



TRANSFORMADA RIPPLET: UM NOVO MÉTODO PARA QUALIFICAÇÃO DE LESÕES DA MAMA

ARTIGO ORIGINAL

BARBOSA, Paulo Henrique de Jesus ¹

BENITES, Cristiano da Silva ²

BARBOSA, Paulo Henrique de Jesus. BENITES, Cristiano da Silva. **Transformada ripplet: um novo método para qualificação de lesões da mama.** Revista Científica Multidisciplinar Núcleo do Conhecimento. Ano 05, Ed. 05, Vol. 09, pp. 14-36. Maio de 2020. ISSN: 2448-0959, Link de acesso: <https://www.nucleodoconhecimento.com.br/saude/lesoes-da-mama>

RESUMO

Este trabalho mostra Um Novo Método para Qualificação de Lesões da Mama Baseado na Transformada *Ripplet*. O câncer de mama é o tipo de câncer mais comum entre as mulheres no mundo e no Brasil, depois do de pele não melanoma. Com os avanços na geração de bancos de dados de tecidos e de estudos histológicos digitalizados, é possível utilizar padrões de tecido histológico com análise de imagem assistida por computador para facilitar a classificação da doença. O objetivo deste trabalho foi investigar um método de extração de descritores de textura, a partir dos coeficientes da transformada *Ripplet-II*; investigar grupos de descritores, combinando os coeficientes das transformadas com as características estatísticas de segunda ordem; estudar o algoritmo ReliefF na seleção de descritores de textura relevantes para classificação com algoritmo Random Forest; identificar quais as características

¹ Graduação Ciências da Computação.

² Mestrado em Engenharia Elétrica e da Computação. Especialização em MBA em Data Center e Computação em Nuvem - Infraestrutura, Arquitetura. Graduação em Redes de Computadores.



mais apropriadas para classificar os diferentes tipos de lesões de mama; e analisar os resultados e comparação com outras abordagens presentes na literatura. A metodologia deste trabalho foi a escolha de imagens de bases histológicas de lesões para extração das características; preparo das imagens através dos procedimentos de pré-processamento, para melhorar a relação sinal/ruído dessas lesões; aplicação da transformada Ripplet-II para extração de descritores e suas variações em sub-bandas; extração de características estatísticas de segunda ordem manipuladas nos conjuntos de combinações de grupos de características; execução de algoritmo para seleção de características, eliminando assim ruídos e características não relevantes; aplicação dos classificadores para avaliação dos grupos de descritores. Diante dos resultados encontrados na literatura, para identificação de lesões em imagens histológicas de mama, utilizando outras abordagens, muitas delas complexas, pode-se destacar que este trabalho apresentou resultados promissores com o uso dos coeficientes da *Ripplet-II*, numa abordagem simples. Conclui-se que o desempenho do método proposto foi melhor utilizando-se as imagens da base UCSB, fato este comprovado também pelos índices obtidos pelos trabalhos existentes na literatura, ou seja, enquanto as análises com as imagens da base UCSB alcançam patamares superiores a 95% de acurácia, com a BreakHis atinge valores em torno de 90%.

Palavras-Chave: Câncer, mama, classificação de Lesões, *Ripplet-II*.

1. INTRODUÇÃO

O câncer de mama é o tipo de câncer mais comum entre as mulheres no mundo e no Brasil, depois do de pele não melanoma (INCA, 2019). A análise de tecidos histológicos é utilizada no diagnóstico mais preciso dentre os diversos exames realizados para o diagnóstico médico de pacientes com lesões de mama. Para o diagnóstico, amostras de tecidos coradas com HematoxilinaEosina (H&E) são analisadas através de microscópio para identificação de células cancerosas; essas análises são essenciais, pois permitem o acompanhamento da doença, a identificação de seu estágio e orientação ao paciente.



Devido ao desenvolvimento dos sistemas multimídia e o avanço de novas técnicas computacionais utilizadas em sinais bidimensionais, o processamento de imagens digitais é uma área da computação que vem crescendo cada vez mais. Com o auxílio de algoritmos computacionais chamados CAD (Computer Aided Diagnosis), permitem-se pesquisas em diagnósticos e prognósticos. O CAD tornou-se parte da constatação clínica do câncer de mama na triagem e em hospitais nos Estados Unidos da América (TANG et al., 2009). Com os avanços na geração de bancos de dados de tecidos e de estudos histológicos digitalizados, é possível utilizar padrões de tecido histológico com análise de imagem assistida por computador para facilitar a classificação da doença (GURCAN et al., 2009).

O reconhecimento de padrões (quantitativa ou estrutural) nas imagens histológicas é uma das fases fundamentais de um sistema CAD. Um padrão é formado por um ou mais descritores, ou seja, características, que podem ser extraídas em processamento de imagem: as morfológicas (extremidades da imagem como bordas e margens) e as não morfológicas (características internas, os pixels que constituem a região). A representação externa é utilizada quando o objetivo é a apresentação da forma e a interna quando o objetivo do problema está na análise de textura (AZEVEDO e CONCI, 2008).

As características não morfológicas são consideradas problemáticas devido à dificuldade de interpretação das variações de níveis pelo olho humano. Através das informações da textura da imagem é obtida uma classe de atributos usada para a representação dessas características. Para obtenção de informações de textura, são utilizados métodos como funções de autocorrelação, métodos baseados em modelos, dimensão fractal, filtros baseados em processamento de sinal e muitos outros. Dentre esses métodos, estudos vem demonstrando as vantagens da aplicação de técnicas baseadas na transformada *Wavelet* na avaliação de lesões em mamogramas (DANTAS et al., 2009, JACOMINI et al., 2012; RAMOS et al., 2012).

Através do uso de partes da imagem que sejam importantes para o seu reconhecimento, a representação esparsa é altamente necessária numa visão computacional e no processamento de imagens. Os métodos como a transformada



de Fourier, a transformada *Wavelet* e a transformada *Ridgelet* são propostos na literatura, porém, eles não conseguem resolver as descontinuidades bidimensionais (2D) ao longo de qualquer curva em uma imagem; no entanto, Xu e Wu em 2010, propõe resolver a questão das descontinuidades 2D ao longo das bordas considerando a diferença significativa nas descontinuidades presentes na região de lesões benigna-maligna e sua transformação circundante, a *Ripplet-II*; com os coeficientes *Ripplet-II*, diversas características de textura estatística de primeira e segunda ordem são provenientes da imagem transformada *Ripplet-II* (XU e WU, 2010).

No estudo feito em 2017 por Rabidas, Chakraborty e Midya, foram analisaram as lesões de mama com a transformada *Ripplet-II*, para classificá-las como benignas e malignas; entretanto, este estudo demonstrou que ainda há muito que se fazer na elaboração de descritor de textura a serem consideradas para resultados que comprovem verdadeiramente as vantagens da *Ripplet-II* em relação às outras transformadas.

O desenvolvimento de sistemas computacional está relacionado com a definição de algoritmos de extração de características para imagens, pois está diretamente relacionado com os resultados obtidos para aplicação dessas ferramentas. Sendo assim, nesse trabalho são apresentados uma nova abordagem de descritores que investiga a transformada de *Ripplet-II*, seus coeficientes combinados com diferentes características estatísticas de segunda ordem, derivadas da imagem, para identificação das lesões benignas e malignas em imagens histológicas de mama.

Com a exploração deste trabalho, pretende-se: investigar um método de extração de descritores de textura, a partir dos coeficientes da transformada *Ripplet-II*; investigar grupos de descritores, combinando os coeficientes das transformadas com as características estatísticas de segunda ordem; estudar o algoritmo ReliefF na seleção de descritores de textura relevantes para classificação com algoritmo Random Forest; identificar quais as características mais apropriadas para classificar os diferentes tipos de lesões de mama; e analisar os resultados e comparação com outras abordagens presentes na literatura.



2. FUNDAMENTAÇÃO TEÓRICA

Expõe-se a seguir os princípios conceituais que sustentam todo o processo de análise de imagens histológicas digitais de lesões de mama, desde a coleta e preparação do material da biopsia, passando pela montagem das lâminas físicas com as amostras, digitalização das mesmas, até chegar à etapa de reconhecimento das lesões.

2.1 CÂNCER DE MAMA

Câncer é definido como o crescimento desordenado de células que invadem os tecidos e órgãos, podendo espalhar-se pelo corpo, estas células tornam-se agressivas e incontroláveis, formando tumores ou neoplasias malignas. As causas do câncer podem ser externas (meio ambiente e aos hábitos ou costumes próprios de um ambiente social e cultural) ou internas (geneticamente pré-determinadas) ao organismo, estando ambas inter-relacionadas. A maioria dos casos de câncer é relacionada ao ambiente em que o indivíduo está inserido, porém, um grande percentual está ligado a alterações herdadas que aumentam a pré-disposição de se desenvolver a doença. O histórico familiar e testes genéticos permitem a identificação em indivíduos que possuem alto risco de desenvolvimento de câncer (INCA, 2019).

A incidência do câncer de mama aumenta a partir dos 35 anos, sendo raro antes dessa idade. A detecção do câncer de mama pode ser feita pelo exame de palpação, também pela mamografia, há também a biópsia realizada por meio de tecidos histológicos, que é uma das formas mais adequadas para verificar se a neoplasia está realmente presente em um tecido (INCA, 2019).

2.2 HISTOLOGIA

A Histologia é o estudo dos tecidos biológicos e dos tecidos plasmáticos de animais e plantas, sua formação, estrutura e função. Uma disciplina essencial das áreas de ciências biológicas e da saúde e outras áreas correlacionadas, o estudo dos tecidos é realizado por meio da preparação de lâminas histológicas. É também um componente na medicina diagnóstica para estudar a patogênese e os processos do



câncer. O objetivo principal no diagnóstico é examinar o tecido e fornecer um diagnóstico correto para a melhor intervenção terapêutica do paciente.

As lâminas de histopatologia do tecido podem ser digitalizadas e armazenadas em forma de imagem digital, assim tornou-se passível de aplicação de técnicas de análise de imagem computadorizada e aprendizagem automática. Com o objetivo de se conservar as propriedades do tecido e evitar que o mesmo se deteriore, o processo da preparação do material extraído de uma biopsia, até se transformar em uma imagem digitalizada da lâmina de vidro com o tecido fixado, passa por diversas etapas, sendo elas: fixação; desidratação e diafanização; inclusão; microtomia; e montagem e coloração. A digitalização da lâmina histológica é realizada através do uso de scanners digitais ou microscópios equipados com câmeras (TOSTA e NASCIMENTO, 2016).

A técnica de digitalização das lâminas físicas para o desenvolvimento de sistemas computacionais baseados nessas imagens, possibilitam o desenvolvimento de algoritmos capazes de permitir a obtenção de diagnósticos e prognósticos para auxílio a especialistas. Essa abordagem computacional proporciona o alcance de rápidos resultados automáticos por meio da utilização de técnicas relacionadas ao processamento digital de imagens aliadas ao conhecimento histológico (TOSTA e NASCIMENTO, 2016).

2.3 SISTEMAS COMPUTADORIZADOS DE AUXÍLIO AO DIAGNÓSTICO

Os sistemas de diagnóstico auxiliado por computador (*Computer-aided Diagnosis – CAD*) deve ser utilizado somente como uma ferramenta para obtenção de informação adicional, sendo o diagnóstico final sempre feito pelo especialista. O CAD tem a finalidade de melhorar o diagnóstico, a resposta do computador pode ser útil, uma vez que o diagnóstico do especialista é baseado em avaliação subjetiva, estando sujeito a variações como a perda de informação devido à natureza sutil do achado patológico, baixa qualidade da imagem, sobreposição de estruturas, fadiga visual ou distração. A proposta do CAD é funcionar como um segundo especialista.



O sistema CAD aplica-se no auxílio à detecção de lesões, a partir da localização de padrões anormais através da varredura da imagem pelo e também no auxílio ao diagnóstico, através da quantificação de características da imagem e sua classificação como correspondendo a padrões normais ou anormais.

Os sistemas CAD utilizam-se de técnicas provenientes de duas áreas do conhecimento: visão computacional (envolve o processamento de imagem para realce, segmentação e extração de atributos) e inteligência artificial (inclui métodos para seleção de atributos e reconhecimento de padrões). O sistema CAD pode ser aplicado em diversas modalidades de obtenção de imagem, incluindo radiografia convencional, tomografia computadorizada, ressonância magnética, ultra-sonografia e medicina nuclear. Também é possível produzir esquemas de CAD para vários tipos de exame das partes do corpo, como crânio, tórax, abdômen, osso e sistema vascular, entre outros. Entretanto, sistemas CAD tem três objetos primordiais de pesquisa: as áreas de mamografia, para detecção do câncer de mama; tórax, para detecção de nódulos pulmonares, lesões intersticiais e pneumotórax; e angiografia, para a análise quantitativa de estenoses e de fluxo sanguíneo (DOI et al., 1999).

2.4 PROCESSAMENTO DE IMAGENS DIGITAIS

De acordo com Pedrini e Schwartz (2008), as técnicas de processamento de imagens digitais são base para um sistema CAD e se desenvolvem em cinco fases:

1. Aquisição da imagem: é a captura a imagem por meio de um dispositivo ou sensor, os principais dispositivos utilizados são as câmeras digitais, digitalizador, aparelhos de radiografia digital e tomógrafos médicos.
2. Pré-processamento: a imagem digital pode apresentar imperfeições ou degradações decorrentes das condições de iluminação ou características dos dispositivos utilizados para obter as imagens. Sendo assim, a etapa de pré-processamento tem como função melhorar a qualidade das características da imagem. As principais técnicas de realce de imagens se dividem em duas categorias: métodos no domínio espacial (refere-se ao próprio plano da imagem e a *Contrast Limited Adaptive Histogram Equalization* (CLAHE) é uma



técnica no domínio espacial utilizada para realçar o contraste de imagens histológicas) e métodos no domínio de frequência (baseadas na modificação da imagem).

3. Segmentação: a imagem passa por um processo de identificação e separação das regiões de interesse ou objetos, que se baseiam na detecção de descontinuidades (bordas) ou de similaridades (regiões) dos valores de intensidade do pixel, tendo o intuito de separar fundos e objetos da imagem.
4. Representação e descrição: extração de características ou propriedades que discriminam as classes de objetos contidas na imagem. Como essa etapa coleta informações específicas da imagem, a mesma se torna um ponto crucial para um bom desempenho dos algoritmos de classificação. Há dois tipos de características: as morfológicas (representam as extremidades da imagem como bordas e margens) e as não morfológicas (representam as características internas da imagem).
5. Reconhecimento e interpretação: reconhecimento ou classificação, onde um identificador é atribuído às imagens de acordo com as informações coletadas previamente. As características extraídas são interpretadas e padrões são gerados de acordo com cada imagem. A base de conhecimento gerada pode variar de tamanho e complexidade, dependendo do sistema em que foi empregada a análise das imagens.

2.5 ESTATÍSTICAS DA MATRIZ DE CO-OCORRÊNCIA DE NÍVEIS DE CINZA (GLCM)

O método da matriz de co-ocorrência de níveis de cinza (*Grey-level Cooccurrence Matrix* - GLCM) é uma das técnicas muito utilizadas na extração de características de textura em imagens médicas e com o passar dos anos diversas variações têm sido propostas quanto a forma de se calcular as matrizes de co-ocorrência (HARALICK; SHANMUGAM E DINSTEIN, 1973).



A construção da GLCM não é suficiente para dar uma assinatura a uma determinada matriz, é necessário extrair informações de propriedades texturais. Haralick; Shanmugam e Dinstein (1973) propuseram 13 medidas estatísticas para esse fim.

3. CLASSIFICAÇÃO DE IMAGENS HISTOLÓGICAS DE MAMA

A classificação de imagens histológicas de lesões de mama utiliza diversos algoritmos de seleção e classificação, com o objetivo de se alcançar uma maior assertividade na distinção dos vários tipos de lesões, em complementação ao diagnóstico resultante da análise da imagem mamográfica.

Para classificação de lesões de mama, foram propostas as seguintes: normais, benignas e malignas, com foco na análise de textura de imagens histológicas e mamográficas, através da extração dos vetores LBP dos descritores da Transformada *Curvelet*. Foi utilizado o algoritmo ANOVA para seleção de características e na etapa de classificação foram utilizados os classificadores *Decision Tree* (DT), *Random Forest* (RaF), SVM e *Polynomial* (PL), com intuito de identificar o que apresentaria melhor desempenho, sendo o classificador PL que apresentou o melhor resultado (BRUNO et. al, 2016).

No estudo de Song et al. (2017), propuseram uma abordagem envolvendo a extração de vários descritores locais da imagem, dentre eles o SIFT (*Scale-Invariant Feature Transform*), o DBN (*Deep Belief Network*) e redes neurais convolutivas (*Convolutional Neural Network*). Com esses descritores, foi gerado um vetor denominado FV (*Fisher Vector*), que incidiu na codificação de características que soma um conjunto de descritores locais em um descritor de alta-dimensionalidade, e que representa as características de nível de imagem. O que apresentou o melhor desempenho foi a combinação de FV com *Convolutional Neural Network*. Ou seja, na combinação simples de descritores foi possível obter melhor desempenho.



3.1 CLASSIFICAÇÃO DE IMAGENS HISTOLÓGICAS COM A TRANSFORMADA *WAVELET*

Por meio de sistema de auxílio ao diagnóstico é realizada a extração de características em imagens digitais de mama, sendo que a transformada *Wavelet* e suas derivações são uma das técnicas mais utilizadas, principalmente nos casos que envolvem o uso da *Wavelet* para análise de imagens histológicas.

No estudo de Krishnan, Chakraborty e Ray (2010), foi apresentado um método baseado na segmentação do epitélio, com extração de características texturais usando transformada *Wavelet*, seleção de características usando divergência de Kullback-Leibler (KL) e a classificação dos dados usando abordagem bayesiana e SVM. Foram consideradas 67 imagens histopatológicas normais e 47 de Fibrose Sub-mucosa Oral (FSO). Cinquenta e sete características foram extraídas, sendo 48 características *Gabor-wavelet* e 9 características *Wavelet* do epitélio. Observou-se que a SVM com função de kernel linear proporcionou melhor acurácia de classificação em relação a bayesiana.

A quantidade de células mitóticas presentes nos tecidos é um parâmetro valioso para examinar e classificar a amostra de biópsia de mama, para assim diagnosticar o câncer de mama. Para detecção automática de mitose, é importante diferenciar as células mitóticas das não-mitóticas em imagens histopatológicas da mama.

Em Wan; Liu e Chen (2014), foi desenvolvido um método para melhorar a exatidão da classificação da mitose através da caracterização de células de tecido em representações de multi-resolução baseadas em *wavelets*. A Transformada *Wavelet* complexa *dual-tree* (DT-CWT) foi executada para decompor os *patches* de imagem em formas de múltiplas escalas. Cinco características estatísticas foram extraídas em cada sub-banda *Wavelet*. Os resultados quantitativos mostraram que a abordagem baseada na DT-CWT obteve desempenho de classificação superior.



3.2 CLASSIFICAÇÃO DE IMAGENS COM A TRANSFORMADA RIPPLET

No estudo de Rabidas, Chakraborty e Midya (2017) foram analisadas as singularidades 2D de lesões de mama e suas regiões circundantes com a transformada *Ripplet-II*, onde foram classificá-las em benignas ou malignas. Tanto as lesões benignas como as malignas podem alterar os padrões de orientação dos tecidos mamários normais. A técnica utilizada combinou os coeficientes da *Ripplet-II* com medidas estatísticas de primeira e segunda ordem, calculadas a partir dos próprios coeficientes da *Ripplet-II*. A partir do banco de dados da *Mammography Image Analysis Society* (MIAS), foram selecionados 58 casos de lesões, incluindo 38 benignos e 20 malignos. Para avaliar o desempenho da *Ripplet-II*, os mesmos experimentos foram realizados com os coeficientes da *Wavelet*. O melhor desempenho foi para a *Ripplet-II*, contra a *wavelet*, para as 58 imagens da base de imagens mini-MIAS. Este estudo demonstrou ainda que há diversas combinações de características a serem consideradas para se obter uma solução que referencie as vantagens da *Ripplet-II* em relação às outras transformadas, como a *Wavelet*.

No trabalho de Nayak, Dash e Majhi (2017) foi desenvolvido um sistema CAD para diagnóstico de imagens de Ressonância Magnética (MRI) do cérebro, como patológicas ou saudáveis. O método proposto foi testado nos três conjuntos de dados: DS-66, DS-160 e DS-255, com 66, 160 e 255 imagens, respectivamente. A *Ripplet-II* Ortogonal é conhecida por formar coeficientes grandes e ao mesmo tempo eliminar ruídos. Os algoritmos aplicados em conjunto foram PCA (*Principal Component Analysis*) e o LDA (*Linear Discriminant Analysis*). Depois da redução das características, dos 65536 descritores obtidos da *Ripplet-II*, apenas 2 características por imagem foram selecionadas, utilizando a combinação PCA+LDA. A conclusão obtida após a classificação, apresentou valores entre 99,61% e 100%.



3.3 CONCLUSÕES DOS ESTUDOS REALIZADOS

Ao analisar os estudos realizados, conclui-se que o câncer de mama é uma das principais doenças que atinge especialmente as mulheres acima de 35 anos. Os estudos que desenvolveram ferramentas de apoio a decisão contribuíram com as análises dos especialistas para melhorar a detecção da doença em seu estágio inicial. Uma das áreas exploradas foi a extração de característica de regiões de interesse. Há diversas abordagens que exploram informações no domínio espacial ou da frequência, que são utilizadas largamente para identificação de características de textura de imagens de mama e que fazem uso da Transformada *Wavelet* em conjunto com outros extratores. A Transformada *Wavelet* é eficiente para identificar características em uma abordagem 1D, dificultando a descrição de bordas e contornos, entretanto, a transformada *Ripplet-II* é uma técnica que pode superar a *Wavelet* na análise das singularidades 2D de lesões e suas regiões circundantes de câncer de mama. Ainda são poucos os estudos que exploram as propriedades da *Ripplet-II* na análise de textura de imagens de mama. Sendo assim, este trabalho contribui com novas abordagens de utilização da *Ripplet-II*, para alcançar uma melhor transparência na identificação de lesões de mama em sistemas CAD.

4. MATERIAIS E MÉTODOS

As etapas dos algoritmos desenvolvidos para a extração, seleção e classificação de anomalias decorrentes de lesões de tecidos histológicos, foram executados da seguinte forma: escolha de imagens de bases histológicas de lesões para extração das características; preparo das imagens através dos procedimentos de pré-processamento, para melhorar a relação sinal/ruído dessas lesões; aplicação da transformada *Ripplet-II* para extração de descritores e suas variações em sub-bandas; extração de características estatísticas de segunda ordem manipuladas nos conjuntos de combinações de grupos de características; execução de algoritmo para seleção de características, eliminando assim ruídos e características não relevantes; aplicação dos classificadores para avaliação dos grupos de descritores.

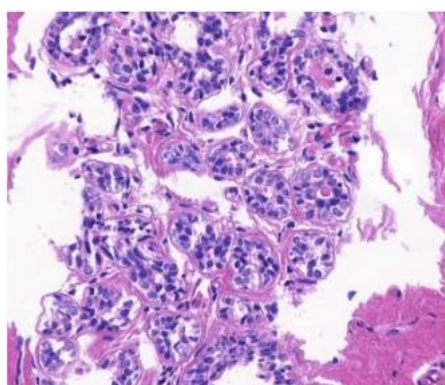
Para este trabalho foram adotadas 2 bases de imagens histológicas, utilizadas na literatura em estudos de classificação de imagens de lesões da mama.

4.1 BASE UCSB

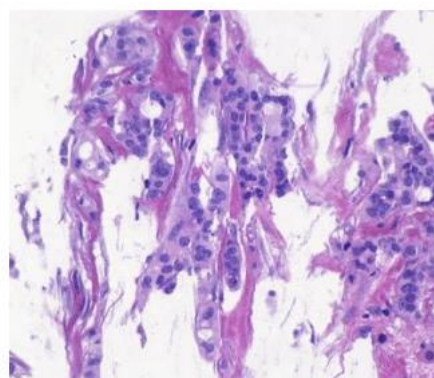
Nesta base foram compreendidas 58 imagens histopatológicas coradas por H&E para detectar células de lesões de mama, do Laboratório de David Rimm em Yale. Separou-se 52 amostras, sendo 26 do tipo maligna e 26 do tipo benigna.

A Figura abaixo apresenta exemplos de imagens de lesões do tipo benigno e maligno utilizadas.

Figura 4.1: Exemplo de imagens histológicas de lesões de mama (a) benigno (b) maligno da base UCSB.



(a)



(b)

Fonte: Gelasca et al., (2009).

4.2 BASE BREAKHIS

A base BreakHis foi composta de 9.109 imagens microscópicas de tecido de tumor de mama coletadas de 82 pacientes usando diferentes fatores de aproximação. Este banco de dados foi construído em colaboração com o Laboratório de P & D - Anatomia Patológica e Citopatologia, da Universidade Federal do Paraná, Brasil.

Dessa base, escolheu-se 2 grupos de pares de lesões, considerando-se as características particulares de cada tipo de imagem. Foram elas: 110 imagens da



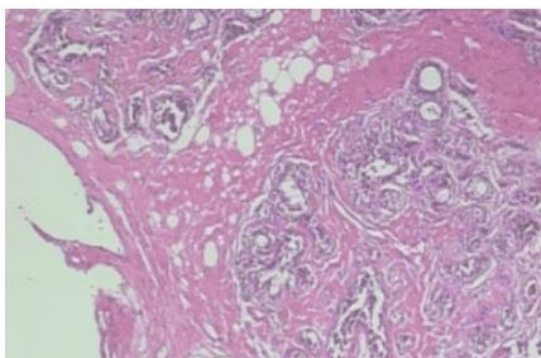
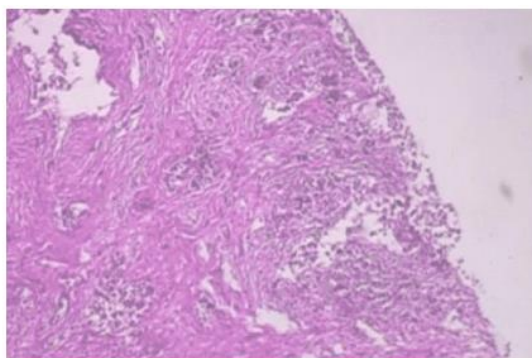
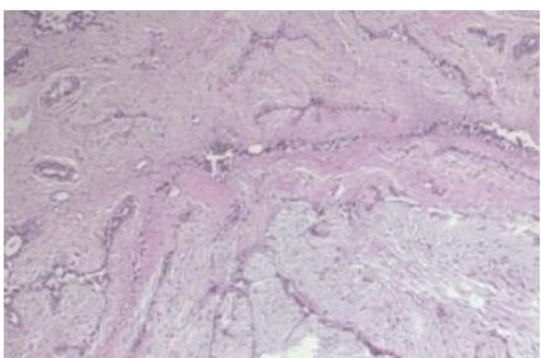
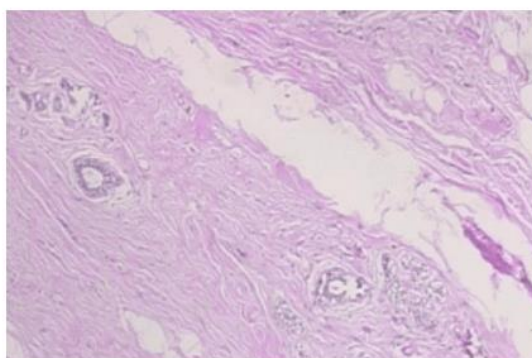
lesão do tipo adenosis (benigna) e 110 da lesão do tipo carcinoma ductal (maligna) - (A x D); 110 imagens da lesão do tipo fibroadenoma (benigna) e 110 da lesão do tipo carcinoma lobular (maligna) - (F x L).

As amostras no conjunto de dados foram coletadas pelo método denominado mastectomia parcial ou biópsia excisional, que comparado a qualquer método de biópsia por agulha, remove um tamanho maior da amostra de tecido e é feito em um hospital com anestesia geral (BreakHis, 2018).

Os grupos de lesões benignas e malignas são compostos por 4 tipos cada: benignas: adenosis (A), fibroadenoma (F), filodes (P) e adenoma tubular (T); malignas: carcinoma ductal (D), carcinoma lobular (L), carcinoma mucinoso (M) e carcinoma papilar (PC).

A Figura 4.2 demonstra uma imagem de cada lesão, de cada um dos grupos, escolhido da base BreakHis.

Figura 4.2: Exemplo de imagens histológicas da base BreakHis, sendo (a) lesão benigna do tipo adenosis, (b) lesão maligna do tipo carcinoma ductal, (c) lesão benigna do tipo fibroadenoma, e (d) lesão maligna do tipo carcinoma lobular.

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

Base: BreakHis (2018).

4.3 SISTEMA COMPUTACIONAL

Para execução dos algoritmos, foi utilizado um computador com processador Pentium, Dual Core 2GHz, com 3 Gbytes de memória RAM e sistema operacional Windows 7 com 32 bits.

4.4 PLATAFORMAS DE SOFTWARE

Todos os algoritmos aplicados foram executados na plataforma MATLAB, versão 2015a. Para as etapas de seleção e classificação das características, foi utilizada a ferramenta WEKA (*Waikato Environment for Knowledge Analysis*), versão 3.7.13.



4.5 MÉTODO COMPUTACIONAL PARA EXTRAÇÃO DOS DESCRITORES

Nesta etapa são descritas as principais técnicas utilizadas no método proposto para extração dos descritores baseado em transformada Ripplet-II e estatísticas de segunda ordem para avaliação quantitativa na base de imagens investigadas.

4.5.1 PRÉ-PROCESSAMENTO DAS IMAGENS HISTOLÓGICAS

Inicialmente, as imagens foram convertidas em escala de níveis de cinza para que as informações presentes nas mesmas pudessem ser extraídas. A conversão em escala de níveis de cinza consiste na eliminação das informações de matiz e saturação da imagem, mantendo a sua luminância.

Em seguida, o filtro CLAHE foi aplicado para aprimorar o contraste e realçar os contornos de regiões de núcleo das células. Com este método, foi possível evitar que a equalização do histograma atinja níveis exagerados de realce, evitando-se ruídos e distorções da imagem.

4.5.2 TRANSFORMADA DISCRETA RIPPLET-II 1D

Nessa etapa, um algoritmo computacional de extração de característica deve ser aplicado sobre as imagens selecionadas na etapa anterior. A Transformada Ripplet-II 1D, foi aplicada a cada ROI. Para extração de características das texturas das imagens, foi utilizada a transformada Ripplet-II 1D discreta com a função mãe Daubechies 4 (DB4).

Nesse método, foram utilizados 3 níveis de decomposição. A primeira decomposição apurou 259 coeficientes Ripplet-II e as subimagens de segunda decomposição apuraram em 133 coeficientes ripplet-II, total de 462 coeficientes para as 6 subimagens de detalhes. Para os grupos de imagens da base BreakHis, o primeiro nível de decomposição da Ripplet-II obteve 131 coeficientes, o segundo mais 69, o terceiro 38, totalizando 238 coeficientes.



4.5.3 EXTRAÇÃO DE CARACTERÍSTICAS ESTATÍSTICAS DE SEGUNDA ORDEM

Para a extração dos descritores estatísticos de Haralick, após obter a ROI com pré-processamento, a matriz de co-ocorrência foi construída. Nessa etapa, as 13 medidas estatísticas de Haralick com 4 ângulos foram aplicadas. Para cada imagem, foram extraídas 52 medidas estatísticas de segunda ordem, sendo 13 para cada uma das 4 direções, consideradas no cálculo da GLCM.

4.5.4 SELEÇÃO DE CARACTERÍSTICAS E CLASSIFICAÇÃO

Nesta etapa, para as 52 amostras da base UCSB, foi utilizado o método *leave-one-out cross-validation*, que é um método de classificação utilizado para realizar validação cruzada, para geração dos conjuntos de dados de treinamento e teste, o qual define partições como conjunto para treinamento, das amostras disponíveis, e apenas uma amostra para teste.

A validação cruzada visa aproveitar o maior número de amostras do conjunto de treinamento para obter tanto a função quanto a estimativa da taxa de erro, apresentando um alto grau de confiança (PEDRINI e SCHWARTZ, 2008).

Com intuito de eliminar ruídos das imagens e características não relevantes para a etapa de classificação, foi utilizado o seletor de características *ReliefF*, que é um método que permite fazer análise multivariada, que envolve parcialmente aprendizado de máquina. Neste estudo o seletor *ReliefF* foi utilizado para obtenção das melhores amostras a serem utilizadas na etapa de classificação.

4.5.5 MÉTRICAS DE AVALIAÇÃO QUANTITATIVA

O desempenho do classificador foi mensurado, através das seguintes taxas:

- Sensibilidade – definida em Verdadeiro Positivo / Total de Positivos;
- Especificidade - definida em Verdadeiro Negativo / Total de Negativos;



- Acurácia - definida em $(\text{Verdadeiro Negativo} + \text{Verdadeiro Positivo}) / \text{Total de Amostras}$.

Uma maneira de representar a relação entre a sensibilidade e especificidade nesse trabalho foi através da métrica da área sob a curva ROC (*Receiver Operating Characteristic* - ROC). Com as medidas de acurácia e área sob a curva ROC pode-se avaliar o desempenho deste método, pois as medidas quanto mais próximas de 100% e 1, respectivamente, tão eficiente é a precisão do método na distinção das imagens de lesões benignas e malignas.

5. RESULTADOS E DISCUSSÕES

Neste tópico é apresentado os resultados obtidos pelos algoritmos computacionais desenvolvidos para as etapas de extração de características, seleção e redução de atributos, a classificação e validação dos métodos propostos.

5.1 PRÉ-PROCESSAMENTO DAS IMAGENS

Na primeira etapa de extração de característica, a imagem foi convertida em escala de níveis de cinza. Em seguida, o filtro CLAHE foi aplicado para melhorar o contraste e realçar os contornos de regiões de núcleo das células.

5.2 MÉTODO DE EXTRAÇÃO DE CARACTERÍSTICAS

Os estudos realizados nos grupos de imagens histológicas utilizaram o vetor de descritores resultantes da extração de 3 níveis de decomposição da transformada Ripplet-II 1D com a função wavelet mãe Daubechies 4. As 52 características estatísticas, extraídas das 4 matrizes GLCM, obtidas para cada ângulo de direção. Esses 2 grupos de descritores foram combinados para formarem o conjunto de características a serem utilizadas nas etapas de seleção e classificação.

Usando como base o p-valor igual a 0,05, aproximadamente 18% das características da base UCSB e em torno de 20% das características dos subgrupos da base BreakHis representam esse valor, ou seja, esses percentuais constituem as



características dentro de todo o conjunto, que podem ser utilizadas para discriminação dos grupos de lesões. A escolha do método *leave-one-out* ocorreu devido ao número de amostras dessa base ser pequena e por utilizar todas as amostras para treino, aproveitando ao máximo o número de amostras do conjunto de treinamento para obter tanto a função quanto a estimação da taxa de erro que apresenta alto grau de confiança (PEDRINI e SCHWARTZ, 2008). Os melhores resultados foram obtidos com o grupo de características da base UCSB, alcançando uma acurácia de 90,38%.

Para os subgrupos da base BreakHis, pode-se notar o resultado de número 6,5, entre falso positivos e falso negativos, ou seja, um percentual médio, 14,8% dos dados testados não conseguiram ser identificados.

5.3 EMPREGO DA SELEÇÃO DE CARACTERÍSTICAS

Na segunda etapa dos experimentos, aplicou-se o seletor *ReliefF*, para eliminar ruídos e outliers, resultantes do processo de extração das imagens. Nessa etapa testes foram realizados para diversas quantidades de características, utilizando-se uma avaliação empírica com subgrupos de apenas 10 características. Esse procedimento foi repetido, com incremento de 10 em 10 características para identificar o melhor ponto de corte, com objetivo de identificar o subgrupo com melhor desempenho na determinação de acurácia.

O ponto de corte para obtenção dos melhores valores com uso do algoritmo *ReliefF* ocorreu com 89 características das 514 do vetor de características, que representa apenas 17,3% do número total de atributos obtidos. Quando essa análise ocorre sobre o grupo de imagens da base BreakHis, observou-se que o ponto de corte para obtenção dos melhores valores com uso do *ReliefF* ocorreu com 274 características, das 290 do vetor de características inicial, representando 94,5% do número total de atributos obtidos.

Os resultados mais significativos foram obtidos com o grupo de características da base UCSB, alcançando uma acurácia de 96,15%. Neste caso, em relação aos resultados



obtidos sem o uso do seletor, obteve-se um avanço em relação ao valor inicial da métrica de 6,4% para a base UCSB, utilizando-se o seletor *ReliefF*.

Nos resultados das métricas de sensibilidade, especificidade, acurácia e valores da matriz de confusão, extraídas da classificação dos grupos de características obtidos das imagens dos sub-grupos da base BreakHis, foram obtidos após aplicação do seletor *ReliefF*, a abordagem apresentada para classificação de lesões histológicas.

Para identificar as características mais importantes nos grupos de amostras de treino e teste, durante o processo de seleção e classificação, calculou-se a frequência com que cada característica esteve presente nos conjuntos de dados. Observou-se que as características extraídas da Ripplet-II 1D influenciaram mais na classificação das imagens da base UCSB e do subgrupo A x D da base BreakHis. O número de características presentes da Ripplet-II 1D foi de 64 e da matriz GLCM foi de 25 de um total de 89 características para UCSB. No caso da BreakHis, o seletor considerou todas as 238 características da Ripplet-II 1D, de um total de 274 características selecionadas, para alcançar o melhor desempenho.

5.4 ANÁLISE DE DESEMPENHO ENTRE RIPPLET-II E WAVELET

A Transformada *Wavelet* foi proposta na literatura para extração de características de textura, porém, não consegue resolver as descontinuidades bidimensionais (2D) ao longo de regiões de curva em uma imagem (XU e WU, 2010). Sendo assim, foram realizados os mesmos experimentos com os descritores das imagens, extraídos da *Wavelet* Discreta 1D, com a finalidade de comparar o seu desempenho em relação a *Ripplet-II*, na identificação dos grupos de lesões.

Os resultados entre a combinação dos descritores da *Wavelet* 1D e a combinação dos descritores da *Ripplet-II* 1D apresentaram diferença em relação ao comportamento das métricas. Os valores de acurácia da *Ripplet-II*, em relação a *Wavelet*, em todos grupos de imagens, foram mais relevantes. Essa diferença fica evidenciada ao comparar os dados da base UCSB, no qual, sem o seletor o desempenho da *Ripplet-*



// foi 4,4% superior a *Wavelet*, e com o seletor *ReliefF* o desempenho da *Ripplet-II* foi 8,7% superior a *Wavelet*.

5.5 ANÁLISE COM ESTUDOS DA LITERATURA

Na literatura existe um grupo expressivo de métodos que são apresentados para investigar o desempenho de propostas de aproximações para classificação de lesões em tecidos histológicos de mama. Esta proposta de aproximação permite a obtenção de valores de acurácia que mostraram relevância diante dos resultados apresentados em outros estudos. Entende-se que a proposta apresentada é pertinente e investigou dados de bases de imagens histológicas de domínio público e pode ser usado em outras bases de imagens na determinação de estratégias de classificação de lesões.

6. CONCLUSÕES

Neste estudo foi exposto uma nova abordagem para descrição de informações de textura baseada em transformada *Ripplet-II*, para avaliação de imagens histológicas de lesões de mama, extraídas de duas bases dispostas na literatura em estudos de classificação de imagens de lesões de mama. Na fase de pré-processamento, as imagens sofreram uma transformação para uma versão em níveis de cinza e Filtro CLAHE de equalização de histogramas. Na fase de extração de descritores, foram captados os coeficientes da *Ripplet-II* 1D Discreta, correspondentes a 3 níveis de decomposição, obtendo-se um conjunto de dados compacto para a fase de reconhecimento e interpretação. Para comparação extraiu-se os mesmos coeficientes das imagens pré-processadas, a partir da *Wavelet* 1D Discreta. A partir das matrizes GLCM das imagens pré-processadas, foram extraídas medidas estatísticas de segunda ordem que contribuíram consideravelmente para a obtenção dos resultados na fase de classificação. Combinando os descritores das transformadas com as características estatísticas da GLCM, formarem-se grupos de dados a serem apresentados para as etapas de seleção e classificação. Os dados de treino e teste, a serem utilizados na etapa de seleção, foram gerados a partir dos métodos *leave-one-out*, que visa aproveitar o maior número de amostras do conjunto de treinamento, apresentando um alto grau de confiança. Além disso, apesar do alto custo



computacional do *leave-one-out*, este foi importante para obtenção dos melhores resultados com as amostras das imagens da base UCSB. De posse dos dados de treinamento e teste, seguiu-se para a fase de seleção, onde o seletor *ReliefF* foi empregado com intuito de eliminar ruídos e outliers, resultantes do processo de extração das imagens. Para os grupos de imagens utilizados neste experimento, os coeficientes da *Ripplet-II* foram predominantes dentro dos conjuntos de características selecionadas pelo *ReliefF*, principalmente nos grupos das imagens da base UCSB e da BreakHis. Em uma verificação das métricas quantitativas de acurácia e área sob a curva ROC, confrontou-se os resultados da *Ripplet-II* com a *Wavelet*, podendo-se observar que a *Ripplet-II* apresenta melhor desempenho que a *Wavelet*, quando o seletor *ReliefF* foi utilizado, alcançando-se 96,15% de acurácia com a *Ripplet-II*, frente a 88,46% da *Wavelet*, para as imagens histológicas da base UCSB. Com as imagens da base BreakHis, o melhor resultado também foi obtido com a *Ripplet-II*, com 86,36% de acurácia, enquanto a *Wavelet* obteve 85,0%. Outro grupo da base BreakHis também obteve um resultado expressivo em relação ao seu desempenho sem o uso do seletor e após o uso do mesmo, passando de 78,18% de acurácia para 85,91% com o *ReliefF*, num avanço de 9,9%. Pode-se constatar que o desempenho do método proposto foi melhor utilizando-se as imagens da base UCSB, fato este comprovado também pelos índices obtidos pelos trabalhos existentes na literatura, ou seja, enquanto as análises com as imagens da base UCSB alcançam patamares superiores a 95% de acurácia, com a BreakHis atinge valores em torno de 90%. Essa diferença de comportamento dos experimentos entre as duas bases utilizadas pode ser atribuída pela diferença na qualidade das imagens, no momento de sua digitalização, ou talvez devido a técnica de biopsia empregada para cada uma delas. Com o uso das duas Transformadas, pode-se afirmar que *Ripplet-II* tem capacidade de representar singularidades 2D de regiões de textura para o contexto investigado, melhor do que as aproximações baseadas em *Wavelet*.

Este trabalho demonstrou uma nova abordagem de descritor de textura, baseada na transformada *Ripplet-II* para classificação de lesões da mama, apoiada por técnicas de preparação de imagens, extração de características, seleção e classificação de dados, largamente utilizados pela literatura, envolvendo poucos elementos para sua



composição. Diante dos resultados encontrados na literatura, para identificação de lesões em imagens histológicas de mama, utilizando outras abordagens, muitas delas complexas, pode-se destacar que este trabalho apresentou resultados promissores com o uso dos coeficientes da *Ripplet-II*, numa abordagem simples. Pode-se constatar neste estudo que o uso de filtros para realce das características da imagem, como o CLAHE, melhora o desempenho para extração das suas características, e que o seletor, *ReliefF*, para eliminação de ruídos e eliminação de características não relevantes, contribuiu também para o desempenho do próprio algoritmo, que tem que processar um número consideravelmente menor de características, em relação a quantidade original, além de proporcionar um resultado promitentes.

REFERÊNCIAS

Azevedo, E., Conci, A., Leta, F. – Computação Gráfica - Vol. 2 – Processamento de Imagens Digitais - Editora Campus 2008.

BreakHis - Breast Cancer Histopathological Database -
<https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/> - acesso em 03/11/2019.

Bruno, D. O. T., Do Nascimento, M. Z., Ramos, R. P., Batista, V. R., Neves, L. A., & Martins, A. S. LBP operators on curvelet coefficients as an algorithm to describe texture in breast cancer tissues. *Expert Systems with Applications*, 2016.55, 329–340.

Dantas, R., Nascimento, M. Z., Ramos, R., e Pereira, D. Análise das variações da matriz de concorrência em imagens derivadas da transformada wavelet haar em mamografia. XIV Congresso Brasileiro de Física Médica, 2009.

Doi K, MacMahon H, Giger ML, Hoffmann KR - Computer-aided diagnosis and its potential impact on diagnostic radiology. Amsterdam: Elsevier Science:11–20, 1999.

Haralick, Robert; Shanmugam, K.; Dinstein, I. Textural Features for Image Classification. *IEEE Transactions on Systems Man and Cybernetics*, SMC-3 No 6, 610–621, 1973.



INCA. Câncer de mama, disponível em: <https://www.inca.gov.br/>, acesso em: 03/11/2019.

Jacomini, R. S., Nascimento, M. Z., Dantas, R., e Ramos, R. Comparison of PCA and ANOVA for information selection of CC and MLO views in classification of mammograms. In Yin, H., Costa, J., and Barreto, G., editors, Intelligent Data Engineering and Automated Learning - IDEAL 2012, volume 7435 of Lecture Notes in Computer Science, pages 117- 126. Springer Berlin / Heidelberg, 2012.

Krishnan, M., Chakraborty, C., & Ray, A. Wavelet based texture classification of oral histopathological sections. Microscopy: Science, Technology, Applications and Education, 897–906, 2010.

Nayak, D. R., Dash, R., & Majhi, B. - Development of pathological brain detection system using Jaya optimized improved extreme learning machine and orthogonal ripplelet-II transform. Multimedia Tools and Applications, pp.1–29, 2017.

Pedrini, H. e Schwartz, W. - Análise de imagens digitais: princípios algoritmos e aplicações. São Paulo: Thomson Learning, 2008.

Rabidas, R., Chakraborty, J., & Midya, A. - Analysis of 2D Singularities for Mammographic Mass Classification. The Institution of Engineering and Technology, Volume 11(Issue 1), 22 – 32, 2017.

Ramos, R. P., do Nascimento, M. Z., and Pereira, D. C. - Texture extraction: An evaluation of ridgelet, wavelet and co-occurrence based methods applied to mammograms. Expert Systems with Applications, 39(12):11036 – 11047, 2012.

Song, Y., Li, Q., Huang, H., Feng, D., Chen, M., & Cai, W. Low Dimensional Representation of Fisher Vectors for Microscopy Image Classification. IEEE Transactions on Medical Imaging, 36(8), 1636–1649. 2017.



Tang, J., Rangayyan, R. M., Xu, J., Naqa, I. E., & Yang, Y. - Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances. IEEE Transactions on Information Technology in Biomedicine, 13(2), 236–251, 2009.

Tosta, T. A. A., Nascimento, M.Z. - Método Computacional para Segmentação não Supervisionada de Imagens Histológicas de Linfoma. Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação. Orientador: Marcelo Zanchetta do Nascimento, 2016.

Xu, J., Wu, D. - 'Ripplet-II transform for feature extraction', Proc. SPIE 7744, Visual Communications and Image Processing, 77441 pp. 77441R-77441R-10, 2010.

Wan, T., Liu, X., Chen, Z. Q. J. - Wavelet-based statistical features for distinguishing mitotic and non-mitotic cells in breast cancer histopathology. Intelligent Computing and Machine Learning Lab School of Biological Science and Medical Engineering No 91 Central Hospital of PLA, Henan. International Conference on Image Processing (ICIP), (61305047), 2290–2294, 2014.

Enviado: Maio, 2020.

Aprovado: Maio, 2020.