos sobre toda a imagem.

Na suavização Gaussiana, a função principal utilizada pelo *kernel* é a curva de Gauss, a qual superficialmente tem por objetivo aumentar a área sob a curva do sinal processado, e manter a amplitude do mesmo. Esse tipo de suavização é útil para imagens com pouca iluminação e alto desvio padrão nos pixels (grande quantia de ruídos). Pelo efeito produzido de preservação de bordas ao manter a amplitude do sinal em regiões com mudança abrupta de intensidade com relação aos vizinhos em objetos de interesse na imagem, esse tipo de filtro é popularmente utilizado antes de algoritmos de detecção de contornos, ou detecção de bordas, elucidado na Seção 3.5.

A seguir são apresentadas as diferenças entre os métodos citados acima. Na Figura 1, tem-se a comparação entre a suavização utilizando funções de média e mediana para remoção de ruídos, e na Figura 4.c há a utilização do filtro Gaussiano sob a imagem original em 4.a para o mesmo fim.

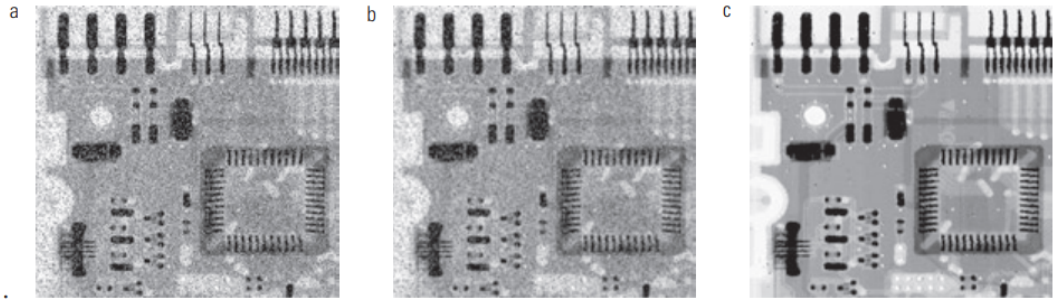

Figura 1 – (a) Imagem de raios X de uma placa de circuito corrompida pelo ruído do tipo *salt and pepper*. (b) Redução de ruído com um filtro de média 3 × 3. (c) Redução de ruído com um filtro de mediana 3 × 3. Fonte: (GONZALEZ; WOODS, 2009).

3.3 Operações de Pré-Processamento

Por pré-processamento entende-se como o uso de técnicas introdutórias ao processo de extração e interpretação de informações relevantes na imagem cuja finalidade é melhorar sua qualidade ao remover ruídos ou ajustar diferenças de iluminação, por exemplo. Dentre as técnicas principais duas categorias possuem destaque, as operantes no domínio de frequência, e as que agem em domínio espacial, sendo muito comum a combinação de ambas para atingir resultados satisfatórios.

Métodos que lidam no domínio de frequência apoiam-se no uso de filtros que atuam sobre o espectro de cores, assim como a técnica de suavização descrita na Seção 3.2. Já os métodos que trabalham com domínio espacial geralmente utilizam-se da morfologia da imagem para operações como abertura ou fechamento de regiões da imagem.

Formalmente, os termos citados são definidos da seguinte maneira de acordo com (GONZALEZ; WOODS, 2009): "A abertura geralmente suaviza o contorno de um objeto, rompe os istmos e elimina as saliências finas. O fechamento também tende a suavizar contornos, mas, ao contrário da abertura, geralmente funde as descontinuidades estreitas e alonga os golfos finos, elimina pequenos buracos e preenche as lacunas em um contorno". Desta forma, a Figura 2(a) utiliza-se da abertura como método auxiliar para a remoção de

pequenos ruídos aleatórios sob a imagem, preservando o contorno. Já na Figura 2(b), o fechamento morfológico é aplicado com o intuito de preencher buracos minúsculos.

(a) Abertura morfológica (b) Fechamento morfológico

Figura 2 – Operações de pré-processamentos sobre a imagem. Fonte: (OPENCV, 2020).

3.4 Segmentação da Imagem

A segmentação de imagem é baseada na homogeneidade (características comuns a um grupo de pixels) e continuidade do sinal (SOBIERANSKI; COMUNELLO; WANGENHEIM, 2010). Tais critérios, quando analisados em conjunto, permitem a identificação de fronteiras, as quais possibilitam que uma determinada região seja discernida na imagem. A seguir é apresentada a técnica de segmentação de imagem utilizada para o presente trabalho.

3.4.1 Binarização

A binarização é o procedimento em que a imagem é transformada de maneira a possuir somente duas cores em sua representação, sendo esta uma classificação bimodal nas cores preto e branco. A nomenclatura atribuída a esse processo é *thresholding* ou limiarização, cuja função principal é segmentar uma imagem de entrada *grayscale* ou escala de cinza, de acordo com a distribuição do espaço de cores. Inúmeros algoritmos são disponibilizados para este fim, porém há destaque especial para 4 tipos de abordagens: *thresholding global* (limiar global), *adaptative Mean thresholdind* (Limiar médio adaptativo), *adaptative Gaussian thresholdind* (limiar adaptativo Gaussiano) e binarização de Otsu.

Em linhas gerais a limiarização global, ou então limiarização simples, consiste em verificar iterativamente se um pixel é maior ou menor que o limiar de intensidade fornecido, de maneira que caso o pixel tenha intensidade maior que o limiar definido seu valor é setado em 255 (branco), caso contrário, o valor é setado em 0 (preto). A intensidade pode ser descrita como o quão luminoso é o pixel. Em imagens do tipo tons de cinza, há somente uma representação de cores, que varia de acordo com a luminosidade do pixel, onde 0 representa a cor preta, sem luminosidade e 255 representa o branco. A Figura 3.b ilustra a utilização desse tipo de binarização, onde o limiar pré definido é de intensidade 127.

Em limiarizações do tipo adaptativa, como limiar adaptativo Gaussiano ou limiar adaptativo médio, diferente da limiarização global, estes não necessitam que a intensidade seja pré-definida, sendo muito úteis para casos onde a imagem tem diferentes condições de iluminação em suas regiões, uma vez que os mesmos calculam o melhor limiar para cada área de acordo com sua função base. Assim como na Seção 3.2, as limiarizações tem como função base, métodos matemáticos seguindo a curva de Gauss, para *thresholding* Gaussiana, e cálculo de média local para *thresholding* médio. Ao analisar a Figura 3, percebe-se na Figura 3.c grande melhora na capacidade de limiarização da imagem original em 3.a se

comparado ao método estático da Figura 3.b, porém é explicito a presença de ruídos aleatórios que podem ser corrigidos ao utilizar a função Gaussiana, assim como na Figura 3.d.

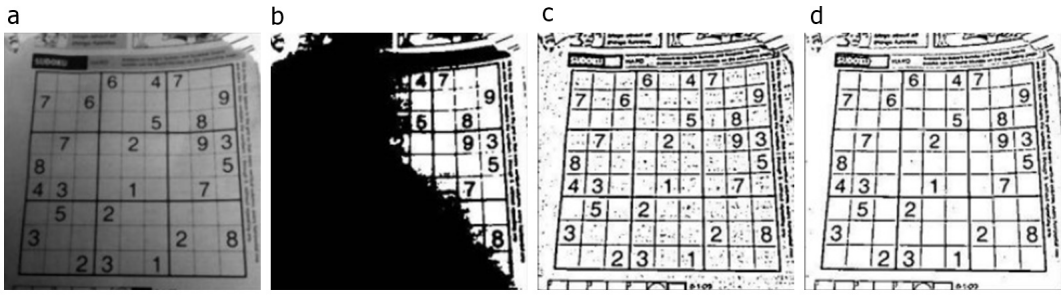

Figura 3 – (a) Imagem original com diferença de iluminação. (b) Limiarização global com intensidade 127 aplicado à imagem original. (c) Limiarização por média adaptativa aplicado à imagem original. (d) Limiarização adaptativa de Gauss aplicado à imagem original. Fonte: (MORDVINTSEV; K., 2013).

Já a binarização de Otsu tem por objetivo o cálculo automático do limite global que é determinado a partir do histograma da imagem. O método matemático por trás do algoritmo não cabe ao escopo deste trabalho, porém, superficialmente, o mesmo retorna um limiar de intensidade ao classificar a imagem em duas classes, plano de fundo ou *background* e primeiro plano ou *foreground*. O limiar é procurado de forma exaustiva de modo que o mesmo minimize a variância entre classes.

Em imagens onde há grande presença de ruídos, vide Figura 4.a, o histograma apresenta variações abruptas em todo o sinal, Figura 4.b, ocasionando erro na binarização por Otsu tal como na imagem 4.c. Neste caso é aconselhável, assim como na ilustração de 4.d, o uso da suavização antes que o limiar seja aplicado, pois as fronteiras dos objetos na imagem serão facilmente identificáveis no histograma em 4.e, permitindo então que o método seja bem sucedido 4.f.

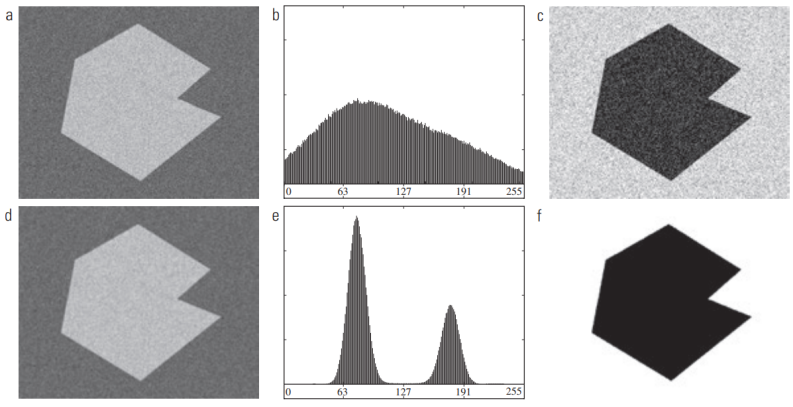

Figura 4 – Comparativo entre histogramas e imagens obtidos a partir da binarização de Otsu. Fonte: (GONZALEZ; WOODS, 2009).

3.5 Detecção de Bordas

Na área de processamento digital de imagens a técnica mais popular de segmentação da imagem é detectar suas bordas. Neste contexto, bordas são fundamentais para mapear os contornos dos objetos de interesse, e pode ser entendida como uma região onde há mudança subida de intensidade.

As bordas podem ser classificadas em três grupos modelados de acordo com a função matemática que melhor a representa. A Figura 5.a representa uma borda do tipo degrau, contida em transições com somente dois níveis de intensidade. Na ilustração 5.b há a borda do tipo rampa, presente em regiões com transição gradual entre cores. Por fim, em 5.c tem-se a borda em forma de telhado (*roof edge*), existente um regiões onde há gradiente bilateral não necessariamente uniforme.

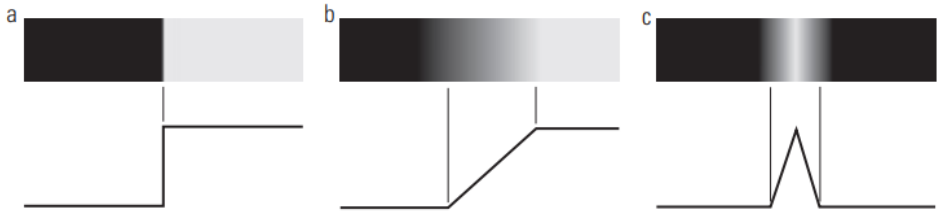

Figura 5 – Tipos de bordas. Fonte: (GONZALEZ; WOODS, 2009)

As imagens anteriores, embora necessárias ao entendimento conceitual do problema, são representações de bordas num mundo ideal, livre de ruídos, característica essa não aplicável em casos reais. Sendo assim, a seguir é descrito o método de detecção de bordas utilizada para este trabalho.

3.5.1 Canny

Possivelmente o algoritmo mais utilizado para a detecção de bordas, Canny foi criado em 1986, sendo desde então um dos métodos mais avançados e de mais alta performance para tal devido a sua complexa formulação matemática a qual tenta encontrar uma solução otimizada com base em três princípios: baixa taxa de erro, boa localização dos pontos de borda e resposta de um único ponto para cada borda. Baixa taxa de erro refere-se a tentar encontrar o ponto de borda o mais similar possível da borda original desejada. Boa localização dos pontos tem como fundamento encontrar o ponto de borda com menor distância do centro da borda original.

A Figura 6 possui em (a) imagem original com plano de fundo propositalmente texturizado e a comparação entre o método matemático de detecção de bordas Canny em relação a outros dois métodos populares que se utilizam de máscaras: Roberts e Sobel. Na Figura 6.b Roberts é utilizado e percebe-se que há grandes quantias de bordas detectadas na imagem, e baixa capacidade de abstração do objeto como um todo. Na Figura 6.c, Sobel é aplicado, embora tenha generalizado a imagem melhor que seu antecessor, é notável que as texturas do plano de fundo foram incorporadas a imagem tornando-se fonte de ruído. Já na Figura 6.d, com a técnica Canny, embora possua contornos abertos, as bordas foram satisfatoriamente identificadas, não havendo interferência da estrutura no plano de fundo.

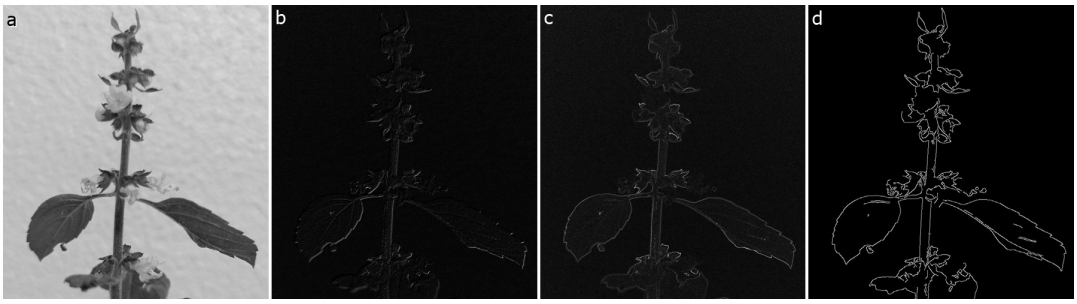

Figura 6 – (a) imagem original, e detecções de bordas com as técnicas Roberts em (b), Sobel em (c) e Canny em (d).

3.6 Detecção de Formas Circulares

Um padrão é algo que se repete de maneira prevista, podendo ser reconhecidos das mais diversas formas, como descritos por fórmulas matemáticas através da análise, ou informalmente através do conhecimento empírico. Em humanos a capacidade de detecção de objetos é adquirida através da vivência, fazendo deste quase sempre um reconhecimento instantâneo. Em máquinas há a necessidade de treinamento, ou criação de sistemas especialistas para que a mesma tarefa seja executa. Antes de qualquer ato relacionado a detecção de objetos por máquinas, um padrão precisa ser estabelecido como parâmetro inicial, podendo ser classificados por múltiplas regiões seguindo um formato geométrico, por objetos com formatos conhecidos e mapeados – como um carro, uma mesa ou uma pessoa–, por uma determinada escala de cores, dentre outras, a depender do objeto de interesse em questão.

Como imagens de células possuem formato correlato a formas circulares, este trabalho irá se limitar a compreensão de métodos que reconheçam esta forma em especial uma vez que os formatos a serem reconhecidos são vastos. Formas circulares são identificadas através da aproximação dos pontos pertencentes a suas curvas a uma equação de segundo grau chamada equação geral da circunferência (Equação 1).

$$(x - x_0)^2 + (y - y_0)^2 = r^2 \tag{1}$$

3.6.1 Excentricidade

Na área de exatas, excentricidade é uma característica associada a quaisquer curvas cuja função matemática obedece um polinômio de segundo grau, que mede seu desvio com relação a uma circunferência.

A excentricidade pode variar de 0 a 1, onde círculos perfeitos possuem excentricidade zero, e elipses muito achatas possuem excentricidade máxima 1. Ao observar a Figura 7, percebe-se que curvas fechadas com alta excentricidade possuem o eixo maior (ma) com comprimento muito superior ao eixo menor (me), e já em contornos com baixa excentricidade, os eixos possuem valores semelhantes, ou muito próximos.

Computacionalmente é possível calcular a excentricidade de regiões circulares de diversas maneiras: (i) utilizando-se de técnicas extremamente manuais – as quais envolvem mapear os pontos de borda do contorno – (ii) traçar linhas sucessivas em todos os ângulos de maneira a encontrar o maior e menor eixo para em seguida utilizá-los na fórmula

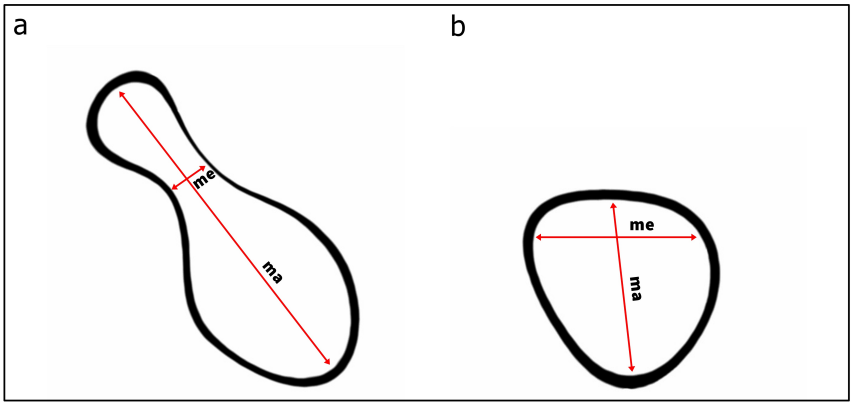

Figura 7 – Ilustração de excentricidade em formas circulares. (a) alta excentricidade. (b) baixa excentricidade.

matemática de excentricidade e obter o resultado. Adicionalmente, podem ser utilizadas bibliotecas de visão computacional, as quais recebem um contorno de entrada (ver técnicas da Seção 3.5) e calculam automaticamente dentre diversos outras métricas a excentricidade da forma circular.

3.6.2 Círculos de Hough

A transformada de Hough foi formulada por Paul Hough em 1962, e tinha como objetivo a identificação de linhas em uma imagem. Uma década após, a transformada foi estendida para compreender além de linhas, círculos e elipses denominando-se círculos de Hough (*Hough Circles*).

Esta técnica é largamente utilizada para a identificação de círculos em imagens que apresentam imperfeições. Os possíveis aspirantes a pontos de borda das circunferências são gerados a partir de uma 'votação'. Esta técnica utiliza dos parâmetros de Hough que é armazenada em uma matriz acumuladora e a qual permite que posteriormente o algoritmo selecione somente os pontos de máximos locais.

Em coordenadas polares, os parâmetros ρ e θ são necessários para a detecção de uma linha, na qual o ρ representa a distância perpendicular de linha da origem em pixels e θ representa o ângulo medido em radianos. Já a detecção de círculos exige as coordenadas bidimensionais do ponto médio do círculo e o raio correspondente. Caso o raio e o centro sejam conhecidos, o algoritmo traça a circunferência e seleciona os pontos de intersecção entre a borda original e a borda segmentada por alguma técnica (ver Seção 3.5) de maneira a encontrar a melhor aproximação possível entre os dois.

De maneira ilustrativa, sempre que um ponto é detectado como possível borda da circunferência, uma unidade é computada na matriz acumuladora. Em pontos de intersecção entre círculos que se sobrepõem, a contagem será maior que 1, permitindo então que o algoritmo os compreenda como formas distintas. Quando não há conhecimento do raio do objeto, a detecção ainda assim se faz possível, pois o algoritmo fixará um centro e estimará iterativamente raios distintos até que a melhor solução seja encontrada.

Embora a teoria apresente uma ideia sólida para detecção de círculos, quando utilizamos a técnica aplicada a um caso real, assim como ilustra a Figura 8, percebe-se dificuldade do algoritmo ao detectar círculos sobrepostos, identificando somente os

contornos mais acentuados em primeiro plano.

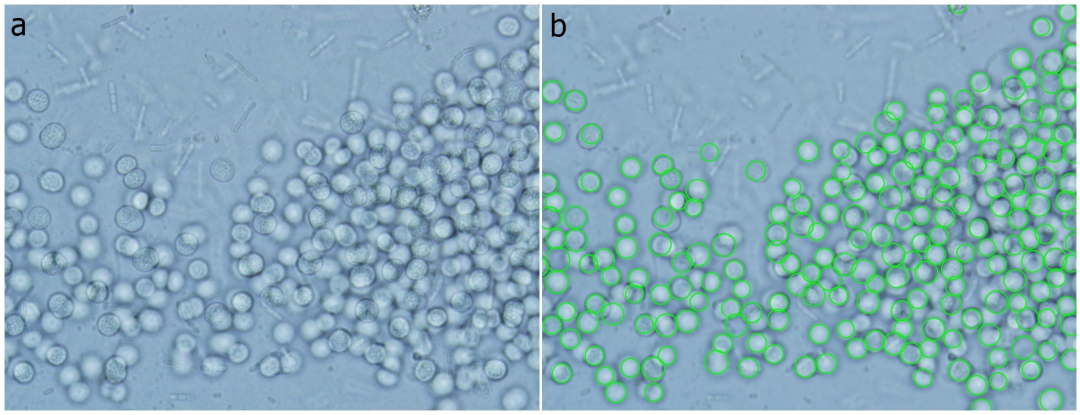

Figura 8 – Detecção de círculos por Hough. Em (a) imagem original, em (b) resultado da detecção. Fonte: (LASKOSKI et al., 2019).

4 Abordagem Proposta

No contexto de imuno-histoquímica para identificação de células cancerosas mamárias, um pigmento é adicionado às amostras. Esta adição de pigmentos, permite caracterizar as células pela diferença de cores entre células positivas - tingidas em tons de marrom, e negativas para câncer - tingidas em tons de azul. Sendo a cor um fator determinante na classificação celular, é de suma importância o refinamento da imagem por seleção de cores antes mesmo do pré-processamento. Em seguida, técnicas descritas na Seção 3 serão aplicadas para solução de diferentes problemas até a conclusão do programa proposto.

O fluxograma utilizado para solução deste trabalho encontra-se descrito na Figura 9, e pode ser agrupado em duas grandes áreas: treinamento e classificação. O algoritmo inicia-se através da etapa de treinamento a qual segmenta a imagem de acordo com uma escala de cores, que pode ser fornecida pelo médico por meio de uma interface, ou utilizar-se de valores pré-definidos, ambos conceitos serão melhor explicados na Seção 4.2. A etapa de classificação terá seus preceitos explicados nas seções que se seguem. Embora mais longa, possui maior contribuição para este trabalho. Nesta etapa é possível refinar as áreas de interesse, torna-las em objetos manipuláveis e operá-las de maneira a obter uma segmentação preliminar satisfatória.

A linguagem utilizada para a construção da interface gráfica e interações com o usuário é o C#, com a utilização de *Windows Forms* para criar a aplicação *Desktop*. Todas as etapas de processamento do algoritmo foram escritas em linguagem *Python*, com o auxílio da biblioteca de visão computacional *OpenCV*, limitando-se a interações externas somente ao receber parâmetros de entrada como escala de cores e imagem, e fornecer as saídas de imagem e resultados do processamento.

As etapas da figura 9 são descritas a seguir.

4.1 Aquisição

A aquisição da imagem servirá como parâmetro de entrada para o algoritmo de detecção. Esta imagem é obtida a partir da captura de imagem da lente de um microscópio

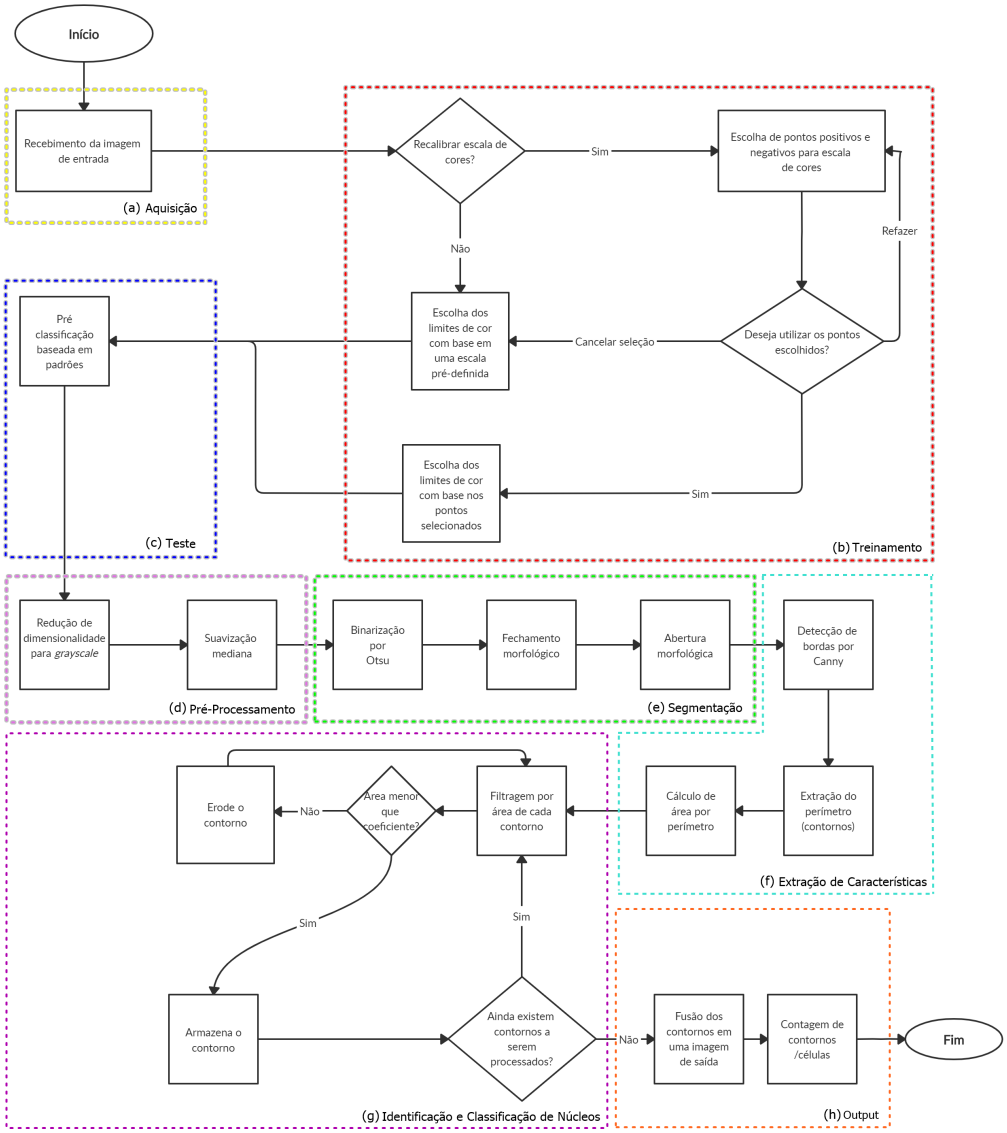

Figura 9 – Fluxograma geral da abordagem proposta para o problema de identificação de núcleos, sumarizado em oito passos principais conforme ilustrado através das linhas pontilhadas.

por meio de uma câmera não profissional cuja lente de observação contém o corte de lâmina desejado. Em seguida, dotado de alguma ferramenta digital de corte de figuras, o médico patologista recorta a imagem de acordo com o local o qual considera ser de *hotspot*, a carrega manualmente para o computador, e a insere no programa.

Durante a captura da imagem da lâmina, a amostra fica suscetível a diversos problemas externos, como não uniformidade de iluminação, *blur* – proveniente da falta de foco da câmera, ou ainda do tremer de mãos do usuário durante a captura –, ou ainda corte da região de *hostpot* contendo partes físicas da lente. Uma abordagem capaz de tratar de maneira generalizada os problemas anteriormente citados podem ser bastante complexos e não cabíveis ao tempo útil para realização deste trabalho, desta forma dentre as amostras recebidas, algumas foram selecionadas (ver Seção 5.2) de maneira a evitar os problemas

citados.

4.2 Treinamento do Modelo

A primeira etapa para a construção do algoritmo é a segmentação das células de acordo com as colorações presentes em seus núcleos. Para isso o programa fornece duas alternativas, a primeira baseia-se no uso de uma escala padrão de cores pré-determinadas definidas por amostragem e a segunda é por meio de uma interface gráfica que auxiliará o patologista na escolha de pontos que irá compor a escala de cores.

A escolha dos valores que constituem a escala pré-definida tanto para células positivas quanto negativas foram obtidas por amostragem, através da coleta e escolha dos pontos de máximos e mínimos, de um conjunto de pontos RGB cuidadosamente selecionados em 35 imagens de carcinoma mamário. Embora a utilização desta escala forneça praticidade por poupar tempo ao profissional, apresenta baixa eficiência na detecção em diferentes tipos de imagens, podendo influenciar negativamente no desenrolar do algoritmo.

Empiricamente, percebeu-se que a boa escolha de pontos para constituir a escala de segmentação possui enorme influência nas etapas seguintes de classificação. Desta forma, fez-se necessário a criação de uma ferramenta que permita ao patologista fazer sua própria seleção de cores. A ferramenta permite que uma imagem seja aberta para a escolha de pontos, em seguida o usuário define quais tipos de células deseja marcar (positivas e/ou negativa) e a quantia de pontos que deseja utilizar, podendo variar de 3 a 10. É recomendado a escolha de pontos visivelmente extremos em pigmentação para cada caso.

As etapas que envolvem o processo de treinamento, bem como os fluxos de decisão para escolha de escala são melhor detalhadas no fluxograma da Figura 9, destacados em vermelho.

4.3 Testing

Obtidos os pontos constituintes da escala de segmentação, a etapa de *testing*, ou teste como ilustrado no fluxograma, tem início com o objetivo de pré-classificar e segmentar os pixels da imagem baseado em padrões, que para esta etapa em questão é relativo a cores.

Este passo possui grande importância para o início do algoritmo, pois permite segmentar a imagem em 3 planos, dos quais consistem em um plano contendo somente células consideradas positivas, um plano contendo somente células consideradas negativas e o plano de fundo. Sendo assim, a cor presente nos pixels será verificada iterativamente de maneira a agrupá-los em um dos planos anteriores.

Ao observar a Figura 10.a, percebe-se que os citoplasmas das n células difundem-se criando um borrão marrom claro ao fundo dos núcleos. Caso a segmentação fosse feita considerando todos os tons de marrom da escala, é perceptível que haveria erro na detecção, uma vez que as células se aglomerariam todas em uma única região. Utilizando a escala pré-definida, ou a ferramenta de recalibragem de escala, o patologista pode precisar com cliques diversos pontos na imagem, de maneira que permita o refinamento da seleção de maneira correta.

A Figura 10.b demonstra o resultado da pré-classificação e segmentação de 10.a com base em uma escala recalibrada.

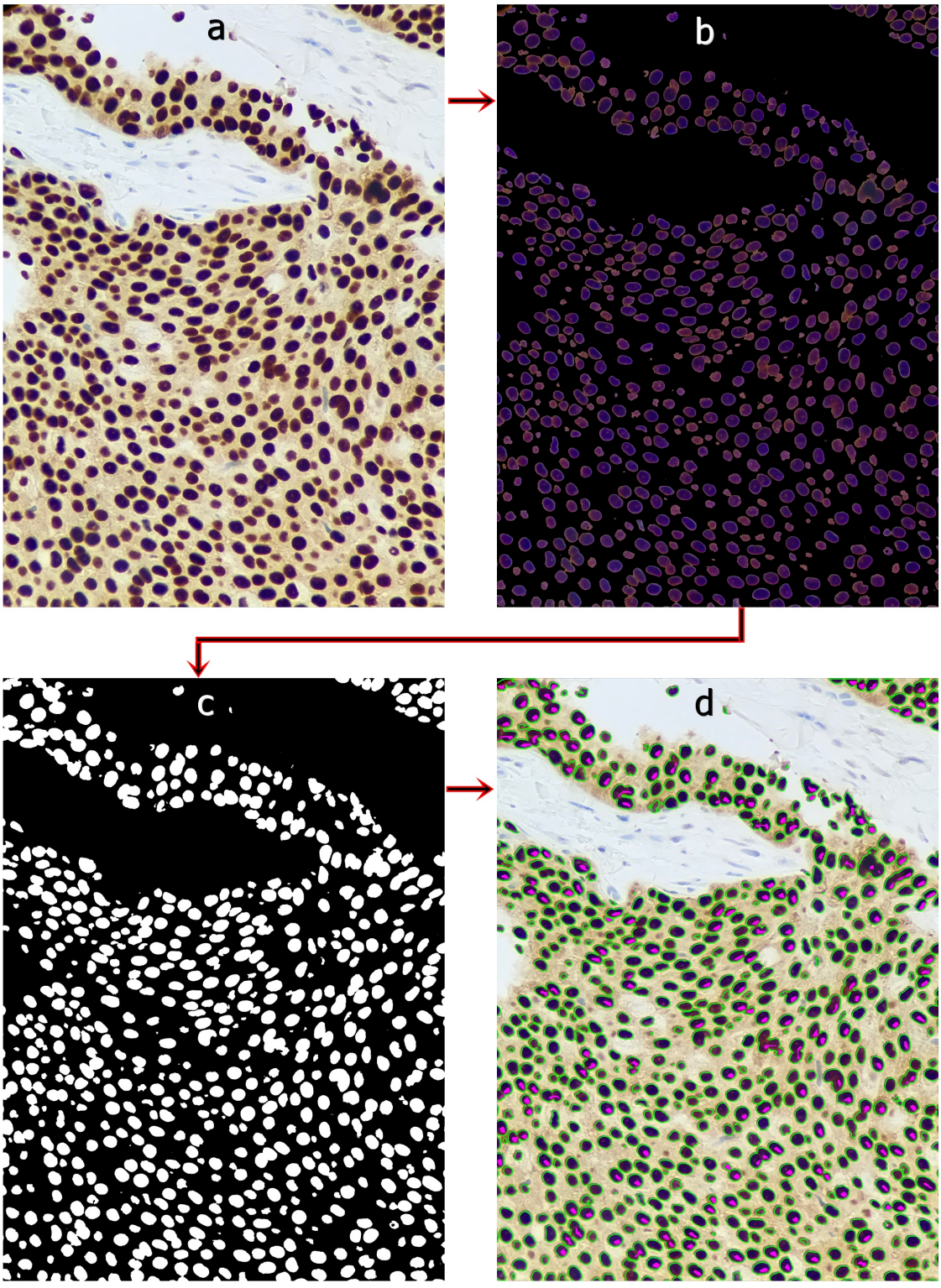

Figura 10 – (a) imagem original. (b) pré classificação de (a). (c) binarização de (b). (d) saída do algoritmo com base na imagem de entrada (a).

4.4 Pré-Processamento

Destacado na cor rosa na Figura 9, o pré-processamento é compreendido por duas etapas: redução de dimensionalidade para *grayscale* e suavização mediana.

A redução de dimensionalidade para tons de cinza, consistem em diminuir o espaço de cores da imagem de saída em *testing*, originalmente RGB, para um único espaço de cores característico por representar somente níveis de luminosidade.

Em seguida, é aplicado suavização mediana com *kernel* quadrado de fator 5 para uniformizar a imagem e reduzir ruídos ínfimos que possam ter sidos agregados a segmentação descrita na seção anterior. Este método de atenuação foi escolhido dentre os demais, pois percebeu-se em testes que possuía esmaecimento mais discreto e menor modificação dos contornos originais. Ao final desta etapa, a imagem deve estar melhorada o bastante para que de fato o processamento seja iniciado.

4.5 Segmentação

Para tratar o problema deste trabalho, a segmentação foi dividida em 2 áreas, sendo uma delas binarização e a outra operações morfológicas com uso de 2 diferentes técnicas. A partir da imagem suavizada obtida anteriormente, será aplicado binarização de Otsu de maneira a limitar mais ainda o espaço de cores da imagem. A Figura 10.c descreve o resultado deste passo, onde percebe-se que houve a uniformização dos diferentes tons apresentados em 10.b.

Embora forneça uma boa visão dos núcleos das células, percebe-se ainda que na Figura 10.c a presença de pequenos ruídos indesejados na imagem, que pode resultar em falso núcleo e afetar negativamente a contagem de células. Em amostras cujas imagens iniciais não possuem núcleos com contornos tão destacados como os demonstrados na Figura 10.a, a presença de ruídos acaba sendo maior. Para solucionar este problema, são aplicados as operações morfológicas. Inicialmente é aplicado o fechamento morfológico com *kernel* 3×3 para preencher pequenos buracos que venham a ocorre no meio das células, dentre tanto acarreta na união de células muito próximas. Dessa forma é aplicado a operação de abertura com *kernel* quadrado de ordem 3, que além de remover ruídos aleatórios na imagem, irá diminuir levemente o formato dos contornos.

Ao final dessa etapa, percebe-se nítida distinção dos núcleos na imagem, os quais já podem ser manipulados individualmente pelas etapas que se seguem.

4.6 Extração de Características

A extração de características é compreendida por 3 estágios cujo objetivo é detectar e segmentar as bordas de cada célula e em seguida calcular alguns atributos espaciais das mesmas.

O algoritmo inicia-se detectando bordas por Canny de maneira a obter somente o esqueleto de todos os contornos inseridos na imagem. Tratando-se de contornos, estes possuem alguns atributos exclusivos que podem vir a ser úteis, como por exemplo classificação de hierarquia de contornos, que com base na teoria de conjuntos avalia quais contornos são pais, e quais são filhos, para casos que há incidência de um ou mais contornos contidos dentro de outro. Embora a operação de fechamento morfológico tenha sido aplicada anteriormente com o intuito de preencher regiões, ainda pode ocorrer a presença de pequenos

buracos no interior das células. Este problema é evitado ao considerar somente o uso dos contornos mais externos de cada células, ou seja os que possuírem hierarquias maior.

Refinadas somente as bordas de interesse, em seguida cada um dos pontos pertencentes aos contornos são mapeados em relação ao espaço vetorial da imagem, armazenados, e então estima-se um a um o valor de pixels de área correspondente com base no perímetro da figura, também em pixels.

Ao fim desta etapa, percebe-se área demasiadamente grande de alguns contornos em relação aos demais. Isso se dá devido a aglomeração de células ocorridos na própria lâmina, ou ainda na suavização ou fechamento morfológico. Por meio de testes constatou-se que regiões agrupadas possuem área maior de 6500 pixels de área para imagens de resolução 12MP (mega pixels), em alguns casos este limite de área também é aplicável para resoluções de no mínimo 1MP. Desta forma, o estágio de segmentação encerra-se ao classificar os contornos em duas categorias: os contornos considerados de núcleos aglomerados e contornos únicos.

Bordas classificadas como possíveis agrupamentos serão plotadas em uma nova imagem de mesma resolução da imagem original, e em seguida passada como parâmetro de entrada para os métodos da seção a seguir, de maneira a resolver a adversidade de núcleos sobrepostos.

4.7 Identificação e Classificação de Núcleos

A identificação e classificação é o passo de maior contribuição por este trabalho, pois ao perceber falhas na segmentação de aglomerações nas imagens do *dataset* por métodos como o *watershed* (SILVA, 2015), criou-se aqui um método iterativo capaz de reduzir os núcleos agrupados de maneira a identifica-los separadamente.

A Figura 9 ilustra de maneira geral a tomada de decisões do algoritmo, a qual consiste em verificar área a área de cada contorno contido na imagem. Os contornos que possuírem área (a) entre dois limites tais que $x < a < y$, são removidos da imagem e armazenados, caso contrário são mantidos e erodidos por uma iteração com um *kernel* quadrado de ordem 9. A quantidade de iterações realizadas dependerá do tamanho de área das aglomerações, desta forma o algoritmo repete-se até que não haja mais contornos na imagem, e ao final, todos os contornos que obedeceram o intervalo de área são plotados em uma única imagem de saída, assim como ilustrado na Figura 10.d.

O fluxograma detalhado de decisões tomadas pelo algoritmo encontra-se na imagem 11. O intervalo de área, ou limites de coeficiente utilizados para comparação, verificam se o pixel possui entre 100 e 1000 pixels de área e foram estabelecidos empiricamente ao perceber tamanho padrão de células não aglomeradas considerando imagens de resolução 3024×4032. Testes demonstraram que estes valores se encaixam mesmo para imagem com resoluções inferiores, como o caso da Figura 13.a com resolução 719×1280.

4.8 Output da Abordagem Proposta

Como etapa final, a saída (*output*) da abordagem proposta possui duas etapas: fusão dos contornos em uma imagem de saída e contagem dos contornos/células.

Como último passo do fluxograma demonstrado na Figura 11, percebe-se que um vetor final F é gerado onde caso houvessem células a serem desagrupadas, o vetor F seria de tamanho maior ou igual a 1 senão possuiria tamanho 0. O vetor F possui as coordenas

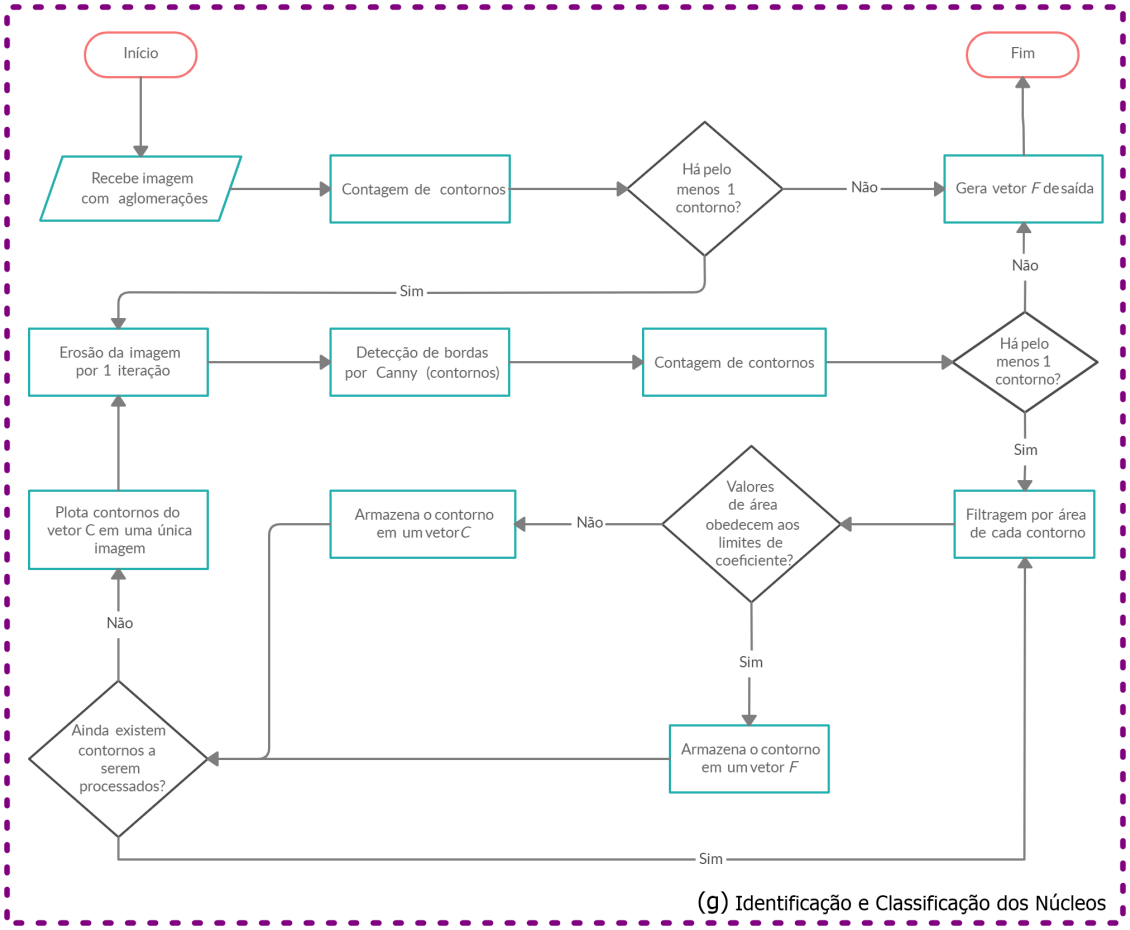

Figura 11 – Fluxograma de Identificação e Classificação de Núcleos obtidos pela expansão da área destacada em roxo da Figura 9.

dos pontos de borda que compõem cada contorno erodido que serão fundidos em uma única imagem de saída com os contornos que não possuíam aglomerações e não foram processados, como descrito na Seção 4.7.

A seguir é realizado a contagem total dos contornos antes e depois da etapa de identificação e classificação de núcleos. Esta etapa tem por objetivo obter a quantidade de células que possivelmente se separaram. Além disso, um *log* completo com diversas métricas referentes ao processamento é gerado. Em um arquivo de dados (.csv) são inseridas informações de resolução da imagem, escala de cores utilizada, tipo de célula em análise (positivas ou negativas), bem como atributos de área, perímetro, momentos, dentre outros, referentes a cada tipo de célula identificada na etapa de extração de características. Por fim, é gerado um arquivo de texto (.txt) com informações relativas ao processo da etapa de identificação e classificação de núcleos. Neste arquivo são inseridos todos os contornos e áreas identificados em cada iteração de erosão, bem como informações se o contorno em questão permanece ou sai do laço de repetição.

Embora o trabalho em questão tenha abordado somente a detecção e análise de um tipo de célula, especificamente células positivas, o processo de identificação para células negativas funciona de maneira similar, mudando somente os limites da escala de cores. Ao fim do algoritmo, deve ser possível a contagem de ambos tipos de células, e estimativa

percentual de uma em relação a outra para análise médica.

5 Resultados Experimentais

5.1 Ambiente Experimental

O *dataset* (conjunto de testes) de imagens utilizadas possui 36 exemplares e correspondem a casos clínicos de carcinoma mamário. As amostras foram fornecidas por Arthur Conelian Gentili, Professor integrante do curso de Medicina da UFSC de Florianópolis, que em conjunto com Marcelo Daniel Berejuck, Professor integrante do curso de Engenharia de Computação da UFSC, identificaram a demanda deste trabalho.

Amostras deste tipo de câncer são classificadas em caráter visual, fazendo-se necessário que os diferentes tipos de células sejam facilmente distinguíveis. Para isso, as amostras são marcadas por receptores hormonais, mais especificamente receptores da proteína estrogênio (ER - *Estrogen Receptor*) do tipo *clone SP1*. O marcador em questão é o anticorpo *Rabbit Monoclonal Primary Antibody* característico por tonalizar o núcleo da célula maligna de carcinomas em marrom (DABBS, 2017), e manter as células negativas com coloração em tons de azul.

O microscópio óptico utilizado para análise da lâmina é o *Primo Star* da marca *Zeiss*® com ampliação de 400 vezes. As imagens que compunham o *dataset* foram retiradas a partir da lente do microscópio óptico por um iPhone® 7 com resolução de 12MP (3024×4032). A ampliação da imagem final é relativa à área de *hotspot* definida pelo profissional.

As configurações de hardware utilizadas para processamento dos códigos são um processador Intel® Core™ i7-5500U CPU de 2.4GHz com 2 núcleos físicos e memória cache de 4MB, memória RAM de 8GB e embora não utilizada, possui placa gráfica NVIDIA® GeForce® 920M com 8GB de memória dedicada.

5.2 Resultados da Detecção

Os resultados da detecção podem ser analisados nas imagens da Figura 12, das quais à esquerda representam imagens originais, e imagens à direita representam as imagens após o processamento.

A primeira saída significativa do algoritmo é destacada em verde, e delimita os contornos totais detectados na etapa de extração de características. Percebe-se boa segmentação dos núcleos em relação ao plano de fundo, baixa presença de ruídos e estimativa satisfatória das bordas celulares. Ainda em verde, é possível observar aglomerações de núcleos em algumas regiões, resultando em um grande contorno envolvendo-as. Desta forma, a etapa de identificação e classificação celular é acionada, fornecendo como saída as marcações em magenta apontando o centro de cada núcleo após o desagrupamento.

A contagem de células presentes na amostra é estimada ao calcular a diferença entre o número de células antes e depois da etapa de identificação e classificação celular, e em seguida somado ao número de células identificadas ao fim da etapa de extração de características.

Estimou-se tempo médio de 20s para execução do algoritmo. Em geral o tempo de execução até a etapa de extração de características é padrão para todos os tipos de amostra

utilizadas, dispendendo maior duração na fase iterativa de identificação e classificação celular a variar de acordo com a quantidade de células aglomeradas.

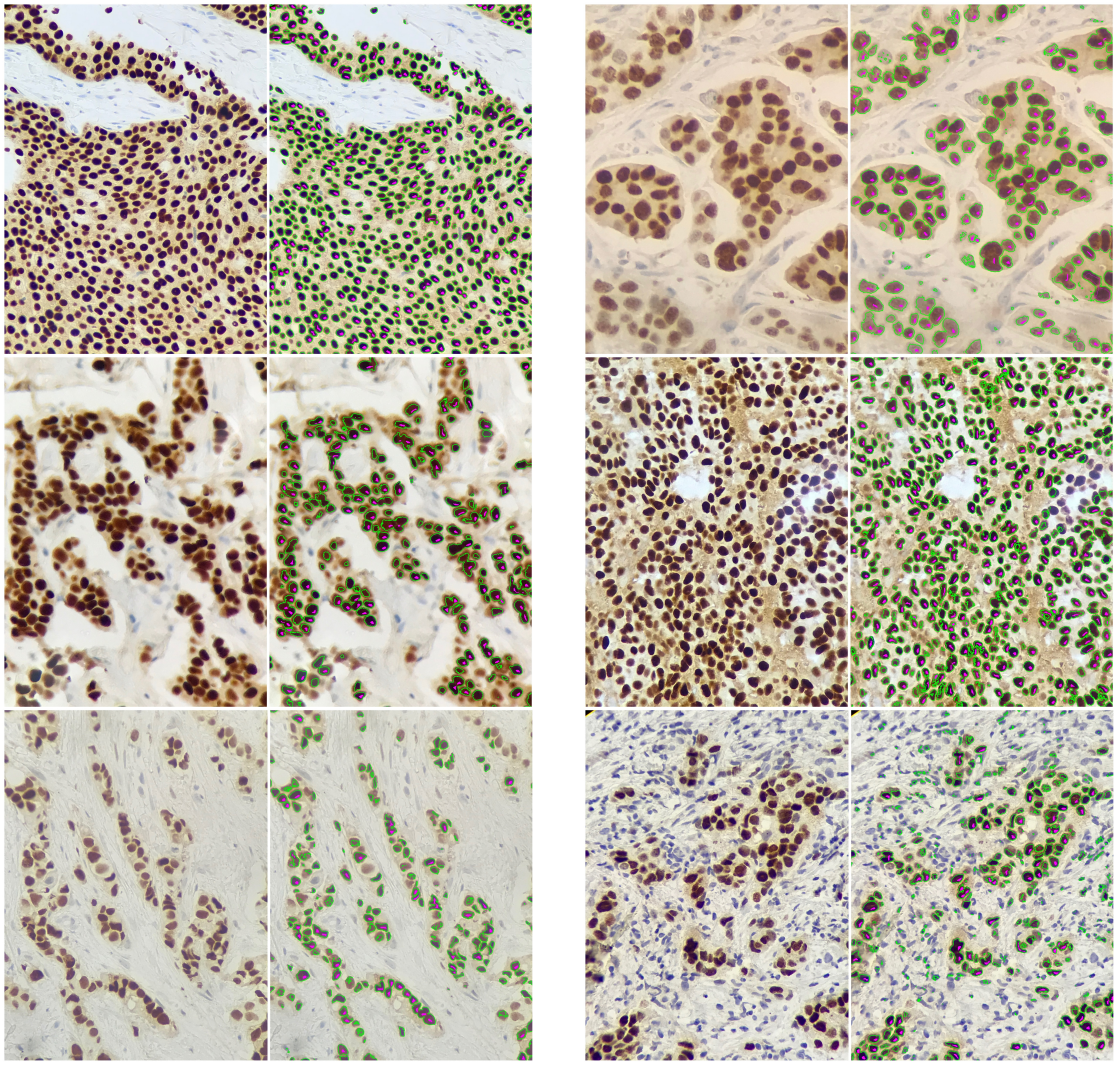

Figura 12 – Resultados da detecção. À esquerda imagens originais, à direita resultados da detecção.

5.3 Validação

Estimou-se precisão de $94,66\%$ para a detecção realizada neste trabalho. Devido a pandemia de COVID-19, a comunicação com o patologista colaborador deste trabalho foi dificultada, logo o fornecimento de amostras marcadas como positivas e negativas não foi possível na totalidade conforme o planejamento inicial. Desta forma, a métrica de precisão obtida pela comparação entre a marcação manual e a do algoritmo, foi estimada com base em uma única imagem, ilustrada na Figura 13.

Além disso, embora não houvessem marcações manuais para estimar com precisão o grau de acerto do sistema apresentado, foi fornecido pelo patologista junto ao conjunto de amostras, uma tabela com o *score Allred* para cada uma das imagens. O *score Allred* tem por objetivo obter a pontuação de carcinoma na amostra numa escala de 0 a 8 de

acordo com o percentual de células positivas (PS - *Proportion Score*) e a intensidade de cor das mesmas (IS - *Intensity Score*). A Tabela 1 exemplifica como a pontuação é realizada, somando-se a pontuação de células positivas presentes na amostra, de 0-5, e a pontuação equivalente de intensidade, de 0-3.

Células Positivas	**Pontuação de proporção**	**Intensidade**	**Pontuação de intensidade**
0	0	Nenhum	0
<1	1	Fraco	1
1-10	2	Intermediário	2
11-33	3	Forte	3
34-66	4		
>= 67	5		

Tabela 1 – Pontuação Allred.

Desta forma, através do índice PS (*Proportion Score*) foi possível comparar visualmente a eficácia do algoritmo ao confrontar os valores deduzidos pelo profissional com o quão tomada de células positivas para câncer uma amostra estava e a quantidade de células identificadas pelo algoritmo.

Na Figura 13.a tem-se a imagem original, e em 13.b há a marcação manual feita pelo patologista com a presença de 75 núcleos. Embora perceba-se distância suficiente entre as células para que não ocorra aglomerações durante as etapas do algoritmo apresentados neste trabalho, a marcação realizada pelo patologista ocasionou em 19 pontos de aglomerações. Estes agrupamentos se dão principalmente pela falta de precisão da ferramenta utilizada para marcação, que para este caso foi o próprio celular. A Figura 13.c representa a detecção de células pelo programa, resultando em 71 núcleos dos quais possuem 0 aglomerações. Em (d) há a sobreposição das marcações feitas pelo médico, em azul e as marcações realizadas pelo algoritmo, em verde.

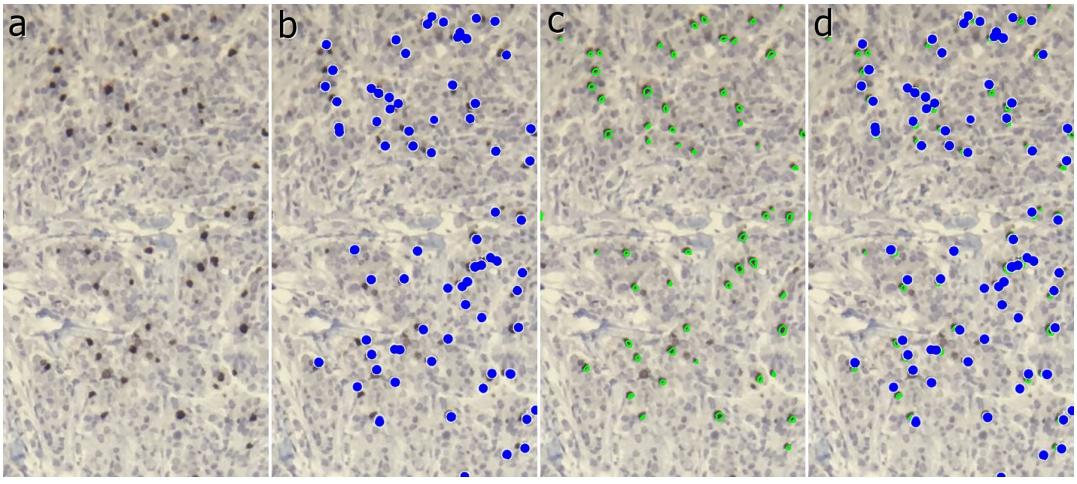

Figura 13 – Comparativo entre imagem marcada manualmente pelo especialista e a detecção realizada computacionalmente pela abordagem proposta. (a) imagem original, (b) marcação manual realizada pelo patologista, (c) marcação realizada pelo algoritmo e em (d) sobreposição das marcações.

6 Conclusões e Discussões

O presente trabalho apresentou um algoritmo que permite a automatização do processo de detecção de células de carcinoma mamário obtidas através de imagens pigmentadas por técnica de imuno-histoquímica. Resultados preliminares demonstraram precisão de 94.66% para o caso de teste utilizado na Figura 13.

Apesar do resultado expressivo, é necessário a implementação de um ambiente mais robusto de testes, onde a acurácia do processamento poderá ser verificada com maior grau de assertividade. Para isso, um *dataset* com maior variedade de marcações manuais se faz necessário. Também é relevante testar o algoritmo em diferentes tipos de biomarcadores, característicos por pigmentarem a amostra em tonalidades diferentes, uma vez que a cor é um parâmetro ajustável de acordo com a escala de cores definida como entrada.

Uma vez que um cenário mais amplo de testes ratificarem a precisão do software apresentado, espera-se como principais vantagens resultantes desta implementação o oferecimento de uma ferramenta de apoio a decisão em diagnósticos médicos, visto que atualmente é estimado em $75,3\%$ a concordância entre profissionais na classificação das amostras. Também é esperado que ocorra a redução no tempo de análise em uma única amostra além da quebra de subjetividade. Não obstante, também é esperado que a ferramenta não se limite ao uso exclusivo de médicos especialistas na área, mas que também possa ser aplicado como material auxiliar de ensino e treinamento em patologia a estudantes.

Da mesma forma, é desejável desvincular o sistema desenvolvido do *desktop* e porta-lo para ambiente *WEB* ou *mobile*, permitindo então o recebimento de *feedbacks* por parte dos usuários com relação a detecção. Trabalhos futuros consideram a utilização das emergentes tecnologias em Redes Neurais Convolucionais para a tomada de decisões em relação ao processo de identificação. Desta maneira ao invés de trabalhar com critérios estáticos, o algoritmo passaria a operar de modo adaptativo, reconhecendo a amostra e lidando com a mesma de maneira única de forma a otimizar a detecção.

Referências

Al-Kofahi, Y. et al. Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering*, v. 57, n. 4, p. 841–852, 2010. Citado na página 5.

AMARANTE, S. *Câncer de mama: a importância do diagnóstico precoce*. 2020. Disponível em: <http://www.iff.fiocruz.br/index.php/8-noticias/274-cancerdemama>. Citado na página 3.

BATTIFORA, H. et al. Estrogen receptor immunohistochemical assay in paraffin-embedded tissue. *Applied Immunohistochemistry*, v. 1, n. 1, p. 39–45, 1993. Citado na página 3.

BRAY, F. et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 68, n. 6, p. 394–424, 2018. Citado na página 2.

CHANG, M. C.; MRKONJIC, M. Review of the current state of digital image analysis in breast pathology. *The Breast Journal*, Wiley Online Library, 2020. Citado 2 vezes nas páginas 3 e 4.

CHOUDHRY, P. High-throughput method for automated colony and cell counting by digital image analysis based on edge detection. *PloS one*, Public Library of Science San Francisco, CA USA, v. 11, n. 2, p. e0148469, 2016. Citado na página 5.

COELHO, G. P. *IMUNO-HISTOQUÍMICA NO CÂNCER DE MAMA*. 2018. Disponível em: <https://www.infomama.com.br/blog/imuno-histoquimica-no-cancer-de-mama/>. Citado na página 3.

DABBS, D. J. *Diagnostic Immunohistochemistry E-Book: Theranostic and Genomic Applications*. [S.l.]: Elsevier Health Sciences, 2017. 1 p. Citado 2 vezes nas páginas 3 e 20.

DPA, D. P. A. *Glossário de termos*. 2012. Disponível em: <https://digitalpathologyassociation.org/glossary-of-terms_1>. Citado na página 4.

ELMORE, J. G. et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, American Medical Association, v. 313, n. 11, p. 1122–1132, 2015. Citado na página 3.

GONZALEZ, R. C.; WOODS, R. C. *Processamento digital de imagens* . [S.l.]: Pearson Educación, 2009. Citado 4 vezes nas páginas 6, 7, 9 e 10.

HOSSEINI, S.; CHEN, H.; JABLONSKI, M. M. Automatic detection and counting of retina cell nuclei using deep learning. *arXiv preprint arXiv:2002.03563*, 2020. Citado na página 5.

LASKOSKI, G. d. A. M. et al. Reconhecimento e contagem automáticos de cianobactérias em água bruta de reservatórios da região de curitiba. In: SBC. *Anais Principais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde*. [S.l.], 2019. p. 258–263. Citado na página 13.

LATTES, P. *Currículo de Arthur Conelian Gentili*. 2020. Disponível em: <http://lattes.cnpq.br/9196132200569809>. Citado na página 4.

LAURO, G. R. et al. Digital pathology consultations—a new era in digital imaging, challenges and practical applications. *Journal of digital imaging*, Springer, v. 26, n. 4, p. 668–677, 2013. Citado na página 4.

LEHR, H.-A. et al. Application of photoshop-based image analysis to quantification of hormone receptor expression in breast cancer. *Journal of Histochemistry & Cytochemistry*, SAGE Publications Sage CA: Los Angeles, CA, v. 45, n. 11, p. 1559–1565, 1997. Citado na página 3.

MOHEBIAN, M. R. et al. A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (hpbcr) using optimized ensemble learning. *Computational and structural biotechnology journal*, Elsevier, v. 15, p. 75–85, 2017. Citado na página 6.

MORDVINTSEV, A.; K., A. *Image Thresholding*. 2013. Disponível em: <https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_thresholding/py_thresholding.html>. Citado na página 9.

MOUELHI, A. et al. Fast unsupervised nuclear segmentation and classification scheme for automatic allred cancer scoring in immunohistochemical breast tissue images. *Computer methods and programs in biomedicine*, Elsevier, v. 165, p. 37–51, 2018. Citado 2 vezes nas páginas 5 e 6.

ONCOGUIA, E. *Biópsia da Mama*. 2020. Disponível em: <http://www.oncoguia.org.br/conteudo/biopsia-da-mama/1390/264/>. Citado na página 3.

OPENCV. *Transformações Morfológicas*. 2020. Disponível em: <https://docs.opencv.org/master/d9/d61/tutorial_py_morphological_ops.html>. Citado na página 8.

PANAGIOTAKIS, C.; ARGYROS, A. A. Cell segmentation via region-based ellipse fitting. In: IEEE. *2018 25th IEEE International Conference on Image Processing (ICIP)*. [S.l.], 2018. p. 2426–2430. Citado na página 5.

RANGAYYAN, R. M.; AYRES, F. J.; DESAUTELS, J. L. A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. *Journal of the Franklin Institute*, Elsevier, v. 344, n. 3-4, p. 312–348, 2007. Citado na página 6.

RICCIO, D. et al. A new unsupervised approach for segmenting and counting cells in high-throughput microscopy image sets. *IEEE Journal of Biomedical and Health Informatics*, IEEE, v. 23, n. 1, p. 437–448, 2018. Citado na página 5.

RUEDEN, C. T. et al. Imagej2: Imagej for the next generation of scientific image data. *BMC bioinformatics*, Springer, v. 18, n. 1, p. 529, 2017. Citado na página 3.

SILVA, B. S. da. *DETECÇÃO AUTOMÁTICA DE CÉLULAS VIA TÉCNICAS DE MORFOLOGIA MATEMÁTICA E PROCESSAMENTO DIGITAL DE IMAGENS*. 2015. Monografia (Bacharel em Engenharia Eletrônica e de Computação), Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil. Citado 2 vezes nas páginas 4 e 18.

SOBIERANSKI, A. C.; COMUNELLO, E.; WANGENHEIM, A. von. *Segmentação supervisionada de imagens pela funcional de Mumford-Shah utilizando métricas de distância não-lineares*. Tese (Doutorado) — Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós . . . , 2010. Citado na página 8.

TIONG, K. H. et al. Quickcount®: a novel automated software for rapid cell detection and quantification. *BioTechniques*, Future Science, v. 65, n. 6, p. 322–330, 2018. Citado 2 vezes nas páginas 4 e 6.