

# Diferenciação de padrões de benignidade e malignidade em tecidos da mama baseado na diversidade taxonômica

Edson Damasceno Carvalho<sup>1</sup>, Antônio Oseas de C. Filho<sup>1</sup>, Alcilene D. de Sousa<sup>1</sup>,  
Patricia Vieira da S. Barros<sup>1</sup>, Patrícia Medyna L. de L. Drumond<sup>1</sup>

<sup>1</sup>Campus Senador Helvídio Nunes de Barros – Universidade Federal do Piauí (UFPI)  
Picos – PI – Brasil

{edsondamasceno11, antoniooseas, alcileneeluzsousa, petibarrosthe}@gmail.com

patymedy@hotmail.com

**Abstract.** *The breast cancer is a disease resulting from the multiplication of abnormal cells of the breast, forming the masses. The tracking through the mammography is the most promising way for the precocious diagnosis. This work presents a method of classification of malignant and benign breast tissues in mammography. In this method, we used texture descriptors based on phylogenetic indexes for extraction of characteristics, then made a classification using the classifiers: J48, RandomForest, J48Consolidated e LMT. The results reached an accuracy of 94,8%, sensibility of 92,9%, specificity of 96,5% and a ROC curve 0,988. The use of the phylogenetic indexes to describe patterns in areas of mammograms images was shown efficient in the categorization of malignant and benign.*

**Resumo.** *O câncer de mama é uma doença resultante da multiplicação de células anormais da mama, formando as massas. O rastreamento por meio da mamografia é o meio mais promissor para o diagnóstico precoce. Este trabalho apresenta um método de classificação de tecidos da mama em maligno e benigno em exames de mamografia. Neste método foram usados descritores de textura baseado em índices filogenéticos para extração de características, em seguida feita uma classificação usando os classificadores: J48, RandomForest, J48Consolidated e LMT. Os resultados alcançaram uma acurácia de 94,8%, sensibilidade de 92,9%, especificidade de 96,5% e uma curva ROC de 0,988. O uso dos índices filogenéticos para descrever padrões em regiões de imagens de mamografias mostrou-se eficiente na categorização de maligno e benigno.*

## 1. Introdução

Segundo o Instituto Nacional do Câncer (INCA), o câncer de mama é o mais comum entre as mulheres no mundo e no Brasil, depois do câncer de pele não melanoma, respondendo por cerca de 25% dos casos novos por ano [INCA 2016].

A detecção precoce do câncer de mama, em grande parte dos casos aumenta as chances de tratamento e cura, podendo ser realizado através da mamografia em mulheres que não possuem sinais e/ou sintomas da doença. No Brasil é recomendado que as mulheres entre 50 e 69 anos façam a mamografia a cada período de 2 anos, em casos que há um histórico de familiares com câncer de mama, deve haver um acompanhamento mais específico.

Com o uso da mamografia para detecção do câncer de mama, foi observada uma redução nas taxas de mortalidades em decorrência do mesmo. Vários fatores podem influenciar na sensibilidade do exame, afetando a qualidade do mesmo. Esses fatores resultam em possíveis erros de diagnósticos pelos radiologistas que podem variar de 10% a 30% dos casos [Braz Júnior 2014]. Na medicina o uso de imagens é um recurso muito importante para o diagnóstico de anomalias. O processamento digital de imagens estuda requisitos para extrair informações técnicas a fim de melhorar os parâmetros para um diagnóstico mais preciso, aumentando a interpretação da mamografia [Silva 2016].

No processamento digital de imagens, a análise de imagens de mamografias pode ser realizada através de textura e forma [Silva 2016]. A análise de textura é caracterizada por variações locais em valores de *pixels* que se repetem de maneira regular ou aleatória ao longo da imagem. A análise da forma caracteriza a geometria dos objetos, como o tamanho, a curvatura e a suavidade dos contornos.

A análise feita através de texturas é útil em aplicações, por se aproximar da avaliação feita pelo sistema visual humano. Cada textura possui um padrão de características. Em imagens mamográficas, atributos de texturas fornecem uma descrição dos *pixels* do tecido mamário, sendo muito importante para descrever tais regiões.

O trabalho proposto tem como objetivo apresentar o desenvolvimento de uma nova abordagem para classificação de tecidos da mama em maligno e benigno, usando os índices de diversidade filogenéticas para extração de características de textura, que foram responsáveis por caracterizar as regiões de interesse. Em seguida, foi feita uma classificação através de múltiplos classificadores.

Este trabalho está dividido como se segue: na Seção 2, são apresentados os trabalhos relacionados; na Seção 3 descreve-se a metodologia proposta; na Seção 4, são descritos os resultados da execução do projeto; e por fim, são apresentadas as conclusões na Seção 5.

## 2. Trabalhos Relacionados

Na literatura especializada, existem trabalhos relacionados à classificação de tecidos da mama em maligno e benigno em exames de mamografia. Para este propósito, utilizam-se características extraídas de imagens médicas, as quais servem como vetores de entrada para os classificadores.

Silva (2016) apresenta uma metodologia para classificação de massas em maligno e benigno, a qual utiliza a base DDSM (*Digital Database for Screening Mammography*). As características são extraídas utilizando as técnicas de *Local Binary Pattern (LBP)*, Função K de *Ripley* e os índices de *Shannon*, *Mcintosh*, *Simpson*, *Gleason* e de *Menhinick*, para extração das características de forma das imagens de mamografias. Para classificação utilizou a Máquina de Vetor de Suporte (SVM), onde as imagens foram divididas em quatro grupos de acordo com a sua densidade, e um quinto grupo contendo todas as imagens. Após os testes obteve como resultado, uma acurácia de 93,70%, sensibilidade 96,29% e especificidade 91,05% para o grupo de densidade 2. Como melhor resultado entre os quatro grupos de densidade, e uma acurácia de 90,18%, sensibilidade de 91,01% e especificidade de 89,94% para o quinto grupo, com todas as imagens.

Fahssi *et al.* (2015) apresenta uma abordagem para classificação das massas

mamárias em maligno e benigno, usando 111 imagens da base mini-MIAS (*Mini Mammographic Image Analysis Society*), sendo 60 benignas e 51 malignas. A abordagem utilizada consiste na Teoria da Transformação Ortogonal Adaptativa que calcula as características informativas das regiões de interesse. A classificação foi realizada por meio da comparação da similaridade entre os vetores das características das regiões de interesse pelo uso do coeficiente da matriz de correlação. Para avaliar a efetividade da metodologia, foram utilizadas as informações fornecidas pela *Mammographic Image Analysis Society* (MIAS), incluindo a classe de imagens e coordenadas de seus centros de regiões de interesse. A metodologia apresentou como resultado um percentual de 93,78% de sensibilidade e 94,54% de especificidade.

Melo *et al.* (2014) apresenta uma abordagem para classificação de massas em maligno e benigno em imagens de mamografias, utilizando a base de imagens DDSM. Foi realizada a segmentação das imagens utilizando um algoritmo baseado na técnica de crescimento de região e árvore de decisão. Os atributos foram extraídos a partir da forma das massas, utilizando a ferramenta *Matlab*. Para classificação, utilizou-se os algoritmos de aprendizagem de Máquina *Multilayer Perceptron*, o *Radial Basis Function Network (RBFN)*, *Naive Bayes*, *K-Nearest Neighbor (KNN)*, o *RandomForest* e o *Functional Trees (FT)*. Após os testes, obteve como melhor resultado, o classificador *Multilayer Perceptron*, apresentando uma acurácia de 85,09%. sensibilidade de 77,59% e especificidade de 89,32%.

Rocha (2014) usa LBP, Geoestatística e Índice de Diversidade para extração das características de textura das imagens de mamografias, utilizando a base DDSM. Para classificação das massas em maligno e benigno utilizou-se a Máquina de Vetor de Suporte (SVM). O melhor resultado obtido com os testes foram os valores de 92,20% para acurácia, 92,26% para sensibilidade, 91,26% para especificidade, 10,63 de razão de probabilidade positiva, 0,07% de razão de probabilidade negativa e uma área sob a curva ROC de 0,92.

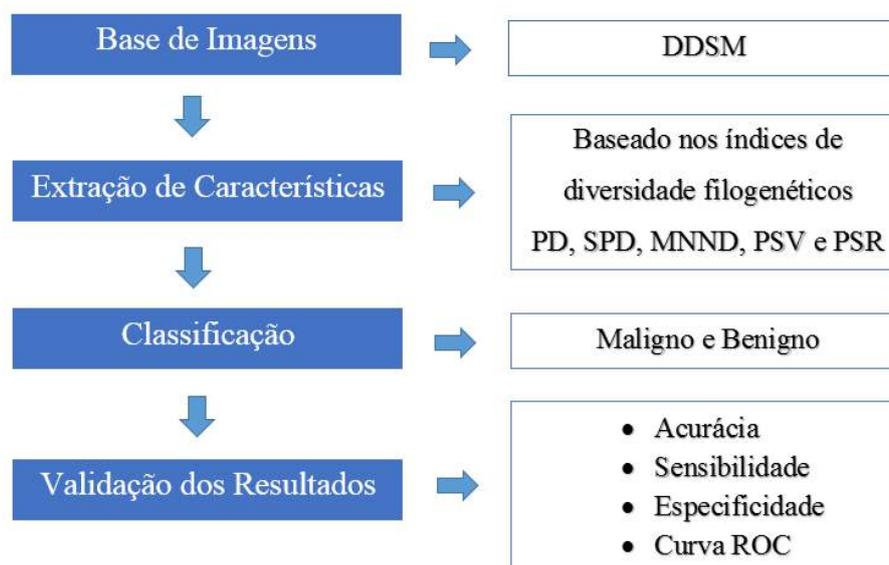
Diante do que foi exposto nos trabalhos relacionados, verificou-se que houve a utilização de vários métodos para extração dos atributos, contando com a intervenção humana e, observou-se a necessidade de melhorar os resultados. Este trabalho tem como objetivo a classificação de massas em maligno e benigno, baseado em características do comportamento dentro de uma comunidade, como o parentesco entre espécies e sua riqueza.

### 3. Materiais e métodos

Para que fosse possível classificar massas em malignas e benignas em imagens de mamografias, foi empregada a seguinte metodologia: Primeiramente, foi feita a aquisição das imagens oriundas da base *Digital Database for Screening Mammography* (DDSM). Seguindo, foi realizada a etapa de extração de características utilizando descritores baseados na textura, através dos índices de diversidade filogenéticas e para a classificação utilizados os classificadores *J48*, *RandomForest*, *J48Consolidated* e LMT. Na Figura 1 é apresentado um resumo de cada etapa.

#### 3.1 Anatomia e Patologia da mama

A mama é formada principalmente por tecido granular, que é a parte responsável pela produção de leite durante o período de amamentação, constituído principalmente



**Figura 1. Metodologia proposta**

por lóbulos, dutos e o tecido de suporte ou conjuntivo, composto por tecidos adiposos e conectivos fibrosos, responsáveis por manter a forma e sustentação da mama [Braz Júnior 2014]. Normalmente, o câncer de mama começa nas células dos lóbulos, ou nos dutos, menos frequentemente, nos tecidos que incluem a gordura e tecidos conjuntivos fibrosos [Silva 2016].

O câncer de mama é formado pela multiplicação de forma desordenada de suas células, formando um tumor maligno, que é um agrupamento de células cancerígenas que podem invadir tecidos adjacentes ou disseminar por outras partes do corpo, tendo como um dos principais fatores de risco o envelhecimento, apresentando uma taxa de 12,5% em mulheres com menos de 45 anos.

Já a Mamografia é a radiologia da mama, que tem por objetivo produzir imagens detalhadas das estruturas internas da mama, permitindo a detecção precoce do câncer, por ser capaz de mostrar lesões muito pequenas, ainda em seu estágio inicial [Silva 2016] e [Rocha 2014].

A mamografia trabalha com intervalos específicos de níveis de radiação, com a finalidade de registrar a imagem da mama, fornecendo uma ampla gama de densidade e maior resolução de contraste, sendo muito importante, devido as estruturas normais e doentes da mama possuírem uma diferença de contraste muito pequenas, com isso a mamografia realça tais diferenças, fornecendo uma resolução de alto contraste.

### 3.2 Base de 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. A base DDSM foi uma doação do *Breast Cancer Research Program of the U.S. Army Medical Research and Materiel Comman*. O banco possui 2500 estudos. Cada estudo contém duas imagens de cada mama, juntamente com informações do paciente, tais como classificação da densidade da mama, sutileza para

anormalidades e informações de imagens, como resolução espacial e imagens contendo áreas suspeitas [Heath et al. 2000].

Este trabalho utiliza *Region of Interest* (ROI) extraídas de imagens da base DDSM, onde cada ROI possui 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.

### 3.3 Índices de Diversidade Filogenética

A filogenia é um ramo da biologia responsável pelo estudo das relações evolutivas entre as espécies, pela verificação dos relacionamentos entre elas, a fim de determinar possíveis ancestrais comuns [Webb 2000]. Diversidade filogenética é uma medida de uma comunidade que incorpora as relações filogenéticas das espécies [Magurran 2004]. A forma mais simples da aplicação do índice de diversidade em imagens é quando a comunidade representa uma imagem ou região da mesma, as espécies sendo os níveis de cinza, os indivíduos sendo os pixels e as distâncias filogenética sendo os número de arestas entre duas espécies [Oliveira 2013].

Neste trabalho, são utilizados cinco índices de diversidade filogenética para extração dos descritores de textura: *Phylogenetic Diversity* (PD), *Sum of Phylogenetic Distances* (SPD), *Mean Nearest Neighbour Distance* (MNND), *Phylogenetic Species Variability* (PSV) e *Phylogenetic Species Richness* (PSR).

O índice de diversidade filogenética PD é o somatório dos comprimentos dos ramos da filogenia de cada espécie, mostrada na Equação 1, onde  $B$  é o número de ramificação da árvore,  $L_i$  é o comprimento do ramo e  $A_i$  é a abundância média de espécies que compartilham ramo  $i$ .

$$B \times \frac{\sum_i^B L_i A_i}{\sum_i^B A_i} \quad (1)$$

O índice de diversidade filogenética SPD é soma das distâncias filogenéticas entre cada par de espécies, que pode ser observado na Equação 2, onde  $d_{mn}$  é a distância entre as espécies  $m$  e  $n$ ;  $a_m$ , abundância de espécies  $m$ ;  $S$ , número de espécies no conjunto focal.

$$\left( \frac{S(S-1)}{2} \right) \times \frac{\sum \sum m < n^2 m n^a m^a n}{\sum \sum m < n^a m^a n} \quad (2)$$

O índice de diversidade filogenética MNND é a distância média do táxon mais próximo, definido na Equação 3. Onde  $d_{mn}$  é a distância entre as espécies  $m$  e  $n$ ;  $a_m$ , abundância de espécies  $m$ .

$$\sum_m^S m i n (d_{mn}) a_m \quad (3)$$

O índice de diversidade filogenética PSV é a variabilidade de espécies filogenéticas, que resume o grau em que as espécies em uma comunidade são filogeneticamente relacionadas, definido na Equação 4, onde, o  $trC$  representa a soma dos valores

da diagonal de uma matriz  $C$ ,  $\sum c$  é o somatório de todos os valores da matriz,  $n$  é o número de espécies e  $'c$  é a média dos elementos da diagonal de  $C$ .

$$PSV = \frac{n \operatorname{tr} C - \sum c}{n(n-1)} = 1 - 'c \quad (4)$$

O índice de diversidade filogenética PSR é a riqueza de espécies e quantifica o número de espécies em uma comunidade. O valor do PSR é encontrado multiplicando-se o número de espécies  $n$  pela variabilidade da comunidade, pode-se verificar os parâmetros na Equação 5.

$$PSR = nPSV \quad (5)$$

### 3.4 Classificação

A classificação foi realizada pela suíte de algoritmos de mineração de dados e Aprendizado de Máquina WEKA, que contém ferramentas para pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização. A classificação também é adequada para o desenvolvimento de novos esquemas de Aprendizado de Máquinas [WEKA 2017].

Foram utilizados os algoritmos *J48* [Quinlan 1993], *RandomForest* [Breiman 2001], *J48Consolidated* [Pérez et al. 2007] e *LMT* [Landwehr et al. 2005], 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.

Foram utilizados os classificadores *J48*, *J48Consolidated*, *LMT* e *RandomForest* devido a utilização de descritores baseados em comportamentos dentro de uma comunidade. Classificadores baseados em Árvores, são robustos em relação a ruídos, apresentando bons resultados na categorização de padrões de malignidade e benignidade em imagens médicas.

### 3.5 Validação dos Resultados

A fim de se considerar a presença ou ausência de massas malignas e benignas em imagens de mamografia, para validação dos resultados, utilizou-se de métricas de avaliação baseadas em estatísticas como, Sensibilidade (S) Especificidade (E) e Acurácia (A).

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, sobre um determinado conjunto de exemplo, como mostra a Tabela 1.

A sensibilidade (S), Equação 6, é a capacidade de um teste diagnóstico identificar os verdadeiros positivos nos indivíduos verdadeiramente doentes. Quando um teste é sensível raramente deixa de encontrar pessoas com a doença.

**Tabela 1. Matriz de Confusão**

Resultado do teste	Doença	
	Presente	Ausente
Positivo	<b>VP</b> Verdadeiro positivo	<b>FP</b> Falso Positivo
Negativo	<b>FN</b> Falso Negativo	<b>VN</b> Verdadeiro Negativo

$$S = \frac{VP}{VP + FN} \quad (6)$$

Especificidade (E), Equação 7, 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.

$$E = \frac{VN}{VN + FP} \quad (7)$$

Acurácia (A), Equação 8, é a proporção de acertos, ou seja, o total de verdadeiramente positivos e verdadeiramente negativos, em relação a amostra estudada.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \quad (8)$$

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, maior é a probabilidade do sistema fazer uma decisão correta [Silva 2016].

A curva ROC representa a dependência entre a sensibilidade e a especificidade de um classificador [Rocha 2014]. 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, como representado na Figura 2.

#### 4. Resultados e Discussões

Para os testes realizados neste trabalho, utilizou-se a base DDSM, descrita na Subseção 3.2. A extração de características de textura foi realizada a partir dos índices de diversidade filogenéticas como especificado na Subseção 3.3 e a classificação das massas em maligno e benigno, utilizou os classificadores e parâmetros definidos na Subseção 3.5.

De acordo com a Tabela 2, o classificador *RandomForest* obteve o melhor resultado, com uma taxa de acertos de 94,8%, uma sensibilidade de 92,9%, especificidade de 96,5% e uma curva ROC de 0,988, considerada ótima, indicando que o classificador apresenta boa capacidade de diagnóstico. A curva ROC apresentada pelo *RandomForest* indica que o classificador consegue detectar um número auto de casos positivos. As taxas

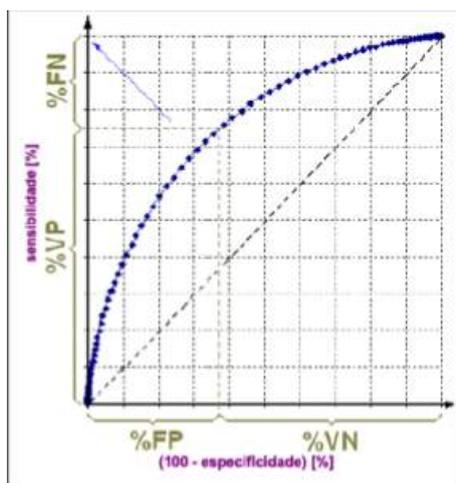


Figura 2. Representação da curva ROC. FONTE: [Rocha 2014]

Tabela 2. Resultados da classificação

Classificador	Base	Acurácia (%)	Sensibilidade (%)	Especificidade (%)	Curva ROC
<b>J48</b>	<b>DDSM</b>	<b>93,7</b>	<b>91,6</b>	<b>95,5</b>	<b>0,957</b>
J48Consolidated	DDSM	93,8	92,1	95,3	0,955
LMT	DDSM	93,8	92,1	95,3	0,971
<b>RandomForest</b>	<b>DDSM</b>	<b>94,8</b>	<b>92,9</b>	<b>96,5</b>	<b>0,988</b>

de erros apresentadas pelo *RandomForest* são muito baixos, devido ser mais robusta a presença de ruídos, possibilitando um melhor resultado na classificação.

O J48 apresentou o pior resultado, com uma taxa de acerto de 93,7% e curva ROC de 0,957. Um fator que leva a erros de classificação é a estrutura normal e doente da mama possuírem características muito parecidas, levando a classificar dados em classes diferentes. Como podemos observar na Tabela 2 o classificador *J48* foi capaz de identificar um número baixo de casos com doença, como mostra a sensibilidade de 91,6% e poucos casos saudáveis foram classificados como doentes, especificidade de 95,5%.

Tabela 3. Comparação da metodologia com os trabalhos relacionados

Trabalho	Base	Acurácia (%)	Sensibilidade (%)	Especificidade (%)	Curva ROC
Silva (2016)	DDSM	90,18	91,01	89,94	0,96
Melo <i>et al.</i> (2014)	DDSM	85,09	77,59	89,32	-
Fahssi <i>et al.</i> (2015)	MIAS	-	93,78	94,54	-
Rocha (2014)	DDSM	92,2	92,26	91,26	0,92
<b>Metodologia</b>	<b>DDSM</b>	<b>94,8</b>	<b>92,9</b>	<b>96,5</b>	<b>0,988</b>

Como pode-se observar na Tabela 3, a metodologia proposta utilizando os índices PD, SPD, MNND, PSV e PSR apresentou resultados melhores em relação aos trabalhos do estado da arte, com uma acurácia de 94,8%, sensibilidade de 92,9% e especificidade de 96,5%. A metodologia proposta apresentou um resultado bem significativo em relação ao

trabalho de [Melo et al. 2014] e em relação ao trabalho de [Fahssi et al. 2015] apresentou uma sensibilidade muito próxima e uma especificidade maior. Comparando a metodologia com os trabalhos de [Silva 2016] e [Rocha 2014] pode-se observar uma melhoria nos resultados obtidos. A curva ROC apresentada pela metodologia desenvolvida é maior que os demais trabalhos, indicando que a metodologia desenvolvida consegue detectar um número auto de casos positivos. Não é possível fazer uma comparação justa entre os trabalhos, primeiramente porque as bases não são as mesmas e nem apresentam o mesmo número de casos.

## 5. Conclusão

O uso de sistemas computadorizados para diagnósticos de anomalias em imagens médicas, vem se tornando cada vez mais frequente, visto que, em muitos casos, os dados serem de difícil avaliação por um especialista.

A análise de imagens de mamografias por meio dos Índices de Diversidade Filogenética, mostrou-se eficiente na categorização de regiões em maligno e benigno. A metodologia proposta neste trabalho apresentou uma taxa de acerto de 94,6%, sensibilidade de 92,9%, especificidade de 96,5% e uma curva ROC de 0,988. A metodologia desenvolvida consegue detectar um número auto de casos que são positivos, ou seja, possuem massas malignas, como mostra a curva ROC.

A partir dos resultados obtidos, pode-se inferir que a utilização de descritores baseados na textura, apresentam resultados eficazes na classificação de tecidos da mama em maligno e benigno. O uso de índices filogenéticos para descrever padrões em regiões de imagens mostrou-se eficiente para a metodologia proposta.

## Referências

- Braz Júnior, G. (2014). Detecção de regiões de massas em mamografias usando índices de diversidade, geoestatística e geometria côncava. Tese de Doutorado. Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Maranhão. São Luís - MA.
- Breiman, L. (2001). "Random forests". *Machine Learning*, v. 45, n. 1, p. 5-32.
- Fahssi, K. E., Elmoufidi, A., Abenaou, A., Jai-Andaloussi, S., e Sekkaki, A. (2015). Benign or malignant lesion classification in mammography images using the adaptive orthogonal transformation and the coefficients of the correlation matrix. *Recent Advances in Electrical Engineering*, ISBN: 978-1-61804-351-1.
- Heath, M., Bowyer, K., e Kopans, D. (2000). The digital database for screening mammography. *Citeseer, Proceedings of the 5th international workshop on digital mammography*. p. 212-218.
- INCA (2016). Tipos de câncer: mama. Disponível em: [http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama/cancer\\_mama](http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama/cancer_mama). Acesso em: 11/11/2016.
- Landwehr, N., Hall, M., e Frank, E. (2005). Logistic model trees. *Machine Learning*, 95(1-2):161-205.
- Magurran, A. E. (2004). Measuring biological diversity. *African Journal of Aquatic Science*; v. 29, n. 2, p. 285-286.

- Melo, M. C., Gajadhar, A., e Batista, L. V. (2014). Análise comparativa de métodos de aprendizagem de máquina para classificação de massas em mamografias. In: Congresso da Sociedade Brasileira de Computação - Workshop de Informática Médica, 2014, Brasília. Anais do XXXIV Congresso da Sociedade Brasileira de Computação - XIV Workshop de Informática Médica. Brasília: Universidade de Brasília. v. 1. p. 1772-1775.
- Oliveira, F. S. S. (2013). Classificação de tecidos da mama em massa e não-massa usando Índice de diversidade taxonômico e máquina de vetores de suporte. Dissertação de Mestrado. Curso de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão. São Luís - MA.
- Pérez, J. M., Maguerza, J., Arbelaitz, O., Gurrutxaga, I., e Martín, J. I. (2007). Combining multiple class distribution modified subsamples in a single tree. *Pattern Recognition Letters*, 28(4): 414-422.
- Quinlan, R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, CA.
- Rocha, S. V. (2014). Diferenciação do padrão de malignidade e benignidade de massas em imagens de mamografias usando padrões locais binários, geoestatística e Índice de diversidade. Tese de Doutorado. Curso de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão. São Luís - MA.
- Silva, T. F. (2016). Diferenciação do padrão de malignidade e benignidade de massas em mamografias utilizando características geométricas e máquina de vetor de suporte. Dissertação de Mestrado. Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Maranhão. São Luís - MA.
- Webb, C. O. (2000). Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *The American Naturalist* 156.2, p. 145-155.
- WEKA (2017). Machine learning group at the university of waikato. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 15 de Março de 2017.