Igor Miranda Grangeiro
Orientador: Prof. Dr. Antônio Oseas de Carvalho Filho

Classificação automática de tecidos mamários em malignos e benignos utilizando Índices de diversidade

Picos - PI
30 de maio de 2019

Igor Miranda Grangeiro
Orientador: Prof. Dr. Antônio Oseas de Carvalho Filho

Classificação automática de tecidos mamários em malignos e benignos utilizando Índices de diversidade

Monografia submetida ao Curso de Bacharelado em Sistemas de Informação como requisito parcial para obtenção de grau de Bacharel em Sistemas de Informação.

Universidade Federal do Piauí
Campus Senador Heuvídio Nunes de Barros
Bacharelado em Sistemas de Informação

Picos - PI
30 de maio de 2019

FICHA CATALOGRÁFICA
Serviço de Processamento Técnico da Universidade Federal do Piauí
Biblioteca José Albano de Macêdo

S757c Grangeiro, Igor Miranda.
Classificação automática de tecidos mamários em malignos e benignos utilizando Índices de diversidade. / Igor Miranda Grangeiro. – Picos, PI, 2019.
44 f.
CD-ROM: il; 4 ¾ pol.

Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) – Universidade Federal do Piauí, Picos, 2019.

Orientador(A): Prof. Dr. Antônio Oseas de Carvalho Filho.

1. Processamento Digital de Imagem. 2. Câncer de Mama - Diagnóstico. 3. Descritores de Textura. I. Título.

CDD 005

CLASSIFICAÇÃO AUTOMÁTICA DE TECIDOS MAMÁRIOS EM MALIGNOS E BENIGNOS UTILIZANDO ÍNDICES DE DIVERSIDADE

IGOR MIRANDA GRANGEIRO

Monografia aprovado como exigência parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Data de Aprovação:

Picos – PI, 12 de junho de 2019

Prof. Dr. Antônio Oseas de Carvalho Filho

Prof. Dr. Romuere Rodrigues Veloso e Silva

Prof. Dr. Flávio Henrique Duarte de Araújo

Agradecimentos

À Deus, todo agradecimento pela força e coragem diária. Aos meus país José Grangeiro da Silva Filho e Maria do Socorro Miranda Grangeiro pelo incentivo e trabalho de ambos cheguei até aqui.

A minha Irmã Iara Miranda Grangeiro pelo apoio e ajudas.

À Universidade Federal do Piauí, querida instituição, por dar todo o preparo e ofertar um corpo docente admirável. Aos mestres, muito obrigada.

Ao meu orientador Prof. Dr. Antonio Oseas de Carvalho Filho por se mostrar sempre disponível a me auxiliar no desenvolvimento deste trabalho e ter me oferecido oportunidades de crescimento acadêmico.

Por fim, a todos que, de forma direta ou indireta, contribuíram à realização deste sonho.

A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê.

Arthur Schopenhauer

Resumo

O câncer é um problema de saúde pública mundial. O câncer de mama é o tipo de câncer mais comum entre as mulheres no mundo e no Brasil, é causada pela multiplicação desordenada de células da mama. O presente estudo apresenta um método de classificação de tecidos mamários em maligno e benigno por meio de exames por imagens. Sendo assim, realizado através de rastreamento, onde o meio mais promissor para o diagnóstico precoce é em exames de mamografia. Na primeira etapa, é feita a aquisição das imagens oriundas da base Digital Data base for Screening Mammography (DDSM). Na segunda, o pré-processamento, através de técnicas de processamento de imagem. Na terceira etapa, a extração de características, através dos índices de diversidade. E na quarta etapa, é a classificação em benignos ou malignos, onde foram utilizados os classificadores Random Forest, Sequential Minimal Optimization (SMO), Multilayer Perceptron (MLP) e Instance Based Learning (IBK). Portanto, o processo é finalizado com a validação dos resultados. Neste método foram usados descritores de textura baseado em índices de diversidade para extração de características. Para todos os testes realizados, o melhor resultado alcançou uma acurácia de 77,92%, sensibilidade de 81,51%, especificidade de 74,88%, curva ROC de 0,852 e uma Kappa de 0,5592. O uso dos índices de diversidade para descrever padrões em regiões de imagens de mamografias mostrou-se moderado na categorização de maligno e benigno. O mesmo contribuiu de forma positiva em duas distintas áreas. Sendo que na área da saúde, disponibilizou uma metodologia automática para auxílio no diagnóstico de tecidos mamários em maligno e benigno, que apesar de não fornecer excelentes resultados mostram-se promissor. E na computação, com a adaptação de técnicas de outras áreas do conhecimento que serviram como descritores de textura, podendo então, serem utilizados em caracterizações em outros tipos de imagens.

Palavras-chaves: Câncer de mama, Mamografia, Descritores de Textura, Índice de Diversidade.

Abstract

Cancer is a global public health problem. Breast cancer is the most common type of cancer among women in the world and in Brazil, it is caused by the disordered multiplication of breast cells. This work has as main objective to develop a methodology for classification of breast regions in malignant and benign. The present study presents a method of classification of malignant and benign mammary tissues through imaging tests. Thus, performed through screening, where the most promising means for early diagnosis is in mammography exams. In the first step, the acquisition of images from the base Digital Data Base for Screening Mammography (DDSM) is done. In the second, the pre-processing, through image processing techniques. In the third step, the extraction of characteristics, through the diversity indexes. And in the fourth stage, it is the classification in benign or malignant, where the classifiers Randow Forest, SMO, MLP and IBK were used. Therefore, the process is finished with the validation of the results. In this method, texture descriptors based on diversity indexes were used to extract characteristics. For all tests performed, the best result achieved an accuracy of 77.92 %, sensitivity of 81.51 %, specificity of 74.88 %, ROC curve of 0.852 and Kappa of 0.5592. The use of diversity indexes to describe patterns in regions of mammography images has been shown to be moderate in the categorization of malignant and benign. The same contributed positively in two different areas. Being that in the health area, it provided an automatic methodology to aid in the diagnosis of malignant and benign breast tissues, which, although not providing excellent results, are promising. And in computing, with the adaptation of techniques from other areas of knowledge that served as descriptors of texture, and can then be used in characterizations in other types of images.

Key words: Câncer de mama, Mamografia, Descritores de Textura, Índice de Diversidade.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

A	Acurácia
CAD	*Computer-Aided Detection*
CADx	*Computer-Aided Diagnostic*
DDSM	*Digital Database for Screening Mammography*
DRLBP	Padrão Binário Local Discriminativo Robusto
DRLTP	Padrão Ternário Local Robustivo Discriminativo
E	Especificidade
ELM	*Extreme Learning Machines*
FN	Falso Negativo
FP	Falso Positivo
GA	Algoritmo Genético
GLAM	Matriz de Aura de Nível de Cinza
GLCM	Matriz de Co-ocorrência de Nível de Cinza
INCA	Instituto Nacional do Câncer
KNN	*k-Nearest Neighbor*
LBP	*Local Binary Patterns*
LDA	Análise Discriminante Linear
MIAS	*Mammographic Image Analysis Society*
MLP	*Multi Layer Perceptron*
OMS	Organização Mundial de Saúde
RNA	Rede Neural Artificial
RNN	Rede Neural Recorrente
ROC	*Receiver Operationg Characteristic*
ROI	Região de Interesse

SMO	*Sequential Minimal Optimization*
SVM	Máquina de vetor de suporte
S	Sensibilidade
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
WEKA	*Waikato Environment for Knowledge Analysis*

Lista de símbolos

H'	Índice Shannon
$\sum$	Somatório
S	Total de espécies da espécies i
i	Espécie
n_i	Número de indivíduos
N	Total de indivíduos
ln	Logaritmo
D	Índice de Simpson
D	Indice de Margalef
D	Indice de McIntosh
U	Distância euclidiana
E	Indice de Camargo
i	Espécie
p_i	Proporção de indivíduos da espécies i
p_j	Proporção de indivíduos da espécies j
π	Pi
E	Smith & Wilson B
E	Índice de Smith & Wilson 1-D
E	Índice de Smith & Wilson -lnD
S_{min}	Índice de Shannon minimum

Sumário

1 Introdução

O câncer é um problema de saúde pública mundial, tornando-se entre as principais causas de morbidade e mortalidade do mundo. A evolução das taxas de incidência e de mortalidade marcam um contínuo crescimento desta doença para as próximas décadas. A Organização Mundial de Saúde (OMS), em consequência do aumento da prevalência dos fatores de risco, estima ter um aumento de 70% de novos casos para as próximas duas décadas, principalmente em países de baixa renda (CARVALHO et al., 2019)

Segundo INCA (2019), o câncer é um conjunto de mais de 100 doenças, que têm em comum o crescimento desordenado de suas células que invadem os tecidos e órgãos. O câncer de mama é o tipo de câncer mais comum entre as mulheres no mundo e no Brasil, correspondendo a cerca de 25% dos casos novos a cada ano, sendo importante ressaltar que no Brasil, esse percentual é de 29%. Em 2018, estimativas indicam 59.700 novos casos, sendo excluído o câncer de pele não melanoma. Os casos de mama são mais frequentes nas mulheres das regiões Sul, Sudeste, Centro-Oeste e Nordeste. INCA (2019) ainda cita, que o câncer de mama também é acometido em homens, porém raro, representando apenas 1% do total de casos da doença.

O câncer de mama não apresenta uma causa única. Contudo, diversos fatores podem estar relacionados ao desenvolvimento da mesma, tais como: fatores genéticos/hereditários (histórico de caso de câncer de mama na família), fatores comportamentais/ambientais (alcoolismo, obesidade), fatores endócrinos/história reprodutiva (história de menarca precoce, nuliparidade), além da idade. Os sintomas iniciais da doença podem ser perceptíveis, através de presença de nódulo, pele da mama avermelhada, retraída ou semelhante a casca de laranja, descarga papilar sanguinolenta, pequenos nódulos no pescoço ou nas axilas.

Tais sintomas citados podem ser palpáveis pelo autoexame da mulher, qualquer aparecimento destes deve ser investigado, procurando o serviço especializado, levando em conta que nem sempre estarão relacionados a um evento maligno. É de grande importância a detecção precoce, pois aumenta as chances de cura e tratamento. O tratamento do câncer de mama pode incluir intervenção cirúrgica, radioterapia, quimioterapia e hormonioterapia. Contudo, mesmo após o tratamento, é necessário o acompanhamento médico dos pacientes (MAZALA, 2019).

Na medicina, o uso de imagens são considerados importantes recurso para a elaboração de diagnósticos médicos. No auxílio desses diagnósticos, as técnicas de análise auxiliadas por computador têm sido utilizadas como segundo parâmetro para a elaboração de diagnósticos com uma maior precisão de acerto (BATISTA et al., 2010). Os Sistemas *Computer-Aided Detection* (CAD) são responsáveis por auxiliar os especialistas a detectar alterações em imagens digitais e em seguida serem analisadas pelos mesmos. Já os sistemas *Computer-Aided Diagnostic* (CADx) são responsáveis por diagnosticar (malignos

ou benignos) um problema em uma imagem digital, auxiliando os especialistas com uma segunda opinião (FILHO et al., 2016).

Os processos envolvidos nos sistemas CAD e CADx geralmente são compostos por quatro etapas. Primeira etapa, é aquisição que se refere à digitalização ou obtenção de regiões de interesse. Após a aquisição pode ser feita o pré-processamento da imagem ou região, cujo objetivo é suprimir os ruídos e melhorar o seu contraste. A segunda etapa, é extração de características e/ou segmentação, que consiste na localização das regiões suspeitas de conterem as massas. Normalmente são geradas muitas regiões suspeitas. Daí tem a necessidade da terceira etapa, a extração de características, para reavaliar as regiões segmentadas, reduzindo os falsos positivos. A última etapa, é a de classificação, que utiliza do conjunto de características previamente extraídas para informar se a região possui características de massa em maligno ou benigno.

A metodologia proposta ira atua em três etapas do sistema CADx, na etapa que consiste na obtenção da região de interesse e no pré-processamento da imagem, na etapa de extração das características através de análise de textura e na etapa de classificação do tecido mamário em maligno e benigno.

A análise de textura, é útil em aplicações, por se aproximar da avaliação feita pelo sistema visual humano. Os índices de diversidade são utilizados como descritores na extração de características de textura e são responsáveis por caracterizar as regiões de interesse, onde através do reconhecimento de padrões utiliza de múltiplos classificadores para a avaliar eminência da metodologia proposta.

1.1 Objetivos

1.1.1 Objetivo Geral

Em vista do contexto apresentado, este trabalho tem como principal objetivo desenvolver uma metodologia para classificação de regiões da mama em maligno e benigno.

1.1.2 Objetivos Específicos

1. Desenvolver e adaptar técnicas para caracterizar propriedades de textura;
2. Classificar as regiões de interesse em maligno e benigno;
3. Avaliar a viabilidade do uso de índices de diversidade como forma de extração de características; e
4. Construir uma metodologia capaz de sugerir ao especialista uma segunda opinião no diagnóstico do câncer de mama.

1.2 Organização do Trabalho

Além da Introdução, este trabalho está dividido seguindo a ordem: no Capítulo 2, são apresentados os trabalhos relacionados; no Capítulo 3, apresenta-se o referencial teórico; no Capítulo 4, descreve-se a metodologia proposta; no Capítulo 5, são descritos os resultados e discussões da execução do projeto; e por fim, são apresentadas as conclusões e sugestões para trabalhos futuro no Capítulo 6.

2 Trabalhos Relacionados

Na literatura é possível encontrar trabalhos relacionados ao tema proposto. A seguir, tem-se resumos de alguns destes.

A metodologia proposta por Rocha et al. (2014), para a extração de características utilizou os índices de diversidades para detectar padrões de coocorrência de espécies. Utilizou os índices *Gleason* e *Menhinick* e na classificação em maligno ou benigno utilizou da Máquina de vetor de suporte (SVM). O melhor resultado obtido foi 86,66% de acurácia, 90% de sensibilidade, 83,33% de especificidade e área sob a curva ROC de 0,86.

No trabalho proposto em Paramkusham et al. (2015), apresenta a classificação dos tecidos mamários em normal, maligno e benigno, utilizou a extração de características de textura *Local Binary Pattern* (LBP) e classificou os tecidos mamários em normal e anormal. E utilizou de descritores de características geométricas na classificação em benignos e malignos. O melhor resultado é na classificação em normal e anormal tendo uma acúracia de 99,27%.

Rouhi et al. (2015), apresentou dois métodos automatizados para diagnosticar massas benignas e malignas extraídas de mamografias digitais. Inicialmente, é feita a etapa de segmentação usando região de crescimento automatizado cujo limiar é obtido por uma rede neural artificial treinada (RNA). No outro método a segmentação é realizada por uma *Cellular neural network* (CHUA; YANG, 1988) cujos parâmetros são determinados por um Algoritmo Genético (AG). As características de intensidade, textura e forma são extraídas das regiões segmentadas. O AG é usado para selecionar características mais relevantes. As RNAs são usadas para classificar as regiões como benignas ou malignas. Para avaliar o desempenho dos métodos propostos foram utilizados diferentes classificadores *Random Forest*, *Naive Bayes*, SVM e *K-Nearest Neighhor* (KNN). A metodologia apresentou resultado um percentual de 96,87% de sensibilidade, 95,94% de especificidade e 96,47% de acurácia.

Foram utilizados Valarmathie et al. (2016), um método para classificar massas mamográficas utilizando características de textura, forma e margem geométricas, selecionadas com classificador *Multilayer Perceptron* (MLP). As massas são segmentadas a partir da mamografia utilizando Limiares de Nível de Cinza e são extraídas as características. Os recursos são *fuzzificados* usando valores de associação difusos. O classificador MLP obteve uma precisão de 100% e 0,99 de curva ROC.

O CAD proposto por Kanadam e Chereddy (2016) modela automaticamente, região de interesse (ROI) identificadas à medida que ocorrem usando matriz esparsa, sendo nomeado como sparse-ROI. A massa é detectada através de uma nova tecnologia de sparse-ROI e o classicador multi-SVM é usado para a classificação. O desempenho do classificador é avaliado na base de imagens MIAS. Para este propósito, dois algoritmos são utilizados

com base nas matrizes estatísticas, Matriz de Co-ocorrência de Nível de Cinza (GLCM) e Matriz de Aura de Cível de Cinza (GLAM). A eficácia do classificador dos dois novos algoritmos desenvolvidos é avaliada em termos de precisão, sensibilidade, tamanho e tempo computacional. Os resultados do estudo tem redução no tempo computacional em 99,93% no GLCM e 75,73% no GLAM com a retenção concomitante da acurácia da classificação de 97,2%

Foi proposto por Rabidas et al. (2016), a introdução de um novo Padrão Binário Local Discriminativo Robusto (DRLBP) e Padrão Ternário Local Robustivo Discriminativo (DRLTP) (GUO; ZHAO; PIETIKÄINEN, 2012) para a classificação de massas benignas ou malignas. O método de Análise Discriminante Linear de Fisher é incorporado com características discriminantes, selecionadas pelo método de regressão logística para a classificação da massas em benignas e malignas. Para avaliar as características DRLBP e DRLTP é usada uma técnica de validação cruzada de dez vezes com 58 massas do banco de dados mini-MIAS, e o melhor resultado é observado com DRLBP tendo uma área sob a curva ROC de 0,982.

No trabalho de Lima et al. (2016), os autores apresentam um método para detectar e classificar lesões mamográficas. A proposta consiste em decompor cada imagem usando Wavelets multi-resolução. Os momentos de Zernike são extraídos de cada componente Wavelets. Com o uso dessa abordagem, puderam combinar características de textura e forma, foram aplicadas tanto na detecção quanto na classificação de lesões mamárias. A classificação utilizou das redes SVM com kernels modificados para otimizar as taxas de precisão. O método apresentou uma precisão de 94,11%.

Em Li et al. (2016), é proposto um método de classificação de massa em mamografias, com base em máscaras concêntricas e Textura discriminante. Utilizando de máscaras concêntricas, onde dividiram cada região de massa na região central e na região periférica. Em seguida para a Análise Discriminante Linear (LDA) com Textura tradicional, a textura discriminante é proposta. A falta de considerar as informações de classe na textura tradicional é melhorada. Finalmente, os recursos são extraídos com textura discriminante para a região central e a região periférica. Assim, o problema de desconsiderar as informações de layout espacial é aliviado. O método proposto é testado em 130 regiões de massa do banco de dados Digital Data base for Screening Mammography (DDSM). O método apresentou acurácia de 86,92% e a curva ROC de 0,91.

O trabalho de Carvalho et al. (2018), apresenta uma metodologia para classificação de tecidos mamários em malignos e benignos. Utilizando na extração de características os índices filogenéticos. A classificação foi conduzida usando vários classificadores. Os resultados mostram que o método atinge 99,73% de precisão, 99,41% de sensibilidade, 99,84% de especificidade e uma curva de características de operação do receptor (ROC) com um valor de um ao usar imagens do Banco DDSM. Uma precisão de 100% é obtida ao usar o banco de dados de imagens do Mammography Imaging Analysis Society (MIAS).

Segundo o método proposto em Wang et al. (2018), utilizou Rede neural convolucional (CNN) genérica para extração de características e assumindo uma configuração de múltiplas visualizações, uma rede baseada em atenção é utilizada para selecionar automaticamente os recursos informativos da massa mamária. O mecanismo de atenção tenta fazer com que a CNN se concentre nas regiões relacionadas à semântica para um resultado de classificação mais interpretável. Então, os recursos de massa dos dados de múltiplas visualizações são efetivamente agregados por uma Rede Neural Recorrente (RNN) apresentando uma precisão de 0,85 e a curva ROC de 0,89.

A Tabela 1 resume as abordagens dos trabalhos relacionados expostos. Nota-se que houve utilização de vários métodos para extração dos atributos. Alguns trabalhos mostraram bons resultados. No entanto, ainda é necessário identificar técnicas que permitam melhorar e consolidar estes resultados. Podemos verificar que a classificação de massa mamária quanto a sua malignidade e benignidade é ainda um problema em aberto, e que medidas de textura se mostram promissoras para essa caracterização. Este trabalho tem como objetivo propor descritores de textura, levando em consideração o comportamento de espécies dentro de uma comunidade, grau de parentesco e riqueza de espécies.

Tabela 1: Resumos dos trabalhos relacionados.

Trabalhos	**Bases**	**Descritores / Classificados**
(ROCHA et al., 2014)	DDSM	Índices de diversidades MVS
(PARAMKUSHAM et al., 2015)	IRMA	Local Binary Pattern Características geométricas SMV
(ROUHI et al., 2015)	MIAS DDSM	Random Forest Naive Bayes SVM KNN
(VALARMATHIE et al., 2016)	MIAS	Limiares de nível de cinza MLP
(KANADAM; CHEREDDY, 2016)	MIAS	sparse-ROI multi-SVM
(RABIDAS et al., 2016)	MIAS	DRLBP DRLTP
(LIMA et al., 2016)	IRMA	SVM ELM
(LI et al., 2016)	DDSM	Textura Discriminante na região central e região periférica KNN
(CARVALHO et al., 2018)	DDSM MIAS	Índices Filogenéticos Random Forest MLP Neural Network SMO
(WANG et al., 2018)	BCDR	CNN RNN

3 Referencial Teórico

Neste capítulo serão apresentados conceitos pertinentes ao referencial teórico necessário ao entendimento e embasamento científico do método proposto.

3.1 Câncer de Mama

O câncer pode ser definido como uma doença degenerativa que se desenvolve no próprio organismo, resultante de um acúmulo de lesões no material genético das células, que induz ao processo de crescimento, reprodução e dispersão anormal destas com controle alterado sobre a proliferação e morte celular. O câncer é uma doença de proporções graves, colocando em risco a vida do indivíduo e podendo afetar qualquer parte de seu organismo, sem predisposição de idade e de maneira quase igualmente proporcional em ambos os sexos (AMORIM; SIQUEIRA, 2017).

O câncer de mama é mais comum após os 40 anos, a partir dessa idade deve-se realizar anualmente exames de mamografia. Embora, menos de 20% dos casos de câncer de mama são relacionados a fatores genéticos, recomenda-se que as mulheres que tenham casos na família próximos, comecem a fazer a mamografia 10 anos antes da idade que seus familiares diagnosticaram a doença (BRUNO et al., 2018).

O câncer de mama é uma doença causada pela multiplicação desordenada de células da mama. Esse processo gera células anormais que se multiplicam, formando um tumor. Por isso, a doença pode evoluir de diferentes formas. Alguns tipos têm desenvolvimento rápido, enquanto outros crescem mais lentamente. Esses comportamentos distintos se devem a características próprias de cada tumor, que pode ser maligno ou benigno (INCA, 2019).

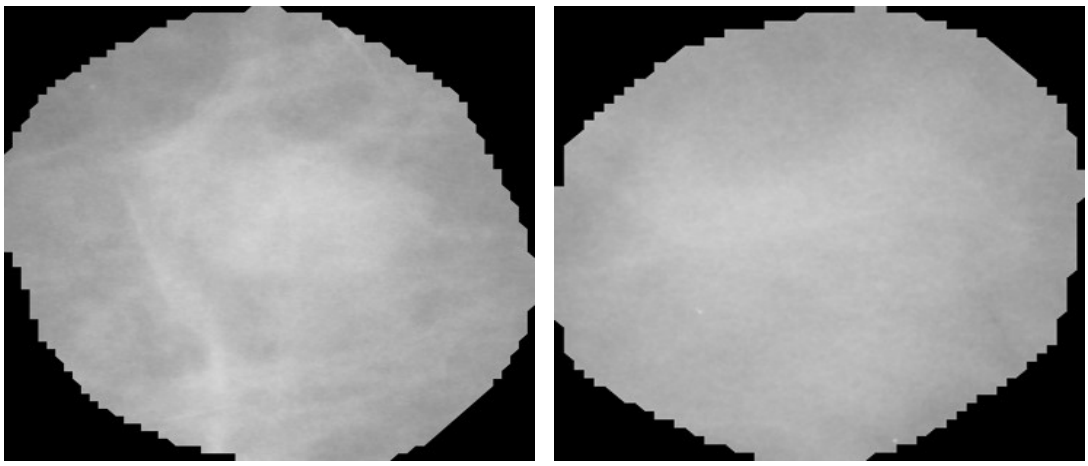
A classificação entre tumores benignos e malignos é feita com base em três pontos principais, que são:

- A aparência;
- A estrutura, e
- O comportamento reprodutivo das células atingidas.

Nos tumores benignos as células sofrem uma pequena mutação, porém não é tão grave a ponto de criar uma expansão acelerada. Já os tumores malignos, as células sofrem grande mutação, essas células se multiplicam rapidamente e causam o surgimento de pequenos vasos que nutrem o tumor e corre o risco de que ele invada outros órgãos, formando novos tumores de crescimento desordenado. Segundo ROCHA et al. (2014) os passos

fundamentais de processamento digital de imagens podem ser divididos em cinco etapas. A Figura 1 mostras a imagem de tumor benigno e maligno.

Figura 1: Exemplos de imagens para câncer (a) benigno e (b) maligno.



(a) (b)

O Ministério da Saúde oferece tratamento para câncer de mama, por meio do Sistema Único de Saúde (SUS). Um nódulo ou outro sintoma suspeito nas mamas deve ser investigado para confirmar se é ou não câncer de mama. Para a investigação, além do exame clínico das mamas, exames de imagem podem ser recomendados, como mamografia, ultrassonografia ou ressonância magnética. A confirmação diagnóstica só é feita, porém, por meio da biópsia, técnica que consiste na retirada de um fragmento do nódulo ou da lesão suspeita por meio de punções (extração por agulha) ou de uma pequena cirurgia. O material retirado é analisado pelo patologista para a definição do diagnóstico (SAÚDE, 2019).

3.2 Sistemas Computacionais de Auxílio à Detecção e Diagnóstico

Diagnóstico auxiliado por computador (CADx) pode ser definido como um diagnóstico utilizando os resultados de análises quantitativas automatizadas de imagens digitais, auxiliando os profissionais na tomada de decisões para o diagnóstico. É importante ressaltar que o computador é utilizado somente como uma ferramenta para obtenção de informação adicional, sendo o diagnóstico final sempre feito pelo profissional da área (AZEVEDO-MARQUES, 2001).

A detecção de anomalias em imagens médicas, em geral é um procedimento demorado, propício a erros e a algum grau de subjetividade devido a várias razões. Sendo a existência de estruturas complexas e de grande número de imagens normais; a grande variação na aparência dos tecidos (mesmo os normais); a sutileza das anormalidades; a superposição

dos tecidos; a necessidade de grande sensibilidade e, ao mesmo tempo, de minimizar o retorno desnecessário dos pacientes.

O CADx alerta o profissional a examinar detalhadamente, padrões suspeitos detectados pelo sistema. Isto melhora a acurácia do diagnóstico e a consistência da interpretação da imagem, servindo como uma "segunda opinião" quando da tomada de decisões diagnósticas (PIRES et al., 2006).

Os Sistemas CAD são responsáveis por auxiliar os especialistas a detectar alterações em imagens digitais e em seguida possam ser analisadas pelos mesmos, já os sistemas CADx são responsáveis por diagnosticar (malignos ou benignos) um problema em uma imagem digital, auxiliando os especialistas, são muito importantes para casos em que a detecção é muito difícil ao olho humano (FILHO et al., 2016).

3.3 Processamento Digital de Imagens

Processamento de imagens, de acordo com Filho e Neto (1999), é um conjunto de métodos e técnicas capazes de transformar imagens de forma adequada para uma melhor compreensão da visão humana ou análise computacional. Podemos classificar em três níveis:

- Baixo nível:
  - Operações primitivas (redução de ruído, aumento de contraste, etc).
  - Entrada: Imagem - Saída: Imagem;
- Nível intermediário:
  - Segmentação, descrição e classificação de objetos.
  - Imagem - Atributos (bordas, contornos, nível de cinza).
- Alto nível:
  - Atribuir "sentido" há um conjunto de objetos reconhecidos.

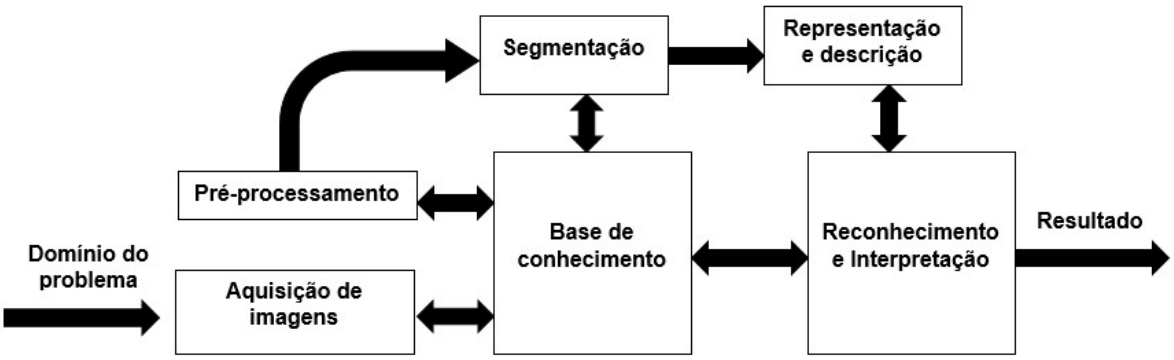
Uma imagem pode ser definida como uma função bidimensional, onde x e y são coordenadas espaciais e a amplitude de f no par de coordenadas f (x;y) é denominado intensidade ou nível de cinza. Quando x e y e a intensidade dos valores de f são finitos e discretos, tem-se então uma imagem digital. Esses elementos possuem uma localização e um valor particular, chamado de pixel (FERREIRA et al., 2017).

Para aplicações práticas, a imagem é uma função contínua, representada por medidas obtidas em intervalos regularmente espaçados. Os valores assumidos em cada ponto medido são quantificados em um número pertencente a uma escala de diferentes cores. Em imagens médicas, geralmente essas cores são relacionadas a níveis de cinza, sendo

atribuído o valor zero à cor mais escura (preto) e o valor máximo M à cor mais clara da escala (branco) (FERREIRA et al., 2017).

Segundo ROCHA et al. (2014) os passos fundamentais de processamento digital de imagens podem ser divididos em cinco etapas conforme é mostrado na Figura 2.

Figura 2: Etapas Fundamentas do Processamento Digital de Imagens



A primeira etapa, é à Aquisição de Imagens essa etapa responsável pela obtenção da representação digital da imagem;

A segunda etapa, é o Pré-processamento que consiste em tornar certas estruturas da imagem mais simples de serem definidas. Para tanto, utiliza-se de técnicas, tais como diminuição de ruídos, realce de contraste, filtros morfológicos e etc;

A terceira etapa, é a Segmentação consiste em qualquer operação que possa distinguir os objetos contidos na imagem ou de alguma forma isolando-os entre si;

A quarta etapa, é a Representação e Descrição é também chamada de extração de características. Tem Como finalidade determinar a características básicas de cada objeto que resultem em informações importantes para discriminação entre classes distintas.

E finalmente a última etapa, que é o Reconhecimento e Interpretação tem por finalidade atribuir um rótulo a um objeto com base nos seus descritores através de uma base de conhecimento que foi construída na etapa anterior.

3.4 Pré-processamento

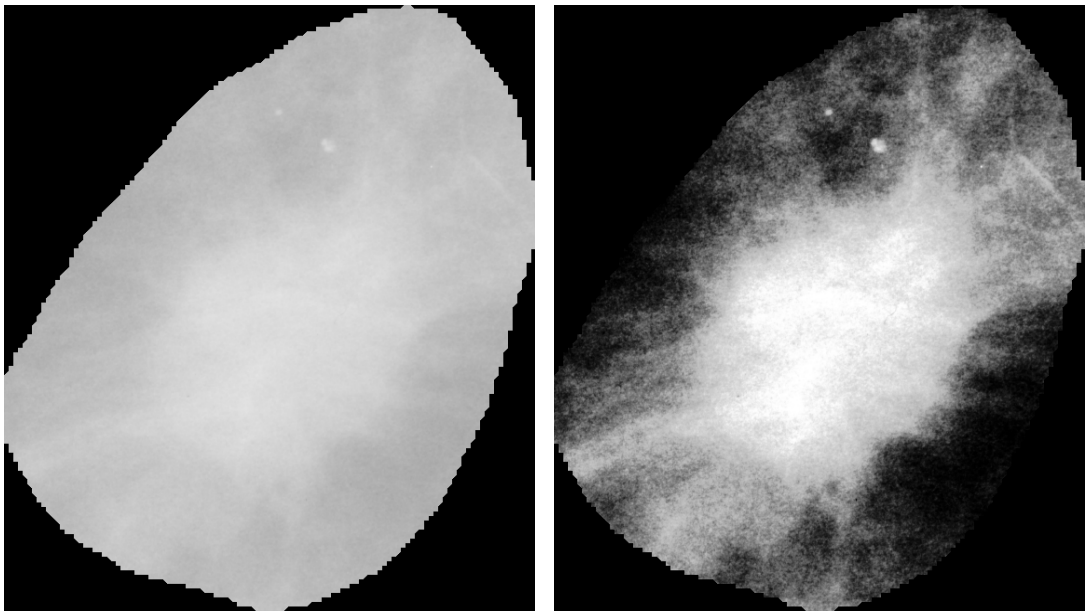
O pré-processamento consiste na utilização de mecanismo de processamento de imagens para o realce das imagens, aumento de contraste, retirada de possíveis ruídos entre outras para que aja uma nitidez na características da imagem. Na metodologia proposta utilizou-se de técnicas visando o melhoramento do contrate das imagens.

A Equalização de Histograma tem como finalidade obter um histograma uniforme através do espalhamento da distribuição dos níveis de cinza ao longo de toda a escala de

contraste, aumentando dessa maneira, a detectabilidade de aspectos da imagem (relacionados a contraste) (NASCIMENTO et al., 2012).

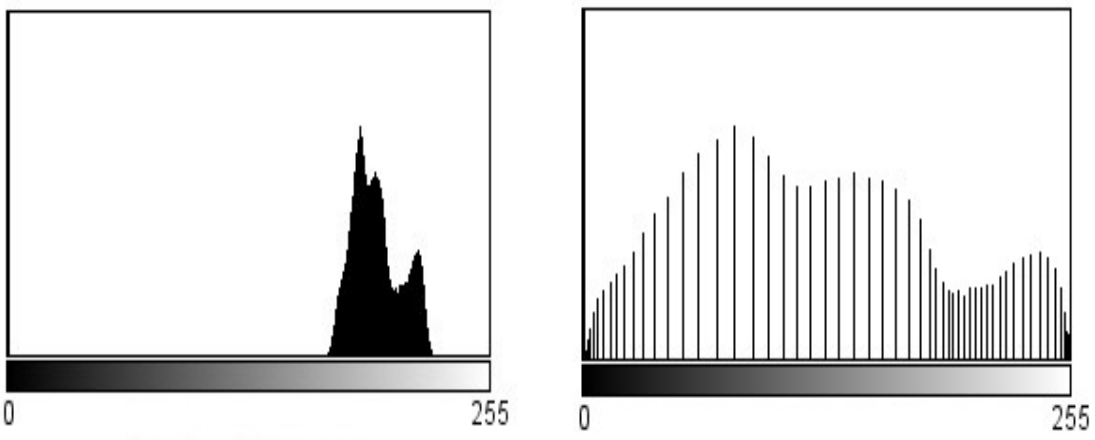
Como dedução dessa distribuição uniforme obtém-se uma imagem com melhor contraste, como demostra a Figura 3 e a Figura 4 mostras os histogramas.

Figura 3: (a) Imagem original e (b) Imagem Equalizada.



(a) (b)

Figura 4: (a) Histograma da Imagem Original e (b) Histograma da Imagem Equalizado.



(a) (b)

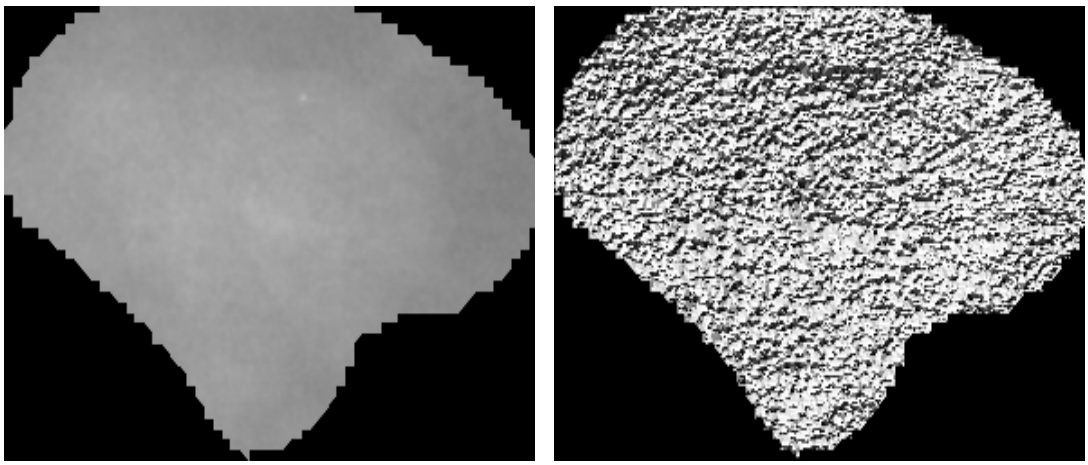
3.4.1 *Local Binary Patterns*

Segundo Pires e Neto (2015), a técnica *Local Binary Pattern* (LBP) é robusta em termos de variações nos níveis de cinza e que discrimina efetivamente uma larga gama de texturas rotacionadas. O LBP é um operador de textura invariante aos níveis de cinza e rotação, baseado em padrões locais binários. O LBP é calculado como mostrar a Equação 3.1,

$$LBP_{k,r} = \sum_{i=0}^{k-1} |S(N_i - N_c)| \, 2^i \tag{3.1}$$

onde, N_c é o pixel central e N_i são os pixels vizinhos, k é o numero de pontos e r o raio do circulo.

Figura 5: (a) Imagem Original e (b) Imagem resultado após aplicação do LBP.



(a) (b)

O LBP possui um padrão de textura vizinho ao redor de cada pixel. Onde o calcular é feito utilizando de uma janelar 3x3. O valor do pixel central é subtraído de cada um dos valores dos pixels vizinhos, o resultado da subtração é utilizado na função $S(x)$ que substituirá o valor anterior do pixel vizinho por 0 ou 1, dependendo do retorno da função $S(x)$. Cada valor da matriz binária é multiplicado pela sua respectiva posição na matriz de pesos. O LBP e o resultado da soma de todos os valores resultantes das multiplicações. A Figura 5 mostra uma imagem original e LBP.

3.5 Índices de Diversidade

O termo diversidade, em ecologia, pode ser definido como a variedade e a variabilidade entre os organismos vivos e os complexos ecológicos em que ocorrem. Uma medida de diversidade é um parâmetro extremamente reducionista que objetiva expressar toda a complexidade estrutural de uma comunidade ecológica através de um único número

(NASCIMENTO et al., 2012). A Tabela 2 resume como a metodologia utiliza os índices de diversidade.

Tabela 2: Índice de Diversidade.

Metodologia Proposta	**Ecologia**
Região de Interesse da Imagem	Comunidade
Níveis de Cinza da Imagem	Espécies
Píxels da Imagem	Indivíduos

3.5.1 Índice de Shannon

Originado da teoria da informação (SHANNON; WEAVER, 1949), o Índice de Shannon assume que os indivíduos são randomicamente amostrados a partir de uma comunidade infinitamente grande, e que todas as espécies estão representadas na amostra. O Índice de Shannon é calculado conforme a Equação 3.2,

$$H' = -\sum_{i=1}^{S} p_i \, ln \, p_i \tag{3.2}$$

onde, S é o total de espécies, p_i é a proporção de indivíduos pertencentes a espécie i na amostra, calculado com a Equação 3.3,

$$p_i = \frac{n_i}{N} \tag{3.3}$$

e n_i é o número de indivíduos na espécie i na amostra e N é o número total de indivíduos na comunidade.

3.5.2 Índice de Simpson

Proposto em 1949, por E. H. Simpson (SIMPSON, 1949), O Índice de Simpson (D) é a medida de probabilidade de dois indivíduos selecionados aleatoriamente de uma comunidade infinitamente grande, pertencerem à mesma espécie. O Índice é calculado conforme Equação 3.4,

$$D = 1 - \sum_{i=1}^{S} p_i^2 \tag{3.4}$$

onde, p_i é a proporção de indivíduos pertencentes a espécie i na comunidade, calculado com a equação 3.3, n_i é o número de indivíduos na espécie i e N é o número total de indivíduos na comunidade. S é o total de espécies predominante na comunidade.

3.5.3 Índice de Margalef

Índice de Margalef (D) (MARGALEF, 1969) é usado na ecologia para estimar a biodiversidade de uma comunidade com base na distribuição numérica dos indivíduos das diferentes espécies em função do número total de indivíduos existentes na amostra analisada. O Índice é calculado conforme Equação 3.5,

$$D = \frac{(S-1)}{ln\,(N)} \tag{3.5}$$

onde, S é o número de espécies e $ln(N)$ é o logaritmo do número de indivíduos pertencente a comunidade.

3.5.4 Índice de McIntosh

O Índice de McIntosh foi proposto por (MCINTOSH, 1967). A comunidade pode ser vista como um ponto em um hiper-volume S-dimensional e a distância euclidiana da comunidade para a origem pode ser utilizada como uma medida de diversidade. E uma medida de probabilidade de dois indivíduos selecionados aleatoriamente de uma comunidade infinitamente grande pertencerem a mesma espécie (ARAÚJO et al., 2017). O índice de diversidade de McIntosh (D) é estimado através das seguintes Equações 3.6,

$$D = \frac{N-U}{N-\sqrt{N}} \tag{3.6}$$

onde, N é o número total de indivíduos na amostra e U é dado pela expressão 3.7,

$$U = \sqrt{\sum n_i^2} \tag{3.7}$$

onde, n_i é o número de indivíduos nas espécies i e o somatório é realizada sobre todas as espécies. U é a distância euclidiana da comunidade desde a origem quando plotada em um hiper volume S-dimensional.

3.5.5 Índice de Camargo

O índice de Camargo (CAMARGO, 1993) é a relação entre pares de espécies i e j, definido pela Equação 3.8,

$$E = 1 - \left[\sum_{i=1}^{s}\sum_{j=i+1}^{s}\frac{Pi - Pj}{S}\right] \tag{3.8}$$

onde, p_i é a proporção de espécies i na amostra, p_j é a proporção de espécies j na amostra e S é o número total de espécies.

3.5.6 Índice de Smith & Wilson B

O índice de B de Smith e Wilson (SMITH; WILSON, 1996). Baseia-se na variação de abundância. Onde variância é calculada usando abundâncias de logaritmo, o que significa que o índice examina as diferenças proporcionais entre as espécies. Com a obtenção da variância, é multiplicada pelo fator -2 / pi arctan para dar um valor entre o intervalo de 0 a 1. O índice é definido pela Equação 3.9,

$$E = 1 - \left[\frac{2}{\pi arctan \left\{ \frac{\sum_{i=1}^{S} \left(ln\, n_i - \sum_{j=1}^{S} \frac{ln\, n_i}{S} \right)^2}{S} \right\}} \right] \tag{3.9}$$

sendo, n_i é o número de indivíduos nas espécies i, n_j é o número de indivíduos nas espécies j; e S é o número total de espécies.

3.5.7 Índice de Smith & Wilson 1-D

Índice de uniformidade de Smith e Wilson 1-D (SMITH; WILSON, 1996), utilizar de dominância D (Simpson), coverter o índice de Simpson (D) de várias maneiras. A mais comum é usar o complemento de D como índice de diversidade e para dados contínuos ou com números grandes de cortes discretos. Definido pela Equação 3.10,

$$E = \frac{1 - D}{1 - \frac{1}{S}} \tag{3.10}$$

onde, D é o índice de diversidade de Simpson representado pela equação 3.4 e S é o número total de espécies.

3.5.8 Índice de Smith & Wilson -lnD

O índice de uniformidade de Smith e Wilson 1 / D (SMITH; WILSON, 1996), converter o índice de dominância de Simpson em uma diversidade de índice. A medida equivalente de uniformidade é a anterior. O índice é definido na Equação 3.11,

$$E = \frac{-ln\, D}{ln\, S} \tag{3.11}$$

onde, D é o índice de diversidade de Simpson representado pela equação 3.4 e S é o número total de espécies.

3.5.9 Índice de Shannon maximum

Este é simplesmente o valor máximo que o índice de Shannon 3.5.1 poderia produzir para o dado conjunto de dados e é dado por $ln(S)$, onde S é o número total de espécies.

3.5.10 Índice de Shannon minimum

Este é simplesmente o valor mínimo que o índice de Shannon 3.5.1 poderia produzir para o dado conjunto de dados e é dado pela Equação 3.12,

$$S_{min} = ln\,(N) - \left\lfloor \frac{(N - S + 1)\,ln\,(N - S + 1)}{N} \right\rfloor \tag{3.12}$$

sendo, N é o número total de indivíduos na amostra e S é o número total de espécies.

3.6 Reconhecimento de padrões

O ato de reconhecer é utilizado diariamente por todos os seres vivos, através dos sentidos (visão, olfato, paladar, audição e tato). Esse ato de reconhecer utiliza-se de padrões que ao longo da vida são aprendidos. Para reconhecer algum objeto, terá que ter um breve conhecimento sobre o mesmo, no caso o padrão. Se esse padrão não for de conhecimento, então o objeto não será reconhecido. Na computação utiliza-se desse reconhecimento em diversas áreas.

Segundo Filho et al. (2016) o reconhecimento de padrão envolve duas tarefas: classificação e reconhecimento. A classificação dividi um conjunto de dados obedecendo métricas, formando classes ou grupos. Reconhecimento trata-se de novo dado reconhecido e atribuído em uma classe gerada.

Para a análise de padrões, na metodologia proposta, foram escolhidas quatro classificadores levando em consideração sua popularidade, eficácia, facilidade de adaptação nos experimentos e sua integração com a ferramenta de extração de características em imagens, sendo eles: Random Forest (BREIMAN, 2001), Sequential Minimal Optimization (SMO) (PLATT, 1998), Multi Layer Perceptron (MLP) (PAL; MITRA, 1992) e K-Nearest Neighbor (KNN) (KELLER; GRAY; GIVENS, 1985) que no weka é chamado de IBK. Os parâmetros utilizados para cada classificador, foram os padrões da ferramenta Weka.

3.7 Métricas de Validação

As métricas de valiação baseiam em estatísticas como, Acurácia (A), Sensibilidade (S), Especificidade (E), Curva ROC e Kappa. A matriz de confusão oferece uma hipótese das medidas efetivas do modelo de classificação, mostrando o número de classificações corretas versus as classificações preditas para cada classe, como mostra a Tabela 3.

Acurácia (A) (Equação 3.13), é a proporção de acertos, ou seja, o total de verdadeiramente positivos e verdadeiramente negativos, em relação a amostra estudada.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.13}$$

Tabela 3: Matriz de Confusão.

Resultado do Teste	Positivo	Negativo
Negativo	Verdadeiros Positivos - VP	Falsos Negativos - FN
Positivo	Falsos Positivos - FP	Verdadeiros Negativos - VN

A sensibilidade (S) (Equação 3.14), é a capacidade de um teste diagnóstico identificar os verdadeiros positivos nos indivíduos verdadeiramente doentes.

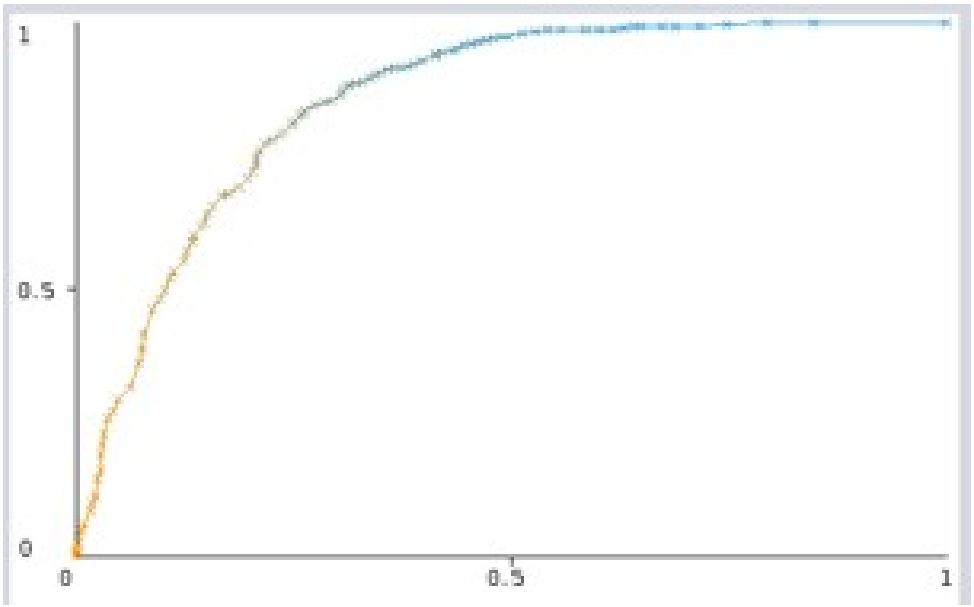
$$S = \frac{VP}{VP + FN} \tag{3.14}$$

Especificidade (E) (Equação 3.15), é a capacidade de um teste diagnóstico identificar os verdadeiros negativos nos indivíduos verdadeiramente sadios. Quando um teste é específico raramente cometerá o erro de dizer que pessoas sadias são doentes.

$$E = \frac{VN}{VN + FP} \tag{3.15}$$

A curva *Receiver Operating Characteristic* (ROC) é uma métrica de avaliação que compara o desempenho de duas ou mais modalidades de imagens. A área sobre a curva ROC representa a probabilidade de que, dado um caso positivo e um negativo, a regra do classificador vai ser mais elevada para o caso positivo. Quanto maior for a curva ROC (Figura 6), maior é a probabilidade do sistema fazer uma decisão correta. A curva ROC representa a dependência entre a sensibilidade e a especificidade de um classificador. Cada ponto é representado por um par de valor, sensibilidade e especificidade, e a linha diagonal um classificador que não consegue discriminar, devido o número de VP ser igual ao percentual de FP.

Figura 6: Área sobre a curva ROC.



A estatística Kappa é freqüentemente usada para testar a confiabilidade entre avaliadores. A importância da confiabilidade do avaliador reside no fato de representar a medida em que os dados coletados no estudo são representações corretas das variáveis medidas. A medição da extensão em que os coletores de dados (avaliadores) atribuem a mesma pontuação à mesma variável é chamada de confiabilidade entre avaliadores. Embora existam uma variedade de métodos para medir a confiabilidade entre avaliadores, tradicionalmente foi medida como concordância percentual, calculada como o número de pontuações de concordância dividido pelo número total de pontuações (MCHUGH, 2012).

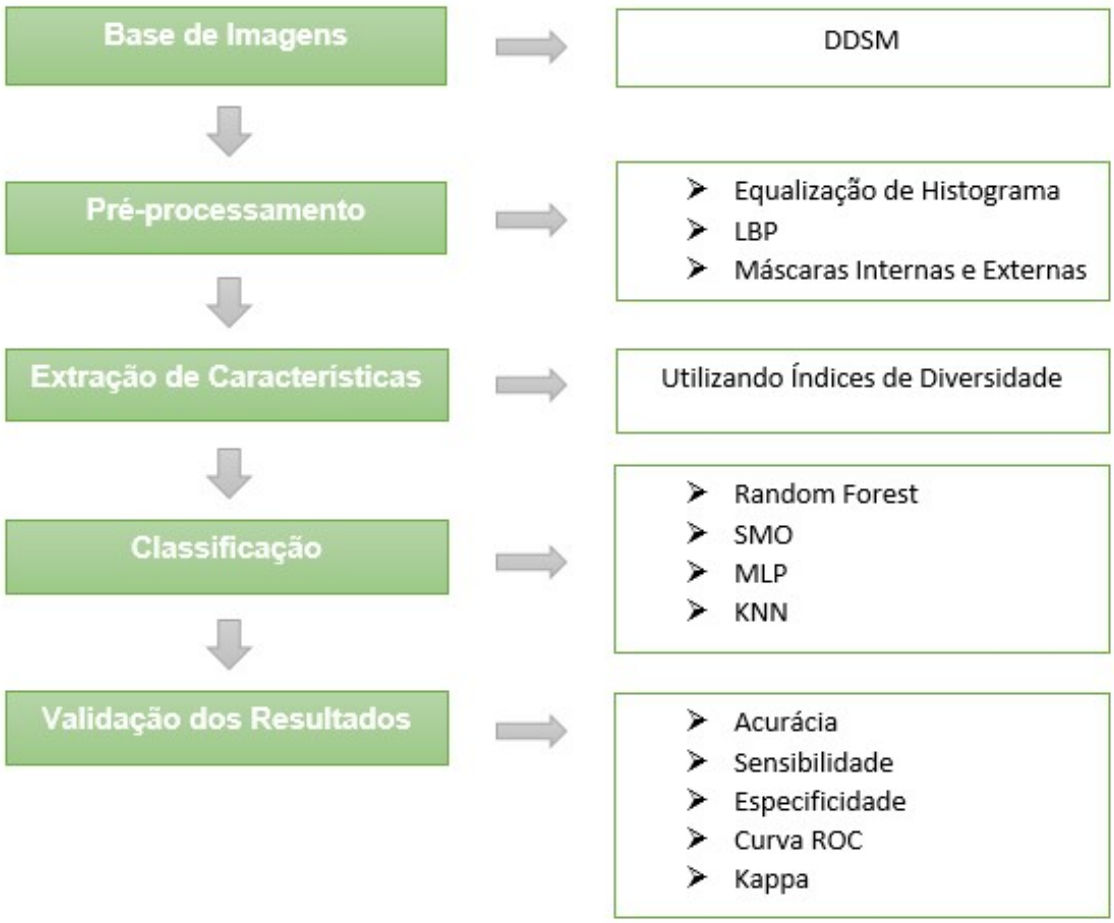
Tabela 4: Interpretação para os valores do índice Kappa.

VALOR DO KAPPA	**NÍVEL DE CONCORDÂNCIA**
< 0	Não exite concordância
Entre 0 e 0,2	Concordância Mínima
Entre 0,21 e 0,4	Concordância Razoável
Entre 0,41 e 0,6	Concordância Moderada
Entre 0,61 e 0,8	Concordância Subtancial
Entre 0,81 e 1	Concordância Perfeita

4 Metodologia

A metodologia proposta neste trabalho segue as etapas da Figura 7. Na primeira etapa, é feita a aquisição das imagens oriundas da base *Digital Data base for Screening Mammography* (DDSM) (HEATH et al., 2000). Na segunda etapa, é o pré-processamento, através de técnicas de processamento de imagem. Na terceira etapa, é a extração de características, através dos índices de diversidade. Na quarta etapa, é a classificação em benignos ou malignos, foram utilizados os classificadores Random Forest, SMO, MLP e KNN. O processo é finalizado com a validação dos resultados.

Figura 7: Metodologia propostar.



4.1 Aquisição das Imagens

A base de imagens DDSM é uma base pública contendo imagens de mamografias, que tem como objetivo facilitar a pesquisa e desenvolvimento de algoritmos para ajudar no diagnóstico de anomalias da mama (HEATH et al., 2000).

Este trabalho utiliza *Region of Interest* (ROI) adquiridas da base DDSM, onde cada ROI possui tamanhos diferentes e apenas uma região de massa. Para a concretização desse trabalho, foram utilizadas 1155 ROIs de imagens de mamografias; sendo 625 ROIs com a presença de massa maligna e 530 ROIs com presença de massa benigna.

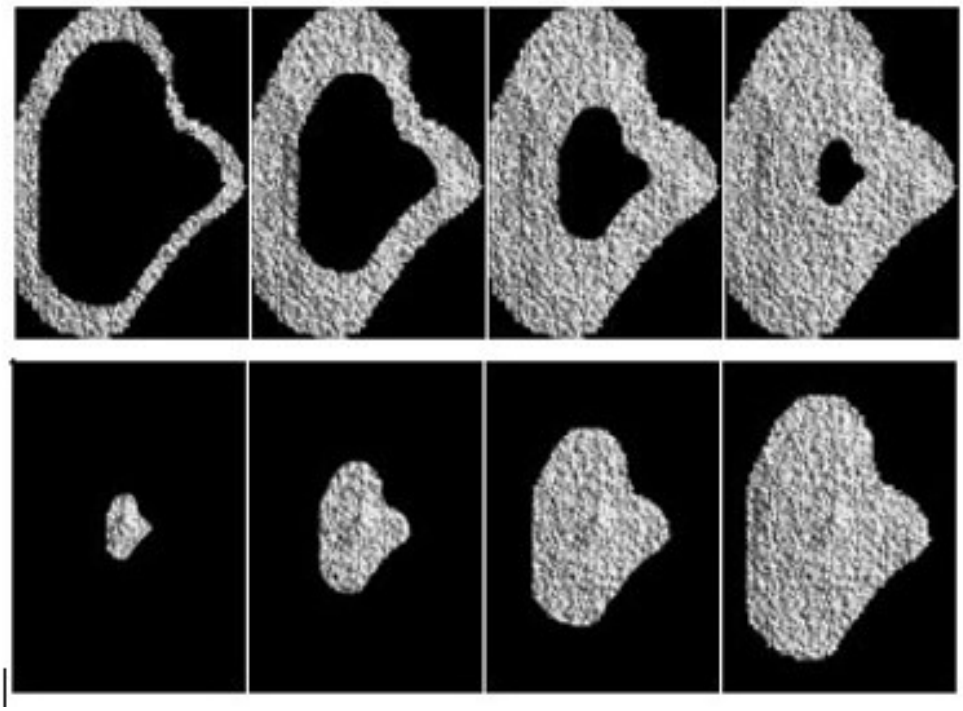
4.2 Pré-processamento

Na metodologia utiliza-se técnica de Equalização de Histograma, para um espalhamento da distribuição dos níveis de cinza ao longo de toda a escala de contraste, com a finalidade de aumentar, a detectabilidade de aspectos da imagem.

A técnica de LBP foi utilizada para gerar uma nova representação da imagem, com intuito de evidenciar propriedades de textura, que apenas com a imagem normal não fossem tão evidentes.

A metodologia utiliza de máscaras internas e externas (Figura 8). A proposta desta abordagem é baseada em descobrir padrões de diversidade nas áreas perto das bordas e internamente. Neste trabalho, os valores de 20%, 40%, 60% e 80% foram definidos para aplicação escalonada. Os tamanhos das máscaras são empiricamente definidos para permitir uma análise viável de regiões em partes isoladas da ROI. As máscaras internas

Figura 8: Máscaras Intenas e Externas.



foram geradas a partir da diminuição da escala em relação a original pelo centro de massa e as sucessoras máscaras foram adquiridas a partir das suas anteriores na sequencia até a mais interna. A máscara externa é similar à abordagem com máscara interna. As máscaras são formandas pela diferença entre as máscaras internas, onde a primeira foi criada pela

junção entre a primeira e a segunda interna, a segunda e a terceira interna. As demais máscaras externas foram criadas seguindo a sequência das diferenças até o último par de máscaras internas.

4.3 Extração de Características

Após o pré-processamento das imagens, inicia-se à fase de extração de características baseada em textura. Para descrever a textura dos nódulos, são utilizados os índices de diversidade. Esses índices descrevem as texturas das imagens e os atributos fornecidos buscam reconhecer as medidas de homogeneidade ou heterogeneidade da ROI. Servindo com base para classifica em maligno ou benigno.

4.4 Classificação e validação dos Resultados

Para a classificação, foi utilizado a ferramenta suíte de algoritmos de mineração de dados e Aprendizado de Máquina WEKA, que contém algoritmos para pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização.

Foram selecionados os algoritmos Random Forest, SMO, MLP e KNN, utilizando os parâmetros com os valores padrões em conjunto com a validação cruzada de k-folds, sendo k = 10; esse método que tem como finalidade dividir as características em 10 grupos, de forma a realizar o treino em 9 grupos e utilizando um grupo para testes, são realizados 10 cruzamentos sempre trocando o grupo de teste, ao final é gerado uma média, onde se configura o resultado.

Por fim, após a etapa de reconhecimento de padrão, a fim de se considerar a presença ou ausência de massas malignas e benignas em imagens de mamografia, é necessário a utilização de métricas para validar os resultados e analisar possíveis melhorias do mesmo. A metodologia proposta, utiliza métricas de avaliação baseadas em estatísticas, como, Acurária (A), Sensibilidade (S), Especificidade (E), Coeficiente Kappa e Área sobre a Curva (ROC).

5 Resultados e Discussões

Para os testes realizados neste trabalho, utilizou-se a base DDSM. A extração de características de textura foi realizada a partir dos índices de diversidade especificados na Seção 3.5 e a classificação das massas em maligno e benigno, utilizou os classificadores e parâmetros definidos na Seção 3.6, o método de teste foi a validação cruzada com $k = 10$.

A Tabela 5 apresenta os resultados para as imagens originais, com e sem Equalização de Histograma. Cada massa é representada por 10 características, isto é, para os 10 (dez) índices propostos nesse método.

Tabela 5: Resultados dos testes utilizando imagens normais

Imagens Originais					
Sem Equalização de Histograma					
Classificador	**A(%)**	**S(%)**	**E(%)**	**ROC**	**Kappa**
Random Forest	64,24	59,81	68,00	0,700	0,278
SMO	**65,45**	**50,75**	**77,92**	**0,643**	**0,292**
MLP	64,33	45,28	80,48	0,710	0,264
KNN	61,21	60,38	61,92	0,607	0,222
Com Equalização de Histograma					
Classificador	**A(%)**	**S(%)**	**E(%)**	**ROC**	**Kappa**
Random Forest	64,50	60,19	68,16	0,712	0,284
SMO	65,37	50,94	77,60	0,643	0,290
MLP	64,59	43,58	82,40	0,714	0,267
KNN	60,26	56,98	63,04	0,589	0,200

De acordo com a Tabela 5, o classificador SMO obteve o melhor resultado nas imagens sem realce, com uma taxa de acurácia de 65,45%, uma sensibilidade de 50,75%, especificidade de 77,92%, uma curva ROC de 0,643 e uma Kappa de 0,292, indicando que o classificador apresenta uma razoável capacidade de diagnóstico, de acordo a Tabela 4. O resultado menos significativo foi do classificador KNN nas imagens comum com realce, com taxa de acurácia de 60,26%, curva ROC de 0,589 e Kappa de 0,2001. Como podemos observar na Tabela 5 o classificador KNN foi capaz de identificar um número baixo de casos com doença, como mostra a sensibilidade de 56,98% e poucos casos sadios foram classificados como doentes, especificidade de 63,04%.

A Tabela 6 demostra os resultados para as imagens originais e máscara internas e externa, com e sem Equalização de Histograma. Cada massa é representada por 90 características, isto é, para os 10 (dez) índices propostos nesse método, sendo aplicado na imagem original obtendo 10 características, nas máscaras internas (20%,40%,60% e 80%) obtendo 40 características e externas (20%,40%,60% e 80%) obtendo 40 características.

Entre os resultados apresentados na Tabela 6, o classificador SMO obteve o melhor resultado nas imagens com realce, com uma taxa de acurácia de 70,48%, uma sensibilidade

Tabela 6: Resultados dos testes utilizando imagens normais e máscaras.

Imagens originais e Máscaras					
Sem Equalização de Histograma					
Classificador	**A(%)**	**S(%)**	**E(%)**	**ROC**	**Kappa**
Random Forest	69,18	66,42	71,52	0,758	0,379
SMO	69,61	66,42	72,32	0,694	0,387
MLP	68,23	67,55	68,80	0,721	0,362
KNN	60,78	58,30	62,88	0,594	0,211
Com Equalização de Histograma					
Classificador	**A(%)**	**S(%)**	**E(%)**	**ROC**	**Kappa**
Random Forest	69,09	65,09	72,48	0,761	0,376
SMO	**70,48**	**66,60**	**73,76**	**0,702**	**0,404**
MLP	66,15	65,28	66,88	0,701	0,320
KNN	61,90	61,32	62,40	0,612	0,236

de 66,60%, especificidade de 73,76%, uma curva ROC de 0,702 e uma Kappa de 0,404. O pior resultado foi do classificador KNN nas imagens sem realce, com taxa de acurácia de 60,78%, curva ROC de 0,594 e Kappa de 0,211. O classificador identificou um número baixo de casos com doença, como mostra a sensibilidade de 58,30% e poucos casos sadios foram classificados como doentes, especificidade de 62,88%.

A Tabela 7 representa os resultados para as imagens LBP, com e sem Equalização de Histograma. Cada massa é representada por 10 características.

Tabela 7: Resultados dos testes utilizando imagens LBP

Imagens LBP					
Sem Equalização de Histograma					
Classificador	**A(%)**	**S(%)**	**E(%)**	**ROC**	**Kappa**
Random Forest	**77,06**	**78,68**	**75,68**	**0,840**	**0,540**
SMO	72,03	91,70	55,36	0,735	0,455
MLP	75,93	78,30	73,92	0,827	0,518
KNN	71,95	72,83	71,20	0,734	0,438
Com Equalização de Histograma					
Classificador	**A(%)**	**S(%)**	**E(%)**	**ROC**	**Kappa**
Random Forest	76,19	78,30	74,40	0,835	0,523
SMO	71,08	94,91	50,88	0,729	0,440
MLP	74,37	77,17	72,00	0,804	0,488
KNN	67,27	66,23	68,16	0,593	0,343

Entre os resultados obtidos na Tabela 7, as imagens sem realce em conjunto com o classificador Random Forest obtiveram o melhor resultado, com uma taxa de acurácia de 77,06%, uma sensibilidade de 78,68%, especificidade de 75,68%, uma curva ROC de 0,840 e uma Kappa de 0,540. A imagens com realce em conjunto com o classificador KNN tiveram o pior resultado, com taxa de acurácia de 67,27%, curva ROC de 0,593 e Kappa de 0,343. Como podemos observar na Tabela 7 o classificador identificou um número baixo

de casos com doença, como mostra a sensibilidade de 66,23% e poucos casos sadios foram classificados como doentes, especificidade de 68,16%.

A Tabela 8 mostra os resultados para as imagens LBP e máscara interna e externa, com e sem Equalização de Histograma. Cada massa é representada por 90 características, isto é, para os 10 (dez) índices propostos nesse método, sendo aplicado na imagem original obtendo 10 características, nas máscaras internas (20%,40%,60% e 80%) obtendo 40 características e externas (20%,40%,60% e 80%) obtendo 40 características.

Tabela 8: Resultados dos testes utilizando imagens LBP e máscaras

Imagens LBP e Máscaras					
Sem Equalização de Histograma					
Classificador	**A(%)**	**S(%)**	**E(%)**	**ROC**	**Kappa**
Random Forest	**77,92**	**81,51**	**74,88**	**0,852**	**0,559**
SMO	74,46	88,68	62,40	0,755	0,498
MLP	77,75	84,72	71,84	0,833	0,558
KNN	71,00	70,19	71,68	0,717	0,417
Com Equalização de Histograma					
Classificador	**A(%)**	**S(%)**	**E(%)**	**ROC**	**Kappa**
Random Forest	77,40	80,19	75,04	0,851	0,548
SMO	76,10	82,64	70,56	0,766	0,525
MLP	72,64	75,09	70,56	0,806	0,453
KNN	67,10	65,47	68,48	0,680	0,339

Dos resultados exibidos na Tabela 8, o classificador Random Forest em conjunto com as imagens sem realce obteve o melhor resultado, com uma taxa de acurácia de 77,92%, uma sensibilidade de 81,51%, especificidade de 74,88%, uma curva ROC de 0,852 e uma Kappa de 0,559. O classificador KNN em conjunto com as imagens com realce tiveram o resultado menos significativo, com taxa de acurácia de 67,10%, curva ROC de 0,680 e Kappa de 0,339. Onde foi observado na Tabela 8, o classificador KNN foi capaz de identificar um número baixo de casos com doença, como mostra a sensibilidade de 65,10% e poucos casos sadios foram classificados como doentes, especificidade de 68,48%.

Na Tabela 9 mostra uma breve comparação entre os resultados encontrados utilizando a metodologia propostas e alguns trabalhos citados no Capítulo 2, que realizam a classificação de tecidos mamários em malignos e benignos. Portanto, a comparação com os trabalhos é meramente subjetiva, pois os casos analisados variam, assim como as bases de imagens utilizadas, sendo estes, fatores primordiais para uma comparação de confiança.

No que diz respeito às métricas de desempenho utilizadas neste trabalho, o ideal para um sistema CADx é ter um bom equilíbrio entre as três métricas de avaliação (acurácia, sensibilidade e especificidade), pois, um bom método deve ser capaz de classificar com sucesso, tantos os casos de positivos (que possuem a doença) como os casos de negativos (que não possuem a doença). Fica claro que na presente metodologia além dos valores de

Tabela 9: Comparação da metodologia com os trabalhos relacionados.

Trabalhos	**Bases**	**A(%)**	**S(%)**	**E(%)**	**ROC**	**Kappa**
(ROCHA et al., 2014)	DDSM	86,66	90	83,33	0,86	-
(PARAMKUSHAM et al., 2015)	IRMA	89,05	-	-	-	-
(ROUHI et al., 2015)	MIAS DDSM	96,47	96,87	95,94	-	-
(VALARMATHIE et al., 2016)	MIAS	100	-	-	0,99	-
(KANADAM; CHEREDDY, 2016)	MIAS	97,2	-	-	-	-
(RABIDAS et al., 2016)	MIAS	-	-	-	0,982	-
(LIMA et al., 2016)	IRMA	94,11	-	-	-	-
(LI et al., 2016)	DDSM	86,92	87,18	86,54	0,91	-
(CARVALHO et al., 2018)	DDSM MIAS	99,73 100	99,41 100	99,84 100	1,00	-
(WANG et al., 2018)	BCDR	85,00	-	-	0,89	-
Metodologia	**DDSM**	**77,92**	**81,51**	**74,88**	**0,852**	**0,559**

validação apresentarem equilíbrio, e de acordo com a curvar roc e a kappa, motivando sua aplicação de acordo com os seus valores

5.1 Discussões

Aqui apresentamos e discutimos os resultados produzidos pelos testes realizados utilizando o método proposto para diagnosticar massa mamária em maligno e benigno. Também realizou-se uma comparação do melhor resultado obtido com outros expostos na literatura.

Assim, utilizou-se técnicas para melhorar o contraste dos objetos de interesse, para prover uma melhor descrição da textura, o LBP se mostrou eficaz, pois em todos os teste, melhoram os resultados alcançados. Convém destacar a eficiência das máscaras em abranger padrões em áreas de borda e interna.

6 Conclusão

O presente estudo apresentou um método automático, com o uso dos Índices de Diversidade, junto com reconhecimentos de padrões, capaz de discriminar e classificar, regiões de tecidos da mama em maligno e benigno.

Os índices de diversidade, apresentaram-se moderado na tarefa de caracterização das regiões em maligno e benigno, apresentando uma taxa de acurácia de 77,92%, sensibilidade de 81,51%, especificidade de 74,88%, curva roc de 0,852 e kappa de 0,5592.

Os resultados obtidos nas imagens LBP, demonstraram o desempenho próspero das técnicas de extração de textura pelos índices de diversidade com o classificador Random Forest, sendo confirmados pelo resultados da estatística Kappa, que é representado na Tabela 4.

No entanto, este trabalho contribuiu de forma positiva em duas distintas áreas. Sendo que na área da saúde, disponibilizou uma metodologia automática para auxílio no diagnóstico de tecidos mamários em maligno e benigno, que apesar de não fornecer excelentes resultados mostram-se promissor. E na computação, com a adaptação de técnicas de outras áreas do conhecimento que serviram como descritores de textura, podendo então, serem utilizados em caracterizações em outros tipos de imagens.

E ainda, como sugestão para trabalhos futuros, pretendendo-se:

- Aumentar a quantidade e variabilidade das amostras de lesões utilizando mais bases de imagens;
- Aplicar os índices propostos para investigação de outras doenças; e,
- Aplicar os índices propostos em conjunto com outros descritores, inclusive, com abordagens de aprendizado profundo.

Referências

AMORIM, M. A. P.; SIQUEIRA, K. Z. Relação entre vivência de fatores estressantes e surgimento de câncer de mama. *Psicologia Argumento*, v. 32, n. 79, 2017. Citado na página 22.

ARAÚJO, J. D. L. et al. Diagnóstico de glaucoma em imagens de fundo de olho utilizando os índices de diversidade de shannon e mcintosh. In: SBC. *17º Workshop de Informática Médica (WIM 2017)*. [S.l.], 2017. v. 17, n. 1/2017. Citado na página 29.

AZEVEDO-MARQUES, P. M. de. Diagnóstico auxiliado por computador na radiologia. *Radiologia Brasileira*, SciELO Brasil, v. 34, n. 5, p. 285–293, 2001. Citado na página 23.

BATISTA, M. d. L. S. et al. Processamento digital de imagens para a detecção e classificação de nódulos em mamografias. 2010. Citado na página 15.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 31.

BRUNO, B. C. R. et al. Câncer de mama: É possível prevenir? *REVISTA UNINGÁ REVIEW*, v. 28, n. 1, 2018. Citado na página 22.

CAMARGO, J. Must dominance increase with the number of subordinate species in competitive interactions? *J Theor Biol*, v. 161, p. 537–542, 1993. Citado na página 29.

CARVALHO, D. S. de et al. Aspectos gerais epidemiológicos da mortalidade por câncer de mama feminino no brasil e no mundo. *Anais do Simpósio de Enfermagem*, v. 1, n. 1, 2019. Citado na página 15.

CARVALHO, E. D. et al. Method of differentiation of benign and malignant masses in digital mammograms using texture analysis based on phylogenetic diversity. *Computers & Electrical Engineering*, Elsevier, v. 67, p. 210–222, 2018. Citado 3 vezes nas páginas 19, 21 e 40.

CHUA, L. O.; YANG, L. Cellular neural networks: Theory. *IEEE Transactions on circuits and systems*, IEEE, v. 35, n. 10, p. 1257–1272, 1988. Citado na página 18.

FERREIRA, M. et al. Processamento digital de imagens mÃ©dicas com python e opencv. In: _____. [S.l.: s.n.], 2017. ISBN 978-85-8320-201-1 331. Citado 2 vezes nas páginas 24 e 25.

FILHO, A. O. d. C. et al. Métodos para sistemas cad e cadx de nódulo pulmonar baseada em tomografia computadorizada usando análise de forma e textura. Universidade Federal do Maranhão, 2016. Citado 3 vezes nas páginas 16, 24 e 31.

FILHO, O. M.; NETO, H. V. *Processamento digital de imagens*. [S.l.]: Brasport, 1999. Citado na página 24.

GUO, Y.; ZHAO, G.; PIETIKÄINEN, M. Discriminative features for texture description. *Pattern Recognition*, Elsevier, v. 45, n. 10, p. 3834–3843, 2012. Citado na página 19.

HEATH, M. et al. The digital database for screening mammography. In: MEDICAL PHYSICS PUBLISHING. *Proceedings of the 5th international workshop on digital mammography*. [S.l.], 2000. p. 212–218. Citado na página 34.

INCA. Tipos de câncer: mama. In: . [s.n.], 2019. Disponível em: <https://www.inca-.gov.br/tipos-de-cancer/cancer-de-mama>. Acesso em: 10.05.2019. Citado 2 vezes nas páginas 15 e 22.

KANADAM, K. P.; CHEREDDY, S. R. Mammogram classification using sparse-roi: A novel representation to arbitrary shaped masses. *Expert Systems with Applications*, Elsevier, v. 57, p. 204–213, 2016. Citado 3 vezes nas páginas 18, 21 e 40.

KELLER, J. M.; GRAY, M. R.; GIVENS, J. A. A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*, IEEE, n. 4, p. 580–585, 1985. Citado na página 31.

LI, Y. et al. Mass classification in mammograms based on two-concentric masks and discriminating texton. *Pattern Recognition*, Elsevier, v. 60, p. 648–656, 2016. Citado 3 vezes nas páginas 19, 21 e 40.

LIMA, S. M. de et al. Detection and classification of masses in mammographic images in a multi-kernel approach. *Computer methods and programs in biomedicine*, Elsevier, v. 134, p. 11–29, 2016. Citado 3 vezes nas páginas 19, 21 e 40.

MARGALEF, R. Diversity and stability: a practical proposal and a model of interdependence. 1969. Citado na página 29.

MAZALA, T. T. Perfil da mortalidade por câncer de mama em sete lagoas–mg no período de 2011-2015. *Revista Brasileira de Ciências da Vida*, v. 7, n. Especial, p. 38–42, 2019. Citado na página 15.

MCHUGH, M. L. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, Medicinska naklada, v. 22, n. 3, p. 276–282, 2012. Citado na página 33.

MCINTOSH, R. P. An index of diversity and the relation of certain concepts to diversity. *Ecology*, Wiley Online Library, v. 48, n. 3, p. 392–404, 1967. Citado na página 29.

NASCIMENTO, L. B. et al. Classificação de nódulos pulmonares em maligno e benigno utilizando os índices de diversidade de shannon e de simpson. Universidade Federal do Maranhão, 2012. Citado 2 vezes nas páginas 26 e 28.

PAL, S. K.; MITRA, S. Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on neural networks*, IEEE, v. 3, n. 5, p. 683–697, 1992. Citado na página 31.

PARAMKUSHAM, S. et al. Novel technique for the detection of abnormalities in mammograms using texture and geometric features. In: IEEE. *2015 International Conference on Microwave, Optical and Communication Engineering (ICMOCE)*. [S.l.], 2015. p. 150–153. Citado 3 vezes nas páginas 18, 21 e 40.

PIRES, A.; NETO, G. Compound local binary pattern para reconhecimento de expressoes faciais. *Universidade Federal do Maranhao, SA£ o Luis, Brasil*, 2015. Citado na página 27.

PIRES, G. M. et al. Arquitetura para um sistema de diagnóstico auxiliado por computador. *Departamento de Informática, Universidade Federal da Paraíba–Brasil*, 2006. Citado na página 24.

PLATT, J. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998. Citado na página 31.

RABIDAS, R. et al. Benign-malignant mass classification in mammogram using edge weighted local texture features. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Medical Imaging 2016: Computer-Aided Diagnosis*. [S.l.], 2016. v. 9785, p. 97851X. Citado 3 vezes nas páginas 19, 21 e 40.

ROCHA, S. V. d. et al. Texture analysis of masses in digitized mammograms using gleason and menhinick diversity indexes. *Revista Brasileira de Engenharia Biomédica*, SciELO Brasil, v. 30, n. 1, p. 27–34, 2014. Citado 3 vezes nas páginas 18, 21 e 40.

ROCHA, S. V. d. et al. Diferenciação do padrão de malignidade e benignidade de massas em imagens de mamografias usando padrões locais binários, geoestatística e índice de diversidade. Universidade Federal do Maranhão, 2014. Citado 2 vezes nas páginas 22 e 25.

ROUHI, R. et al. Benign and malignant breast tumors classification based on region growing and cnn segmentation. *Expert Systems with Applications*, Elsevier, v. 42, n. 3, p. 990–1002, 2015. Citado 3 vezes nas páginas 18, 21 e 40.

SAÚDE, M. D. Ministério da saúde: Câncer de mama. In: . [s.n.], 2019. Disponível em: <http://portalms.saude.gov.br/saude-de-a-z/cancer-de-mama>. Acesso em: 10.05.2019. Citado na página 23.

SHANNON, C. E.; WEAVER, W. *The mathematical theory of communication (Urbana, IL*. [S.l.]: University of illinois Press IL, 1949. Citado na página 28.

SIMPSON, E. H. Measurement of diversity. *Nature*, Nature Publishing Group, v. 163, n. 4148, p. 688, 1949. Citado na página 28.

SMITH, B.; WILSON, J. B. A consumer's guide to evenness indices. *Oikos*, JSTOR, p. 70–82, 1996. Citado na página 30.

VALARMATHIE, P. et al. Classification of mammogram masses using selected texture, shape and margin features with multilayer perceptron classifier. Biomedical Research, 2016. Citado 3 vezes nas páginas 18, 21 e 40.

WANG, H. et al. Breast mass classification via deeply integrating the contextual information from multi-view data. *Pattern Recognition*, Elsevier, v. 80, p. 42–52, 2018. Citado 3 vezes nas páginas 20, 21 e 40.

TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DIGITAL NA BIBLIOTECA "JOSÉ ALBANO DE MACEDO"

Identificação do Tipo de Documento

( ) Tese

( ) Dissertação

(x) Monografia

( ) Artigo

Eu, **Igor Miranda Grangeiro**, autorizo com base na Lei Federal nº 9.610 de 19 de Fevereiro de 1998 e na Lei nº 10.973 de 02 de dezembro de 2004, a biblioteca da Universidade Federal do Piauí a divulgar, gratuitamente, sem ressarcimento de direitos autorais, o texto integral da publicação **Classificação automática de tecidos mamários em malignos e benignos utilizando Índices de diversidade** de minha autoria, em formato PDF, para fins de leitura e/ou impressão, pela internet a título de divulgação da produção científica gerada pela Universidade.

Picos-PI 05 de Julho de 2019.

Igor Miranda Grangeiro

Assinatura