



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PIAUÍ
CAMPUS SENADOR HELVÍDIO NUNES DE BARROS
COORDENAÇÃO DO CURSO SISTEMAS DE INFORMAÇÃO
Rua Cícero Eduardo S/N – Bairro Junco – 64.600-000 – Picos –PI.
Fone (89) 3422-1087 – Fax (89) 3422-1043



**REDUÇÃO DE FALSOS POSITIVOS EM IMAGENS DE MAMMOGRAFIAS
DIGITAIS USANDO OS ÍNDICES *PHYLOGENETIC SPECIES VARIABILITY*,
PHYLOGENETIC SPECIES RICHNESS E MÚLTIPLOS CLASSIFICADORES**

**FALSE POSITIVE REDUCTION IN IMAGES OF DIGITAL MAMMOGRAMS
USING PHYLOGENETIC SPECIES VARIABILITY INDEXES, PHYLOGENETIC
SPECIES RICHNESS AND MULTIPLE CLASSIFIERS**

Autores: Laércio N. Mesquita¹, Antonio O. de C. Filho², Alcilene D. de Sousa³, Patrícia M. L.
de L. Drumond³

PICOS - PI

2016

¹ Graduando em Sistemas de Informação, UFPI;

² Professor do curso de Sistemas de Informação, UFPI - Orientador;

³ Professor (a) do curso de Sistemas de Informação, UFPI.

LAÉRCIO NASCIMENTO MESQUITA
ANTONIO OSEAS DE CARVALHO FILHO
ALCILENE DALÍLIA DE SOUSA
PATRÍCIA MEDYNA LAURITZEN DE LUCENA DRUMOND

**REDUÇÃO DE FALSOS POSITIVOS EM IMAGENS DE MAMOGRAFIAS
DIGITAIS USANDO OS ÍNDICES *PHYLOGENETIC SPECIES VARIABILITY*,
PHYLOGENETIC SPECIES RICHNESS E MÚLTIPLOS CLASSIFICADORES**

Trabalho de Conclusão de Curso submetido à
Coordenação do Curso de Bacharelado em
Sistemas de Informação da Universidade
Federal do Piauí, Campus Senador Helvídio
Nunes de Barros, no período 2016.1 como
requisito para obtenção do título de Bacharel
em Sistemas de Informação.

Orientador: Prof. Me. Antonio Oseas de
Carvalho Filho

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

R321

Redução de falsos positivos em imagens de mamografias digitais usando os índices *Phylogenetic Species Variability*, *Phylogenetic Species Richness* e múltiplos classificadores / Laércio Nascimento Mesquita [et al.]. – 2016.

CD-ROM : il.; 4 ¾ pol. (15 f.)

Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) – Universidade Federal do Piauí, Picos, 2016.

Orientador(A): Prof. Me. Antonio Oseas de Carvalho Filho

1. Índices de Diversidade Filogenética. 2. Imagens Digitais-Redução de Falsos Positivos. 3. Câncer de Mama. I. Sousa, Alcilene Dalília de. II. Drumond, Patrícia Medyna Lauritzen de Lucena. III. Título.

CDD 005

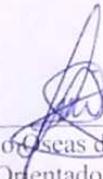
REDUÇÃO DE FALSOS POSITIVOS EM IMAGENS DE MAMOGRAFIAS
DIGITAIS USANDO OS ÍNDICES *PHYLOGENETIC SPECIES VARIABILITY*,
PHYLOGENETIC SPECIES RICHNESS E MÚLTIPLOS CLASSIFICADORES

LAÉRCIO NASCIMENTO MESQUITA

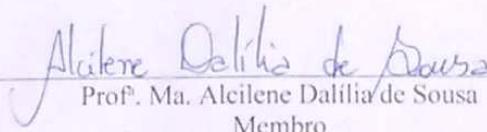
Artigo apresentado como exigência parcial para obtenção do grau de Bacharel em
Sistemas de Informação.

Data de Aprovação

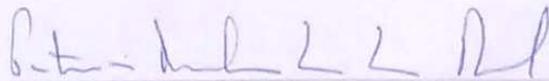
Picos – PI, 14 de julho de 2016.



Prof. Me. Antonio Osceas de Carvalho Filho
Orientador



Prof.ª. Ma. Alcilene Dalízia de Sousa
Membro



Prof.ª. Ma. Patrícia Medyna Lauritzen de Lucena Drumond
Membro

RESUMO: Os sistemas *Computer-Aided Detection* (CAD) têm ganhado cada vez mais atenção devido sua importância na área médica, tornando o diagnóstico por parte dos especialistas mais preciso. Para criação dessas ferramentas são utilizados métodos computacionais, processamento digital de imagens e conhecimentos sobre a doença. Neste trabalho utilizam-se índices de diversidade filogenética para extração de características baseada na textura. Tais índices são utilizados como características para os classificadores: *Support Vector Machine*, *Random Forest*, *Random Basis Function* e *MultiLayer Perceptron* identificando em tecidos de mamografias a presença de massa e não massa, perfazendo assim parte de um CAD. Para validação da metodologia foram utilizadas 200 imagens de mamografia, onde 100 contêm massas e as demais não massas. Os resultados mostram-se promissores, pois os melhores resultados alcançam uma acurácia de 91,5%, sensibilidade de 89,5%, especificidade de 94% e uma taxa de falsos positivos de 0,085 por exame.

PALAVRAS-CHAVE: Redução de falsos positivos. Câncer de mama. Índices de Diversidade Filogenética.

ABSTRACT: The Computer-Aided Detection (CAD) systems has gained more attention because of his importance in the medicinal area, making the diagnosis by the experts more accurate. For the creation of these tools are used computational methods, digital image processing and knowledge about the disease. In this work we are using indices of phylogenetic diversity for extraction of features based on texture. Such indexes are used as features for the classifiers: Support Vector Machine, Random Forest, Random Basis Function and MultiLayer Perceptron, identifying in mammograms tissues the presence of mass and not mass, making thus part of a CAD. For the validation of the methodology, was used 200 mammography images, where 100 contain masses and the others not masses. The results showed promising, because the best results reach an accuracy of 91.5%, sensitivity of 89.5%, specificity of 94% and a false positives rate of 0.085 per examination.

KEYBOARDS: False Positive Reduction. Breast Cancer. Phylogenetic Diversity Indexes.

1 INTRODUÇÃO

Segundo o INCA (2016), o câncer de mama é um tumor maligno formado por um grupo de células cancerosas que crescem rapidamente em tecidos e órgãos, atingindo-os ao se espalharem desordenadamente por todo o corpo. O câncer de mama é mais comum entre as mulheres, depois do câncer de pele não melanoma, respondendo por cerca de 25% dos novos casos a cada ano. O câncer de mama também acomete homens, porém é raro, representando apenas 1% do total de casos da doença. Para o ano de 2016 a estimativa é de 57.960 novos casos de câncer de mama, com risco estimado em 56,20 casos a cada 100 mil mulheres.

Este tipo de câncer é relativamente raro antes dos 35 anos, acima desta idade sua incidência cresce progressivamente, especialmente após os 50 anos. Estatísticas indicam um aumento da sua incidência tanto nos países desenvolvidos quanto nos em desenvolvimento. O método mais indicado é que a mulher realize o autoexame de mama a fim de encontrar possíveis nódulos. A detecção precoce do câncer de mama também pode ser feita pela mamografia, quando realizada em mulheres sem sinais e sintomas da doença, numa faixa etária em que haja um balanço favorável entre benefícios e riscos. A recomendação no Brasil, atualizada em 2015, é que mulheres entre 50 e 69 anos façam uma mamografia a cada dois anos. Os benefícios da mamografia de rastreamento incluem a possibilidade de encontrar o câncer em fase inicial e ter um tratamento menos agressivo, como também uma menor chance

de morrer pela doença (INCA, 2016).

Para que esse rastreamento ocorra de forma mais eficiente, os profissionais da área da saúde utilizam de ferramentas que darão uma segunda opinião quando usadas em exames de mamografias. Essas ferramentas são conhecidas como *Computer-Aided Detection (CAD)*, cujos sistemas são responsáveis por diminuir um diagnóstico errôneo, aumentando assim, as chances de detecção do câncer de mama em estágios iniciais.

Esse trabalho tem como objetivo descrever uma nova abordagem para extração de características baseado em índices de diversidade filogenética, que serão responsáveis por caracterizar as regiões de interesse em um sistema CAD. Outra contribuição desse trabalho é a redução da taxa de falsos positivos nesse tipo de sistema. Serão usados descritores de textura baseados em índices filogenéticos, em seguida, será feita uma classificação através dos classificadores: *MultiLayer Perceptron* (HAYKIN *et al.*, 2009), *Radial Basis Function* (HAYKIN *et al.*, 2009), *Random Forest* (BREIMAN, 2001) e *Support Vector Machine* (VAPNIK, 1998) com intuito de diferenciar massas e não massas em exames de mamografias e, por fim, reduzir a quantidade de falsos positivos para o pré-diagnóstico do câncer de mama.

2 TRABALHOS RELACIONADOS

Na literatura especializada, existem diversos trabalhos relacionados à diminuição de falsos positivos em exames de mamografia e classificação de massas e não massas. Para este propósito, utilizam-se características extraídas de imagens médicas, as quais servem como vetores de entrada para os classificadores.

Sampaio (2015) apresenta uma metodologia computacional que ajuda o especialista na descoberta de massas mamárias, baseando-se na densidade da mama. Para a segmentação foi utilizado um Micro Algoritmo Genético (μ AG) com o objetivo de criar uma máscara de proximidade de textura e selecionar regiões suspeitas em conter lesão. Para reduzir o número de regiões suspeitas erroneamente segmentadas, utilizaram-se duas etapas de redução de falsos positivos. A primeira redução de falsos positivos usa o *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) e um *ranking* de proximidade de textura extraídos das regiões de interesse. Na segunda, as regiões resultantes têm as suas texturas e formas analisadas pela combinação de Árvores Filogenéticas e descritores geométricos, Padrões Binários Locais e SVM. Um μ AG foi utilizado para escolher as regiões suspeitas que gerem os melhores modelos de treinamento e maximizem a classificação de massas e não

massas usadas na SVM. Os melhores resultados obtidos produziram uma sensibilidade de 94,02%, especificidade de 82,28% e acurácia de 84,08%, com uma taxa de 0,85 falsos positivos por imagem e uma área sob a *Free-Response ROC Curve* (curva FROC) de 1,13 nas análises de mamas não densas. Para mamas densas, obteve-se uma sensibilidade de 89,13%, especificidade de 88,61% e acurácia de 88,69%, com uma taxa de 0,71 falsos positivos por imagem e uma Área sob a Curva FROC de 1,47.

Oliveira (2013) propõe uma metodologia de discriminação e classificação de regiões extraídas da mama em massa e não massa. O banco de imagens *Digital Database for Screening Mammography* (DDSM) é usado neste trabalho para a aquisição das mamografias, onde são extraídas as regiões de massa e não massa. Na descrição da textura da região de interesse são usados os Índices de Diversidade Taxonômica (Δ) e Distinção Taxonômica (Δ^*), provenientes da ecologia. O cálculo destes índices é baseado nas árvores filogenéticas, sendo aplicados neste trabalho na descrição de padrões em regiões das imagens da mama com duas abordagens de regiões delimitadoras para análise da textura: círculo com anéis e máscaras internas com externas. Para classificação das regiões em massa e não massa, utilizou-se o classificador SVM. A metodologia apresenta resultados promissores para a classificação de massas e não massas, alcançando uma acurácia média de 99,67%.

Carvalho (2012) também propõe uma metodologia de discriminação e classificação de regiões extraídas de mamografias em massa e não massa. Neste estudo, a base usada é DDSM. Para descrever a textura da região de interesse é aplicado o Índice de Diversidade de *McIntosh*, comumente usado em ecologia. O cálculo deste índice é proposto em quatro abordagens: através do Histograma, da Matriz de Co-ocorrência de Níveis de Cinza, da Matriz de Comprimentos de Corrida de Cinza e da Matriz de Comprimentos de Lacuna de Cinza. Para classificação das regiões em massa e não massa foi utilizado o classificador supervisionado SVM. A metodologia apresenta resultados eficazes para a classificação de massas e não massas, alcançando uma acurácia de 93,68%.

Sousa (2011) apresenta uma metodologia de discriminação e classificação de regiões de tecidos de mamografias em massa e não massa. Para este propósito utiliza-se o Índice de Diversidade de *Shannon-Wiener*, comumente aplicado para medir a biodiversidade em um ecossistema, o qual descreve padrões de regiões de imagens de mama com quatro abordagens: global, em círculos, em anéis e direcional. Em seguida, utiliza-se o classificador SVM para classificar estas regiões em massa e não massa. A metodologia apresenta resultados

promissores para a classificação de regiões de tecidos de mamografia em massa e não massa, obtendo uma acurácia máxima de 99,85%.

Esse trabalho tem como objetivo a redução de falsos positivos em exames de mamografias usando índices de diversidade filogenéticos *Phylogenetic Species Variability* (PSV) e *Phylogenetic Species Richness* (PSR) e múltiplos classificadores. O presente estudo apresenta uma abordagem inovadora, em que são utilizadas características que se baseiam em comportamentos presentes em comunidades, como o parentesco entre espécies e sua riqueza. Essa metodologia foi aplicada em imagens de mamografia para identificação de massa e não massa de forma automática.

3 METODOLOGIA PROPOSTA

Para que fosse possível identificar a presença de massa e não massa em exames de mamografia foi empregada a seguinte metodologia: utilização da base de imagens pública DDSM; para a extração de características foi utilizado o descritor baseado na textura utilizando os índices de diversidade filogenética *Phylogenetic Species Variability* e *Phylogenetic Species Richness* em (HELMUS *et al.*, 2007); e para a classificação utilizamos a Máquina de Vetores de Suporte em (VAPNIK, 1998), *Random Forest* (BREIMAN, 2001), RBF (HAYKIN *et al.*, 2008), MLP (HAYKIN *et al.*, 2008), como está demonstrado na Figura 1.

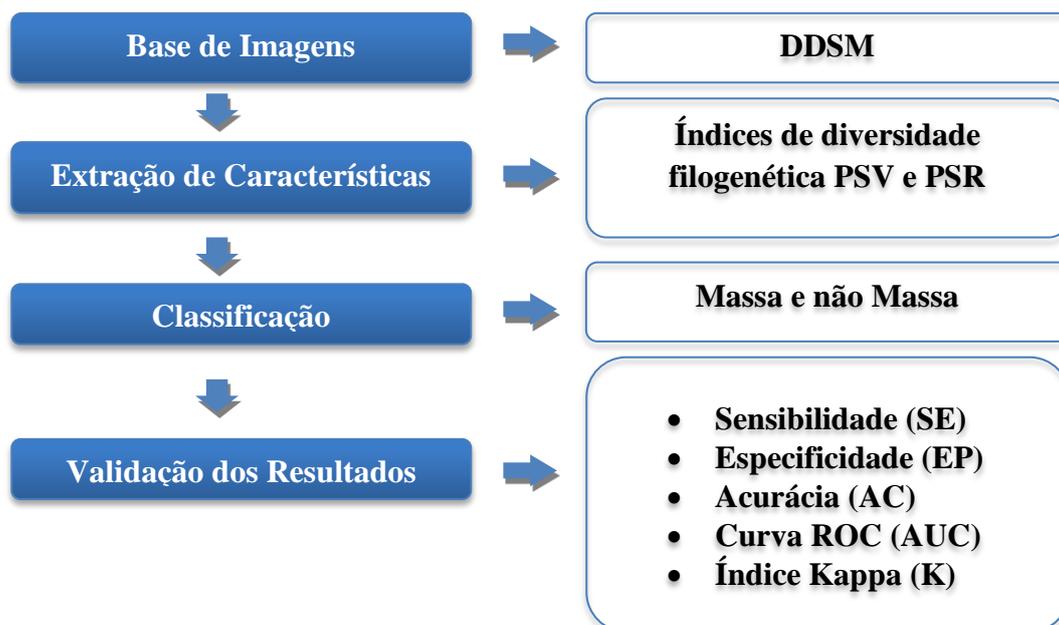


Figura 1. Metodologia proposta para redução de falsos positivos.

3.1 Base de Imagens

A mamografia é a radiografia da mama que permite a detecção precoce do câncer, sendo capaz de mostrar lesões em fase inicial, muito pequenas (de milímetros). Este exame é realizado em um aparelho de raio-X apropriado chamado mamógrafo. Nele a mama é comprimida de forma a fornecer melhores imagens e, portanto, melhor capacidade de diagnóstico. Estudos sobre a efetividade da mamografia sempre utilizam o exame clínico como exame adicional o que torna difícil distinguir a sensibilidade do método como estratégia isolada de rastreamento. A sensibilidade varia de 46% a 88% e depende de fatores tais como: tamanho e localização da lesão, densidade do tecido mamário (mulheres mais jovens apresentam mamas mais densas), qualidade dos recursos técnicos e habilidade de interpretação do radiologista. A especificidade varia entre 82% e 99% e depende igualmente da qualidade do exame (INCA, 2016).

Normalmente são feitas duas radiografias, sendo uma para cada mama, em duas projeções: Médio Lateral Oblíqua (MLO) e uma Crânio-Caudal (CC). Na Figura 2, podemos ver dois exemplos de mamografia.

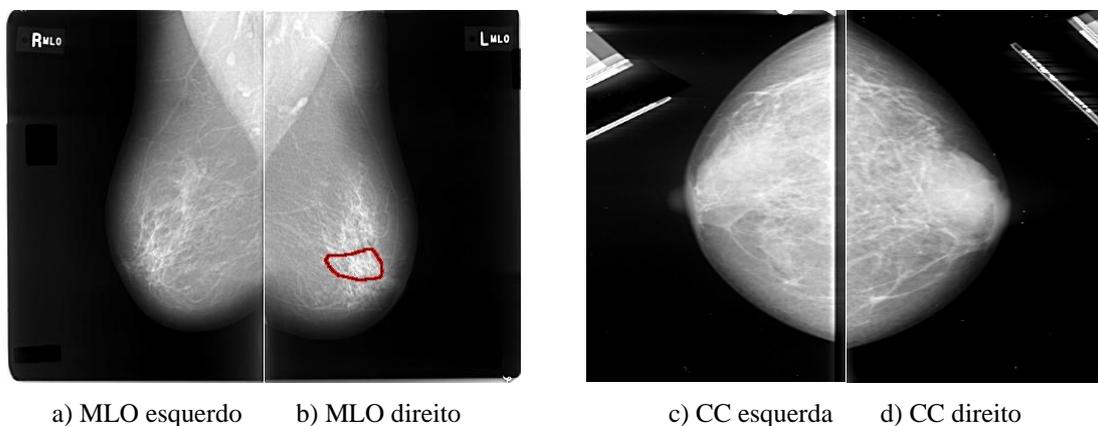


Figura 2. Exemplos de Mamografia.

Fonte: Heath *et al.* (1996).

Pode ser observado na Figura 2. a) e b) Mamografias com incidência Médio-Lateral Oblíquo em ambas as mamas, sendo que em (b) mostra delimitado em vermelho uma massa maligna (HEATH *et al.*, 1996); e temos em (c) e (d) Mamografias com incidência Crânio-Caudal em ambas as mamas e que não apresentam região suspeita de anormalidade (HEATH *et al.*, 1996).

A Digital Database for Screening Mammography é uma base pública de imagens de

mamografias, a qual foi uma doação do *Breast Cancer Research Program of the U.S. Army Medical Research and Materiel Comman*. O objetivo desse banco é facilitar a pesquisa e desenvolvimento de algoritmos para ajudar no diagnóstico, além de ser uma ferramenta de ensino para formação profissional. 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.*,1996). Para a concretização desse trabalho, foram utilizadas 200 imagens de mamografias; sendo 100 imagens com a presença de massa e 100 de não massa.

3.2 Índices de Diversidade Filogética

Segundo Oliveira (2003), a diversidade é um termo muito utilizado na área da ecologia. O seu objetivo é informar a variedade de espécies presentes em uma comunidade ou área. Para Webb *et al.* (2000) a filogenia pode ser utilizada para investigar de forma contemporânea os processos ecológicos e a composição das espécies nas comunidades. Tendo isso em mente, foram utilizados os índices PSV e PSR com intuito de encontrar padrões em regiões de imagens de mamografias. Nesse sentido os índices foram utilizados para descrever a textura das regiões de interesse do inglês *Region of Interest* (ROI) de massa e não massa extraída de exames de mamografias. Para a obtenção dos índices foram usadas apenas as informações das espécies (níveis de cinza) presentes na ROI analisada e a quantidade de indivíduos de cada espécie.

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 e os indivíduos sendo os *pixels* (OLIVEIRA, 2013). Por isso é preciso fazer uma correlação entre a biologia e a metodologia proposta, como na Tabela 1.

Tabela 1. Correspondência entre os termos da biologia e o trabalho

Biologia	Metodologia
Comunidade	ROI
Espécies	Níveis de cinza da ROI
Indivíduos	Quantidade de <i>pixels</i> da ROI

Na árvore em estrela mostrada na Figura 3 podem ser vistas as métricas filogenéticas das espécies, que são: PSV (variabilidade), PSR (riqueza) e *Phylogenetic Species Evenness* (PSE) que representa a equitatividade⁴. Quando a árvore tem o formato em estrela as métricas PSV e PSR são máxima, ou seja, igual 1, sabendo-se que esses valores variam entre 0 e 1. Pode-se notar que o lado esquerdo de A os valores de PSV e PSR são maiores que o lado direito de A, já no lado direito B os valores PSV e PSR são maiores que os do lado esquerdo de B. Segundo Helmus *et al.* (2007) toda vez que uma espécie é adicionada na comunidade, os valores de PSV e PSR dependem do comprimento do ramo onde a espécie é adicionada. Nota-se que houve um aumento do lado direito B em relação ao lado esquerdo B. Já a métrica PSE quantifica enxertos individuais de uma determinada comunidade na ponta de cada ramo com comprimentos 0. Verifica-se que o lado esquerdo C apresenta as métricas PSE e PSV iguais, porque a abundância de todas as espécies são iguais, quando a abundância difere o valor de PSE é menor que o de PSV. Já no caso esquerdo de A quando as espécies são filogeneticamente iguais o valor de PSE = 1.

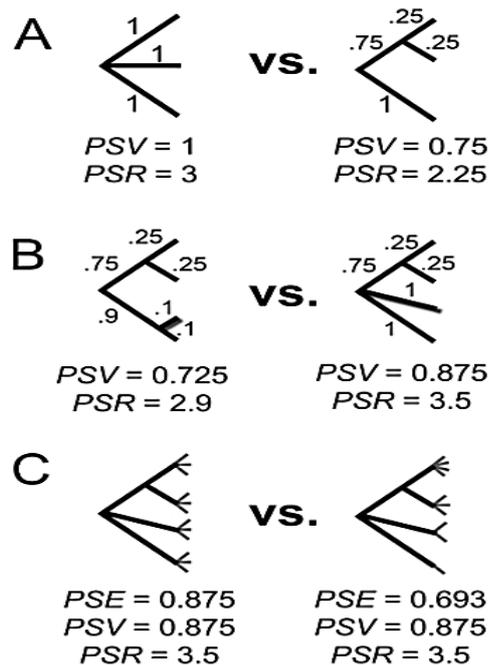


Figura 3. Árvore em forma de estrela.

Fonte: Helmus *et al.* (2007, p.2).

⁴ Em ecologia equitatividade é o termo empregado para definir a uniformidade, ou homogeneidade, da distribuição de abundância de espécies em uma comunidade.

3.3 Phylogenetic Species Variability

Segundo Helmus *et al.* (2007) a variabilidade quantifica o parentesco filogenético, diminuindo a variação das características compartilhadas por todas as espécies da comunidade. A variabilidade de espécies filogenéticas resume o grau em que as espécies de uma comunidade são filogeneticamente relacionadas. Na Equação 1 PSV, o trC representa a soma dos valores da diagonal de uma matriz C , $\sum C$ é o somatório de todos os valores da matriz fora da diagonal de C , n é o número de espécies e \bar{c} é a média dos elementos da diagonal de C .

$$PSV = \frac{ntrC - \sum C}{n(n - 1)} = 1 - \bar{c} \quad (1)$$

3.4 Phylogenetic Species Richness

A riqueza de espécies quantifica o número de espécies em uma comunidade. Podemos encontrar o valor do PSR através da Equação 2, onde multiplicamos o número espécies n pela variabilidade da comunidade (Helmus *et al.*, 2007).

$$PSR = nPSV \quad (2)$$

3.5 Classificação e Validação

A classificação foi realizada pelo *software Waikato Environment for Knowledge Analysis (WEKA)*, que é um conjunto de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Contém ferramentas para pré-processamento de dados, classificação, regressão, *clustering*, regras de associação e visualização. Ele foi desenvolvido por um grupo de pesquisadores da Universidade de Waikato, Nova Zelândia (HALL *et al.*, 2009).

A fim de considerarmos a presença ou ausência de massa em imagens de mamografia, para validação dos resultados, utilizou-se de métricas de avaliação baseadas em estatísticas como, a Área sob a Curva ROC, que mensura o quanto o algoritmo é eficiente, usualmente referida como AUC do inglês *Area Under the Curve* é referida nas tabelas 2, 3,4 como (AUC) (Metz, 1986), Sensibilidade (SE) (Martinez; Louzada-Neto; Pereira, 2003), Especificidade (EP) (MARTINEZ; LOUZADA-NETO; PEREIRA, 2003), Acurácia (AC) (METZ, 1986) e índice Kappa (K) (LANDIS; KOCH, 1977). Segundo Metz (1986) a sensibilidade (SE) é definida como a probabilidade do teste sob investigação fornecer um resultado positivo, dado

que o indivíduo é realmente portador da enfermidade. Ainda em Martinez; Louzada-Neto; Pereira (2003) a especificidade (EP) é definida como a probabilidade do teste fornecer um resultado negativo, dado que o indivíduo está livre da enfermidade. Em Metz (1986) acurácia é a métrica que calcula o total de acertos em relação a todas as instâncias classificadas corretamente. Essa medida mensura o quão bem um classificador reconhece instâncias de diversas classes. Foram obtidos também os Falsos Positivos por imagem (FP/i), para uma análise mais detalhada. O índice Kappa (K) de acordo com Landis; Koch (1977) é uma medida de concordância que pode ser formulada para medir o desacordo de um conjunto de respostas, baseada em pesos, medindo a concordância entre um número de resposta baseado em observadores, chegando a um consenso.

Os classificadores testados utilizaram o método *k-fold cross validation* para a obtenção dos resultados. Os dados foram divididos em 10 conjuntos, sendo 9 deles para treinamento e 1 para testes. Este processo é repetido 10 vezes, de forma que o conjunto escolhido para o teste será diferente do anterior e no final é gerada uma média dos resultados.

4 RESULTADOS

Para os testes realizados neste trabalho, utilizou-se a base DDSM. As características foram extraídas a partir dos índices de diversidade filogenéticos de variabilidade e riqueza de espécies. Os classificadores *MultiLayer Perceptron* (MLP), *Radial Basis Function* (RBF) *Random Forest* (RF) e *Support Vector Machine* (SVM) foram usados para classificação de massa e não massa em imagens de mamografias, utilizou-se vários parâmetros tendo em vista melhores resultados, os melhores parâmetros para o MLP foi utilizando 6 camadas ocultas, para o RBF foi usando 11 *clusters* e para o RF foi utilizando 155 árvores e para o SVM foi utilizando o *kernel* linear e como parâmetros os valores padrões, através do *software* WEKA. A classificação foi realizada com validação cruzada de *k-folds*, sendo $k = 10$.

Tabela 2. Resultados da classificação usando PSV

Classificador	SE	EP	FP/i	AUC	AC	K
SVM	61,6%	77%	0,355	0,645	64,5%	0,29
RF	56,2%	59%	0,435	0,632	56,5%	0,13
RBF	62,2%	74%	0,355	0,621	64,5%	0,29
MLP	63,7%	65%	0,360	0,661	64%	0,28

Como pode ser visto na Tabela 2 o melhor resultado foi obtido através do classificador SVM, com uma acurácia de 64,5%, sensibilidade de 77% e especificidade de 61,6%, a Área

sob a Curva ROC foi de 0,645 é um índice Kappa de 0,29, sendo considerado razoável de acordo com (LANDIS; KOCH, 1977). O resultado menos significativo foi obtido através do classificador RF com uma acurácia de 56,5%, sensibilidade de 56,2%, especificidade de 59% e um índice Kappa de 0,13.

Tabela 3. Resultados da classificação usando PSR

Classificador	SE	EP	FP/i	AUC	AC	K
SVM	75,9%	82%	0,220	0,780	78%	0,56
RF	72,3%	68%	0,290	0,820	71%	0,42
RBF	73,2%	90%	0,215	0,836	78,5%	0,57
MLP	75%	84%	0,220	0,845	78%	0,56

Pode-se observar na Tabela 3 que, para o índice PSR, o melhor resultado obtido foi com o classificador RBF, o qual atingiu uma acurácia de 78,5%, sensibilidade de 73,2% e especificidade de 90%, a Área sob a Curva ROC foi de 0,836 e um índice Kappa de 0,57, sendo considerado bom de acordo com (LANDIS; KOCH, 1977). O resultado menos relevante foi obtido pelo classificador RF com uma acurácia de 71%, sensibilidade de 72,3%, especificidade de 68% e um índice Kappa de 0,42.

Tabela 4. Resultados da classificação usando PSV e PSR

Classificador	SE	EP	FP/i	AUC	AC	K
SVM	82,9%	92%	0,135	0,865	86,5%	0,73
RF	88,3%	91%	0,105	0,934	89,5%	0,79
RBF	88,5%	92%	0,100	0,957	90%	0,80
MLP	89,5%	94%	0,085	0,950	91,5%	0,83

Como descrito na Tabela 4 a junção dos índices PSV e PSR conseguiu atingir resultados muito promissores, no melhor resultado obteve-se 91,5% de acurácia, sensibilidade de 89,5% e especificidade de 94%. A taxa de falsos positivos foi de 0,085 por imagem, com uma Área sob a Curva ROC de 0,952 e índice Kappa de 0,83, sendo considerado excelente de acordo com a tabela do índice Kappa (LANDIS; KOCH, 1977). O resultado menos significativo para análise foi conseguido através do SVM, como uma acurácia de 86,5% e um índice Kappa de 0,73, o resultados mostraram-se com uma maior taxa de acerto sobre o conjunto de características de indivíduos sem a presença da neoplasia.

Tabela 5. Comparação da metodologia com os trabalhos relacionados

Trabalhos	Base	SE	EP	FP/i	AC
Sampaio (2015)	DDSM	94,02%	82,28	0,85	84,08%
Oliveira (2013)	DDSM	89,5%	-	-	99,67%
Carvalho (2012)	DDSM	90,10%	-	-	93,68%
Sousa (2011)	DDSM	91,50%	-	-	99,85%
Metodologia proposta	DDSM	89,5%	94%	0,085	91,5%

Como podemos observar na Tabela 5, a metodologia proposta utilizando os índices PSV e PSR apresentou uma pequena melhoria em relação ao trabalho de Sampaio (2015) com uma acurácia de 91,5%, especificidade de 94% e uma taxa de falsos positivos similar. Sabendo-se que os resultados de falsos positivos retornados se assemelham aos de Sampaio (2015), a utilização dos índices PSV e PSR teria uma maior eficácia na sua metodologia, reduzindo ainda mais essa taxa. Nas tabelas 2 observa-se que o classificador SVM apresenta os melhores resultados, já nas tabelas 3 e 4 o classificador SVM não atingiu os resultados esperados, tendo em vista que todos os trabalhos relacionados atingiram resultados muito expressivos. Na tabela 4 era esperado que o SVM fosse o mais eficiente dentre os classificadores utilizados usando a combinação dos dois índices de diversidade. Os resultados mostraram-se promissores em relação aos trabalhos recentes na literatura relacionados à classificação de tecidos de mamografias nas classes de massa e não massa.

5 CONCLUSÃO

A partir dos resultados obtidos, pode-se inferir que a utilização de descritores baseados na textura apresentam resultados eficazes. O uso de índices filogenéticos para descrever padrões em regiões de imagens mostrou-se eficiente na identificação de massas em exames de mamografias.

6 TRABALHOS FUTUROS

Para os trabalhos futuros, pretende-se utilizar novos índices filogenéticos, testar uma abordagem com novos classificadores, como forma de se obter uma melhor redução da taxa de falsos positivos, como também garantir uma maior eficiência nesse método.

7 PUBLICAÇÕES

Trabalho Original

MESQUITA, L. N.; CARVALHO FILHO, A. O. ; SOUSA, A. D. ; DRUMOND, P. M. L. L.

Redução de falsos positivos em imagens de mamografias digitais usando os índices phylogenetic species variability, phylogenetic species richness e múltiplos classificadores. In: Workshop de Informática Médica, 2016, Porto Alegre. Anais do XXXVI congresso da sociedade brasileira de computação: #Computaçãoeinterdisciplinaridade, 2016.

Disponível em: <<http://ebooks.pucrs.br/edipucrs/anais/csbc/assets/2016/anais-csbc-2016.pdf>> Acesso em 10 de jul. de 2016.

Outras Publicações

MESQUITA, L. N.; CARVALHO FILHO, A. O. ; SOUSA, A. D. **Redução de falsos positivos em imagens de mamografias digitais, usando os índices phylogenetic species variability e phylogenetic species richness e classificador SVM.** In: Escola Regional de Informática do Piauí, 2016, Teresina. Anais da I Escola Regional de Informática do Piauí, 2016.

MESQUITA, L. N.; SOUSA, A. D. ; MOURA, M. J. C. ; OLIVEIRA, I. A. **Uma Análise Comparativa entre Classificadores para Detecção de Nódulos Pulmonares.** In: Escola Regional de Informática do Piauí, 2015, Parnaíba. Anais da I Escola Regional de Informática do Piauí, 2015.

MOURA, M. J. C.; SOUSA, A. D. ; MESQUITA, L. N. ; OLIVEIRA, I. A. **Segmentação de Candidatos a Nódulos Pulmonares em Imagens de Tomografia Computadorizada.** In: Workshop de Informática Médica, 2015, Recife. Anais do Workshop de Informática Médica, 2015.

REFERÊNCIAS

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

CARVALHO, P. M. S. **Classificação de Tecidos da Mama a partir de imagens mamográficas em massa e não massa usando índices de diversidade de McIntosh e Máquina de Vetores de Suporte.** 2012.75f. Dissertação (Mestrado na área de Ciência da Computação) - Programa de Pós-Graduação em Engenharia de Eletricidade, Universidade Federal do Maranhão, São Luís, 2012.

HAYKIN, S. **Neural Networks and Learning Machines.** 3ª ed. New Jersey: Prentice Hall, 936 p. 2008.

HEATH, M.; BOWYER, K.; KOPANS, D.; KEGELMEYER, JR. P.; MOORE, R.; CHANG, K.; MUNISHKUMARAN, S. The digital database for screening mammography. In: **Third international workshop on digital mammography**, p. 27, 1996.

HELMUS, M. R.; BLAND, T. J.; WILLIAMS, C. K.; IVES, A. R. Phylogenetic measures of biodiversity. **The American Naturalist**, v. 169, n. 3, p. E68-E83, 2007.

Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). Mistério da Saúde, **Estimativa 2016. Incidência de Câncer no Brasil.** Disponível em:<<http://www.inca.gov.br/estimativa/2016/estimativa-2016-v11.pdf>>. Acesso em 15 de mar. de 2016.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, p. 159-174, 1977.

MAGURRAN, A. E. **Measuring biological diversity**. John Wiley & Sons, 2013.

MARTINEZ, E. Z.; LOUZADA-NETO, F.; PEREIRA, B. B. A curva ROC para testes diagnósticos. **Cad Saúde Coletiva**, v. 11, n. 1, p. 7-31, 2003.

METZ, C. E. ROC methodology in radiologic imaging. **Investigative radiology**, v. 21, n. 9, p. 720-733, 1986.

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**. 2013.70f. Dissertação (Mestrado na área de Ciência da Computação) - Programa de Pós-Graduação em Engenharia de Eletricidade, Universidade Federal do Maranhão, São Luís, 2013.

SAMPAIO, W. B. **Detecção de massas de imagens mamográficas usando uma metodologia adaptada à densidade da mama**. 2015.127f. Tese (Doutorado na área de Ciência da Computação) - Programa de Pós-Graduação em Engenharia de Eletricidade, Universidade Federal do Maranhão, São Luís, 2015.

SOUSA, U. S. **Classificação de massas na mama a partir de imagens mamográficas usando o índice de diversidade de shannon-wiener**. 2011.69f. Dissertação (Mestrado na área de Ciência da Computação) - Programa de Pós-Graduação em Engenharia de Eletricidade, Universidade Federal do Maranhão, São Luís, 2011.

VAPNIK, V.N.; VAPNIK, V. **Statistical learning theory**. New York: Wiley, 1998.

WEBB, C. O. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. **The American Naturalist**, v. 156, n. 2, p. 145-155, 2000.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, v. 11, n. 1, p. 10-18, 2009.



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

Identificação do Tipo de Documento

- Tese
- Dissertação
- Monografia
- Artigo

Eu, **Laércio Nascimento Mesquita**, 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 **Redução de falsos positivos em imagens de mamografias digitais usando os índices *Phylogenetic Species Variability*, *Phylogenetic Species Richness* e múltiplos classificadores** 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 28 de Julho de 2016.

A assinatura manuscrita de Laércio Nascimento Mesquita é escrita em uma linha decorativa. Abaixo da assinatura, o nome 'Assinatura' é impresso em uma fonte simples.

Assinatura