Edson Damasceno Carvalho

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

Picos - PI
13 de Novembro de 2017

Edson Damasceno Carvalho

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

Monografia submetida ao Curso de Bacharelado em Sistemas de Informação como requisito parcial para obtenção de grau de Bacharel em Sistemas de Informação. Orientador: Prof. Dr. Antonio Oseas de Carvalho Filho

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

Picos - PI
13 de Novembro de 2017

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

**C331d** Carvalho, Edson Damasceno

Diferenciação de padrões de benignidade e malignidade em tecidos da mama baseado na diversidade taxonômica / Edson Damasceno Carvalho.– 2017.

CD-ROM : il.; 4 ¾ pol. (26 f.)

Trabalho de Conclusão de Curso (Curso Bacharelado em Sistemas de Informação) – Universidade Federal do Piauí, Picos, 2018.
Orientador(A): Prof. Dr. Antônio Oseas de Carvalho Filho

1. Câncer de Mama 2. Diversidade Filogenética. 3. Tecidos Maligno e Benigno. I. Título.

**CDD 005**

DIFERENCIAÇÃO DE PADRÕES DE BENIGNIDADE E MALIGNIDADE EM TECIDOS DA MAMA BASEADO NA DIVERSIDADE TAXONÔMICA

EDSON DAMASCENO CARVALHO

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

Data de Aprovação

Picos – PI. 27 de novembro de 2017

Prof. Dr. Antonio Oseas de Carvalho Filho
Orientador

Prof. Esp. Ismael de Holanda Leal
Membro

Leonardo Pereira de Sousa
Prof. Esp. Leonardo Pereira de Sousa
Membro

# Agradecimentos

Agradeço primeiramente a Deus por conceder a realização deste trabalho. A Ele, toda a gratidão.

Aos meus pais, Rodrigo Lopes de Carvalho e Jeruza Francisca Damasceno Carvalho pelo incentivo e trabalho de ambos cheguei até aqui.

A meu irmão Edelson Damasceno Carvalho.

À UFPI pelo apoio durante os 4 anos de curso.

Ao meu orientador Profº. 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.

As Profª. Ma. Patricia Medyna Lauritzen de Lucena Drumond, Profª. Ma. Alcilene Dalília de Sousa e Profª. Ma. Patricia Vieira da Silva Barros pela disponibilidade e disposição a ajudar-me, meus eternos agradecimentos.

Aos meus professores de graduação, muito obrigado.

Aos meus amigos de Iniciação Científica Marcos Vinícius, Antônio Carvalho e Kelly Maria pelo apoio e esclarecimentos às minhas dúvidas.

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

*Sonhos determinam o que você quer. Ação determina o que você conquista.*

*Aldo Novak*

# 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.

**Palavras-chaves**: Câncer de mama, Reconhecimento de padrões, Diversidade filogenética).

# 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.

**Key words**: Breast Cancer, Pattern Recognition, Phylogenetic Diversity).

# Lista de ilustrações

# Lista de tabelas

# Lista de abreviaturas e siglas

| | |
|---|---|
| A | Acurácia |
| DDSM | Digital Database for Screening Mammography |
| E | Especificidade |
| FN | Falso Negativo |
| FP | Falso Positivo |
| FT | Functional Trees |
| INCA | Instituto Nacional do Câncer |
| KNN | K-Nearest Neighbor |
| LBP | Local Binary Pattern |
| LMT | Logistic Model Tree |
| MVS | Máquina de Vetor de Suporte |
| MIAS | Mamographic Image Analysis Society |
| MNND | Mean Nearest Neighbour Distance |
| mini-MIAS | Mini Mamographic Image Analysis Society |
| PD | Phylogenetic Diversity |
| PSV | Phylogenetic Species Variability |
| PSR | Phylogenetic Species Richness |
| RBFN | Radial Basis Function Network |
| ROC | Receiver Operating Characteristic |
| ROI | Region of Interest |
| S | Sensibilidade |
| SPD | Sum of Phylogenetic Distances |
| VN | Verdadeiro Negativo |
| VP | Verdadeiro Positivo |
| WEKA | Waikato Environment for Knowledge Analysis |

# Sumário

# 1 Introdução

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. Segundo o Instituto Nacional do Câncer (INCA), o câncer de mama é o mais comun 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 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.

Em (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 (MVS), 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.

No trabalho de (FAHSSI et al., 2015) apresenta uma abordagem para classificação das massas mamárias em maligno e benigno, usando 111 imagens da base mini-MIAS (*Mini Mamographic 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 *Mamographic 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.

Em (MELO; GAJADHAR; BATISTA, 2014) apresentam 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%.

Em (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 (MVS). 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.

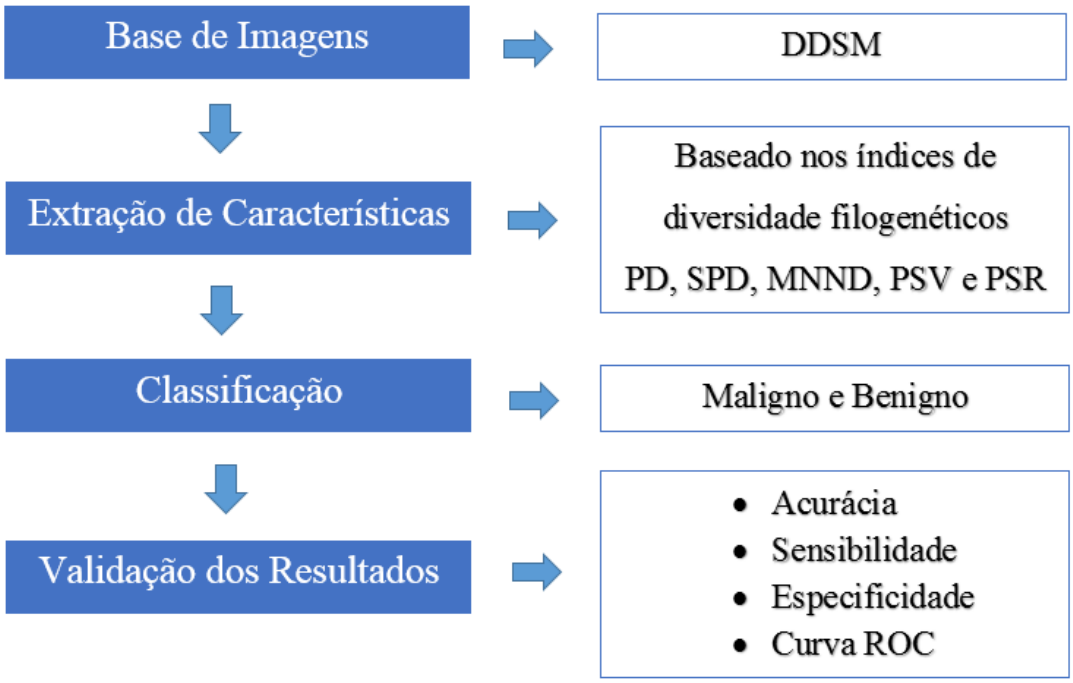



Figura 1: Metodologia proposta

## 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 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 estagio 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 contrate.

## 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; BOWYER; KOPANS, 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 Extração de Características

Após a aquisição das ROIs, as massas são submetidas à fase de extração de características baseada em textura. Para descrever a textura dos nódulos, utilizamos índices que calculam a pura diversidade filogenética, a soma das distâncias filogenéticas, a distância média do táxon mais próximo, a variabilidade de espécies filogenéticas e a riqueza de espécies filogenéticas. Nas seções seguintes, descrevemos os fundamentos de como a árvore foi organizada e dos índices de diversidade filogenética.

### 3.3.1 Árvore Filogenética - Cladograma

As árvores filogenéticas são utilizadas na biologia para descrever as relações filogenéticas evolutivas entre as espécies. Uma árvore filogenética, ou simplesmente filogenia, é uma árvore onde as folhas representam os organismos e os nós internos representam supostos ancestrais. As arestas da árvore denota as relações evolutivas. Para construir e organizar a árvore, precisamos estar cientes de quais espécies estão presentes nos nódulos.

Essas árvores devem expressar similaridade, ancestralidade ou parentescos evolutivos entre espécies, onde as folhas representam os organismos e os nós internos correspondem aos seus ancestrais hipotéticos. Uma das formas de representar a árvore filogenética é através do cladograma (MOURA; VIANA, 2011), que é um diagrama representativo das relações ancestrais entre organismos. Na Figura 2 tem-se uma representação de uma árvore enraizada na forma de cladograma inclinado, onde os valores dentro dos círculos são as espécies, os valores dentro dos quadrados são a quantidade de indivíduos de cada espécie, os nós internos são os ancestrais comuns e as bordas entre cada nó, as distâncias.

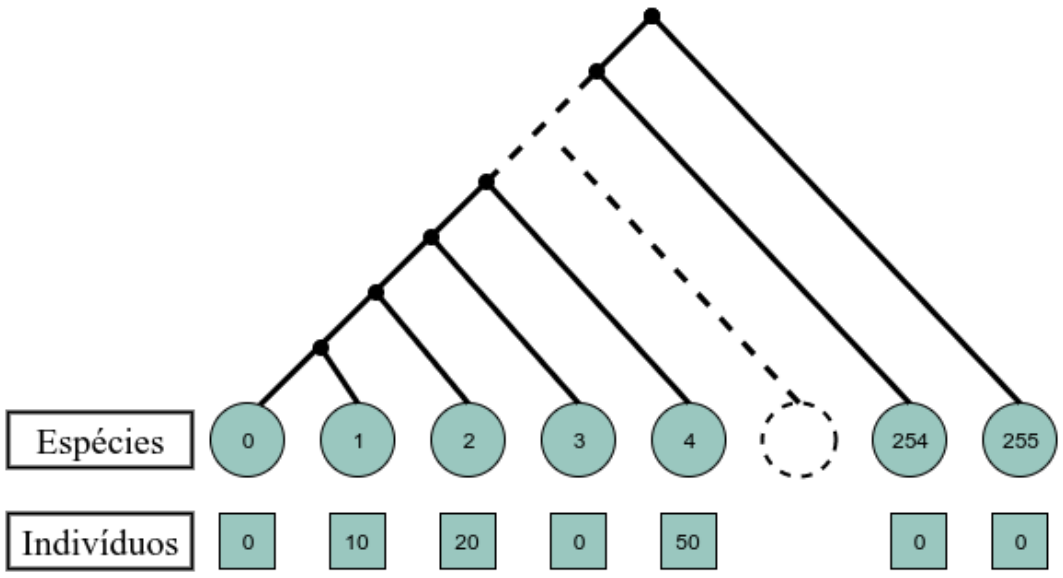



Figura 2: Exemplo de uma árvore enraizada na forma de cladograma inclinado

As árvores filogenéticas combinadas com os índices de diversidade filogenética são empregadas na biologia para comparar amostra de comportamento entre as espécies de diferentes áreas (CIANCIARUSO; SILVA; BATALHA, 2009).

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). A Tabela 1 mostra a correspondência entre a biologia e o nosso trabalho.

A informação requerida de um determinado cladograma para o cálculo da diversidade

Tabela 1: Correspondência entre a biologia e a nossa metodologia.

| **Biologia** | **Metodologia** |
|---|---|
| Comunidade | Região de interesse da imagem da mama |
| Espécies | Níveis de cinza da imagem |
| Indivíduos | Pixels da imagem |
| Distância filogenética | Número de arestas entre duas espécies |

filogenética pode ser resumida por uma matriz das distâncias entre as espécies, retirada do cladograma (FAITH, 1992). Nesta árvore, os nós da folha são as espécies analisadas, os nós internos correspondem a um antepassado comum e as arestas indicam a distância filogenética entre duas espécies. Os valores dentro da matriz de distância podem ser interpretados como a distinção entre cada par de espécies ou entre uma espécie específica e todas as outras (IZSÁK; PAPP, 2000) (RAO, 1982). Em nosso trabalho, as distâncias serão representadas pelo número de arestas entre as espécies. Além disso, a quantidade de indivíduos também é considerada para caracterizar a textura das ROIs.

### 3.3.2 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 premissa principal dessa medida é que a diversidade é maior em uma comunidade em que as espécies são filogeneticamente mais distintas, em outras palavras, pode-se afirmar que isso representaria uma imagem que possui muita variabilidade de níveis de cinza, ou seja, muitas especies. A percepção dominante em ecologia evolutiva é que espécies coexistindo devem diferir significativamente e que a maior parte da variação entre espécies aparentadas é uma resposta adaptativa à competição no passado, quando os traços não diferiam (HARVEY; RAMBAUT, 2000). Esses índices consideram a quantidade de indivíduos, calculada a partir da arquitetura da árvore.

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 3.1, 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 $i$.

$$B \times \frac{\sum_i^B L_i A_i}{\sum_i^B A_i} \tag{3.1}$$

O índice de diversidade filogenética SPD é soma das distâncias filogenéticas entre cada par de espécies, que pode ser observado na Equação 3.2, 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.

$$\left(\frac{S\,(S-1)}{2}\right) \times \frac{\Sigma\Sigma m < n\ d_{mn} a_m a_n}{\Sigma\Sigma m < n\ a_m a_n} \tag{3.2}$$

O índice de diversidade filogenética MNND é a distância média do táxon mais próximo, definido na Equação 3.3. Onde $d_{mn}$ é a distância entre as espécies $m$ e $n$; $a_m$, abundância de espécies $m$.

$$\sum_{m}^{S} mi\,n\,(d_{mn})\,a_m \tag{3.3}$$

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 3.4, onde, o $trC$ representa a soma dos valores da diagonal de uma matriz $C$, $\Sigma c$ é o somatorio de todos os valores da matriz, $n$ é o número de espécies e $'c$ é a média dos elementos da diagonal de $C$.

$$PSV = \frac{n\,trC - \Sigma c}{n\,(n-1)} = 1 - {}'c \tag{3.4}$$

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 3.5.

$$PSR = nPSV \tag{3.5}$$

Foram utilizados os índices de diversidade filogenética para descrever a textura das ROIs. A análise de textura dessas regiões tem como intuito encontrar padrões que serão utilizados para categorizar as massas em maligna e benigna.

## 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; HALL; FRANK, 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 2.

Tabela 2: Matriz de Confusão

| Resultado do Teste | Doença | |
|---|---|---|
| | Presente | Ausente |
| Positivo | Verdadeiro Positivo - VP | Falso Positivo - FP |
| Negativo | Falso Negativo - FN | Verdadeiro Negativo - VN |

A sensibilidade (S), Equação 3.6, 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.

$$S = \frac{VP}{VP + FN} \tag{3.6}$$

Especificidade (E), Equação 3.7, 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.

$$E = \frac{VN}{VN + FP} \tag{3.7}$$

Acurácia (A), Equação 3.8, é a proporção de acertos, ou seja, o total de verdadeiramente positivos e verdadeiramente negativos, em relação a amostra estudada.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.8}$$

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 3.

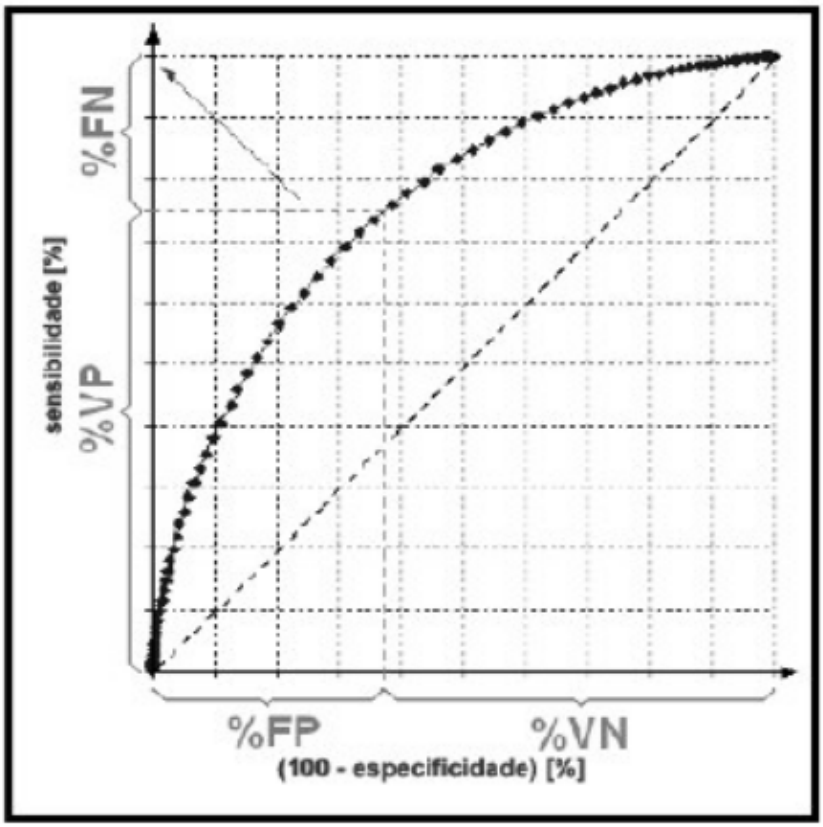



Figura 3: Representação da curva ROC. FONTE: (ROCHA, 2014)

# 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 3.3.2 e a classificação das massas em maligno e benigno, utilizou os classificadores e parâmetros definidos na Subseção 3.4.

Tabela 3: Resultados da classificação

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

De acordo com a Tabela 3, 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, 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 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 3 o classificador *J48* foi capaz de identificar um número baixo de casos com doença, como mostra a sensibilidade de 91,6% e poucos casos sadios foram classificados como doentes, especificidade de 95,5%.

Tabela 4: 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; GAJADHAR; BATISTA, 2014) | DDSM | 85,09 | 77,59 | 89,32 | - |
| (FAHSSI et al., 2015) | MIAS | - | 93,78 | 94,54 | - |
| (ROCHA, 2014) | DDSM | 92,2 | 92,26 | 91,26 | 0,92 |
| **Metodologia** | **DDSM** | **94,8** | **92,9** | **96,5** | **0,988** |

Como pode-se observar na Tabela 4, 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; GAJADHAR; BATISTA, 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, auxiliando os especialistas na tomada de decisão para um diagnóstico mais preciso, 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.

# 6 Publicações

- CARVALHO, E. D.; CARVALHO FILHO, A. O.; SOUSA, A. D.; BARROS, P. V. S.; DRUMOND, P. M. L. L. Diferenciação de padrões de benignidade e malignidade em tecidos da mama baseado na diversidade taxonômica. In: Congresso da sociedade brasileira de computação / Workshop de Informática Médica, 2017, São Paulo. Anais do XXXVII congresso da sociedade brasileira de computação, 2017. v. 37. p. 1911-1920.
- CARVALHO, E. D.; CARVALHO FILHO, A. O. ; SOUSA, A. D. ; BARROS, P. V. S. Classificação de Tecidos da Mama em Maligno e Benigno baseado em Mamografias Digitais usando Descritores de Textura. In: III Escola Regional de Informática do Piauí, 2017, Picos. Livro Anais - Artigos e Minicursos, 2017. v. 1. p. 100-105.
- CARVALHO, E. D.; CARVALHO FILHO, A. O.; SOUSA, A. D.; BARROS, P. V. S.; DRUMOND, P. M. L. L. COMPUTER AIDED DIAGNOSIS PARA TECIDOS DA MAMA, BASEADO EM MAMOGRAFIA DIGITAL USANDO ANALISE DE TEXTURA. In: XIII Simpósio Brasileiro de Automação Inteligente, 2017.

# Referências


BRAZ JÚNIOR, G. Detecção de regiões de massas em mamografias usando ídices de diversidade, geoestatística e geometria côncava. In: . [S.l.]: Tese de Doutorado. Programa de Pós-Graduaçao em Ciência da Computação da Universidade Federal do Maranhão. São Luís - MA, 2014. Citado 2 vezes nas páginas 12 e 15.

BREIMAN, L. "Random forests". In: . [S.l.]: Machine Learning, v. 45, n. 1, p. 5-32, 2001. Citado na página 19.

CIANCIARUSO, M. V.; SILVA, I. A.; BATALHA, M. A. Diversidades filogenética e funcional: novas abordagens para a ecologia de comunidades. *Biota Neotropica*, SciELO Brasil, v. 9, n. 3, p. 93–103, 2009. Citado na página 17.

FAHSSI, K. E. et al. Benign or malignant lesion classification in mammography images using the adaptive orthogonal transformation and the coefficients of the correlation matrix. In: . [S.l.]: Recent Advances in Electrical Engineering, ISBN: 978-1-61804-351-1, 2015. Citado 3 vezes nas páginas 13, 22 e 23.

FAITH, D. P. Conservation evaluation and phylogenetic diversity. *Biological conservation*, Elsevier, v. 61, n. 1, p. 1–10, 1992. Citado na página 18.

HARVEY, P. H.; RAMBAUT, A. Comparative analyses for adaptive radiations. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, The Royal Society, v. 355, n. 1403, p. 1599–1605, 2000. Citado na página 18.

HEATH, M.; BOWYER, K.; KOPANS, D. The digital database for screening mammography. In: . [S.l.]: Citeseer, Proceedings of the 5th international workshop on digital mammography. p. 212-218, 2000. Citado na página 16.

INCA. Tipos de câncer: mama. In: . [S.l.]: Disponível em: http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama/cancer_mama. Acesso em: 11/11/2016, 2016. Citado na página 12.

IZSÁK, J.; PAPP, L. A link between ecological diversity indices and measures of biodiversity. *Ecological Modelling*, Elsevier, v. 130, n. 1, p. 151–156, 2000. Citado na página 18.

LANDWEHR, N.; HALL, M.; FRANK, E. Logistic model trees. In: . [S.l.]: Machine Learning, 95(1-2):161-205, 2005. Citado na página 19.

MAGURRAN, A. E. Measuring biological diversity. In: . [S.l.]: African Journal of Aquatic Science; v. 29, n. 2, p. 285-286, 2004. Citado na página 18.

MELO, M. C.; GAJADHAR, A.; BATISTA, L. V. Análise comparativa de métodos de aprendizagem de máquina para classificação de massas em mamografias. In: . [S.l.]: 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, 2014. Citado 3 vezes nas páginas 13, 22 e 23.

MOURA, H. A. S.; VIANA, G. V. R. Anacê: Phylogenetic trees drawing web service. Citeseer, 2011. Citado na página 17.

OLIVEIRA, F. S. S. Classificação de tecidos da mama em massa e não-massa usando Índice de diversidade taxonômico e máquina de vetores de suporte. In: . [S.l.]: Dissertação de Mestrado. Curso de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão. São Luís - MA, 2013. Citado na página 17.

PÉREZ, J. M. et al. Combining multiple class distribution modified subsamples in a single tree. In: . [S.l.]: Pattern Recognition Letters, 28(4): 414-422, 2007. Citado na página 19.

QUINLAN, R. C4.5: programs for machine learning. In: . [S.l.]: Morgan Kaufmann Publishers, San Mateo, CA, 1993. Citado na página 19.

RAO, C. R. Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, Elsevier, v. 21, n. 1, p. 24–43, 1982. Citado na página 18.

ROCHA, S. V. Diferenciação do padrão de malignidade e benignidade de massas em imagens de mamografias usando padrões locais binários, geoestatística e Índice de diversidade. In: . [S.l.]: Tese de Doutorado. Curso de Pós-Graduação em Engenharia de Ele- tricidade da Universidade Federal do Maranhão. São Luís - MA, 2014. Citado 6 vezes nas páginas 8, 14, 16, 21, 22 e 23.

SILVA, T. F. Diferenciação do padrão de malignidade e benignidade de massas em mamografias utilizando caracteríticas geométricas e máquina de vetor de suporte. In: . [S.l.]: Dissertação de Mestrado. Programa de Pós-Graduaçao em Ciência da Computação da Universidade Federal do Maranhão. São Luís - MA, 2016. Citado 7 vezes nas páginas 12, 13, 15, 16, 21, 22 e 23.

WEBB, C. O. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. In: . [S.l.]: The American Naturalist 156.2, p. 145-155, 2000. Citado na página 18.

WEKA. Machine learning group at the university of waikato. In: . [S.l.]: Disponível em: http://www.cs.waikato.ac.nz/ml/weka/. Acesso em: 15 de Março de 2017, 2017. Citado na página 19.

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

**Identificação do Tipo de Documento**

( ) Tese

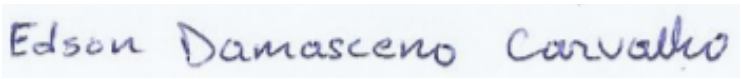
( ) Dissertação

( x ) Monografia

( ) Artigo

Eu, **Edson Damasceno Carvalho**, 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 **Diferenciação de padrões de benignidade e malignidade em tecidos da mama baseado na diversidade taxonômica** 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 10 de Janeiro de 2018.



Assinatura