Redução de falsos positivos em imagens de mamografias digitais usando os índices *phylogenetic species variability*, *phylogenetic species richness* e múltiplos classificadores

Laércio N. Mesquita[1], Antônio O. de C. Filho[1], Alcilene D. de Sousa[1], Patrícia M. L. de L. Drumond[1]

[1]Campus Senador Helvídio Nunes de Barros (CSHNB) – Universidade Federal do Piauí (UFPI) – Picos – PI – Brasil

{laerciomesquita90, antoniooseas, alcileneluzsousa}@gmail.com, patymedy@hotmail.com

Abstract: *The CADx 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 Computer-Aided Diagnosis. For the validation of the methodology, was used 200 mammography images, where 100 contain mass and the others not mass. 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.*

Resumo*: Os sistemas CADx 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 Computer-Aided Diagnosis. Para validação da metodologia, foram utilizadas 200 imagens de mamografia, onde 100 contêm massa e as demais não massa. 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.*

1. Introdução

Segundo o INCA (2015), 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.

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. A detecção precoce do câncer de mama pode também 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 saúde utilizam de ferramentas que darão uma segunda opinião quando usadas em exames de mamografias. Essas ferramentas são conhecidas como *Computer-Aided Diagnosis* (CADx), cujos sistemas são responsáveis por diminuírem 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 filogenéticos, que serão responsáveis por caracterizar as regiões de interesse em um sistema CADx, 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 2008], *Radial Basis Function* [Haykin 2008], *Random Forest* [Breiman 2001], *Support Vector Machine* (SVM) [Vapnik 1998] com intuito 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.

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

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 (μAG) objetivando 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 μAG 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 *Digital Database for Screening Mammography* (DDSM) é usada. 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 baseiam-se 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. Materiais e métodos

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 *Digital Database for Screening Mammography* (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 2008], MLP [Haykin 2008] como está demonstrado na Figura 1.

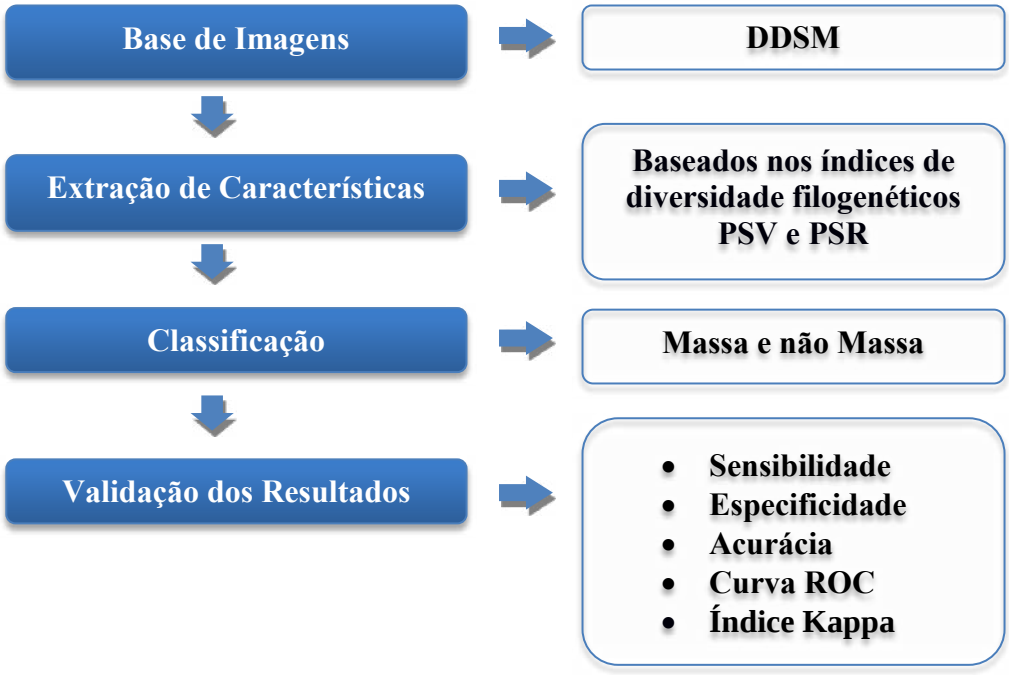


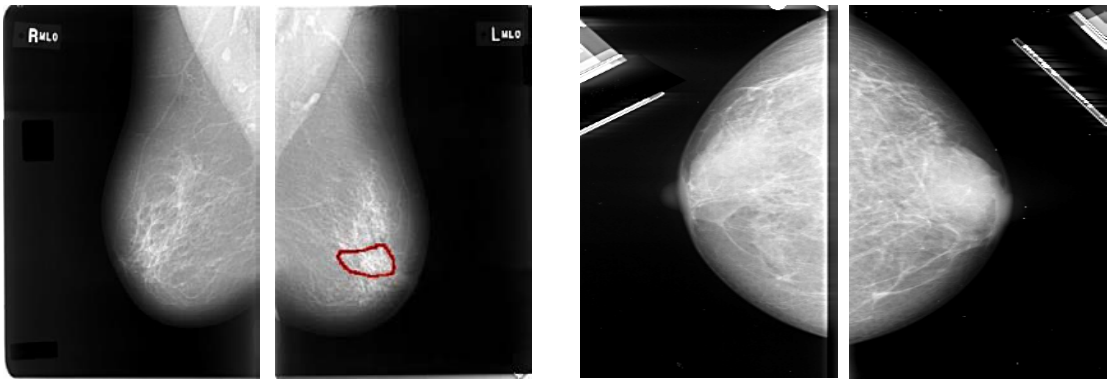
Figura 1. Metodologia proposta para redução de falsos positivos

3.1. Referencial Teórico

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.



a) MLO esquerdo **b) MLO direito** **c) CC esquerda** **d) CC direito**

Figura 2. Exemplos de Mamografia

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*. 2000]; 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*. 2000].

3.2 Base de Imagens

A *Digital Database for Screening Mammography (*DDSM) [Heath *et al*. 2000] 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. Para a concretização desse trabalho, foram utilizadas 200 imagens de mamografias; sendo 100 imagens com a presença de massa e 100 de não massa.

3.3 Índices de Diversidade Filogenética

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 isto 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* (ROIs) 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.

3.4 *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 fórmula 1 do 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, *n* é o número de espécies e $\bar{c}$ é a média dos elementos da diagonal de C.

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

3.5 Phylogenetic Species Richness

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

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

3.6 Classificação e Validação

A classificação foi realizada pelo *software 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 [WEKA 2016].

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 na tabela como (AUC) [Metz 1986], Sensibilidade (SE) [Martinez *et al.* 2003], Especificidade (EP) [Martinez *et al.* 2003] , Acurácia (AC) [Metz 1986] e índice Kappa (K) [Landis e Koch 1977]. Segundo [Martinez *et al.* 2003] a sensibilidade (SE) é definida como a probabilidade do teste sob investigação fornecer um resultado positivo, dado que o individuo é realmente portador da enfermidade. Ainda em [Martinez *et al.* 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 mede 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 e 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 e Discussões

Para os testes realizados neste trabalho, utilizou-se a base DDSM [Heath *et al*. 2000]. As características foram extraídas a partir dos índices de diversidade filogenéticos de variabilidade e riqueza de espécies. Os classificadores *MultiLayer Perceptron* (MLP) [Haykin 2008], *Radial Basis Function* (RBF) [Haykin 2008], *Random Forest* (RF) [Breiman 2001], *Support Vector Machine* (SVM) [Vapnik 1998] 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* [WEKA 2016]. A classificação foi realizada com validação cruzada de k-*folds,* sendo *k = 10*.

Tabela 1. 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

Nos testes realizados, 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 e 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 2. 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

Podemos observar na Tabela 2 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 e 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 3. 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 3, 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 e 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.

Tabela 4. 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 4, 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 1 vemos que o classificador SVM apresenta os melhores resultados, já nas tabelas 2 e 3, o classificador SVM não atingiu os resultados esperados, tendo em vista que todos os trabalhos relacionados atingiram resultados muito expressivos. Na tabela 3 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 e Trabalhos Futuros

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 para o que foi proposto. Para os trabalhos futuros, pretende-se utilizar novos índices filogenéticos a fim de uma melhor redução da taxa de falsos positivos, como também garantir uma maior eficiência nesse método.

Referências

Breiman, L. (2001) “Random forests”. Machine Learning, v. 45, n. 1, p. 5-32.

Carvalho, P. M. S. (2012) “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”. Dissertação de Mestrado na área de Ciência da Computação, (Programa de Pós-Graduação em Engenharia de Eletricidade), Universidade Federal do Maranhão, São Luís.

Digital Database for Screening Mammography. DDSM (2016). Disponível em: <http://marathon.csee.usf.edu/Mammography/Database.html>. Acesso em 29 de fevereiro de 2016.

Helmus MR, Bland TJ, Williams CK, Ives AR (2007) “Phylogenetic measures of

biodiversity". American Naturalist, 169, E68–E83.

Heath, M.; Bowyer, K.; Kopans, D.; et al. The digital database for screening mammography. Citeseer, Proceedings of the 5th international workshop on digital mammography. p. 212–218, 2000.

Haykin, S. (2008) "Neural Networks and Learning Machines". 3ª ed. New Jersey: Prentice Hall, 936 p.

Instituto Nacional do Câncer – INCA (2015), Mistério da Saúde, Estimativa 2016. Incidência do Câncer no Brasil, http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama, março, 2016.

Instituto Nacional do Câncer – INCA (2016), Mistério da Saúde, Câncer de mama, http://www.inca.gov.br/conteudo_view.asp?id=1932, março, 2016.

Landis, J. Richard, and Gary G. Koch. "The measurement of observer agreement for categorical data." biometrics (1977): 159-174.

Lima, B. V. A. (2014) "Rotulação de Dados com Aprendizado Semi-Supervisionado", Dissertação de Mestrado na área de Ciência da Computação, (Programa de Pós-Graduação em Ciência da Computação), Universidade Federal do Piauí, Teresina.

Martinez E. Z; Louzada-Neto F; Pereira B. B (2003). "A curva ROC para testes diagnósticos". Cadernos Saúde Coletiva 11 (1): 7-31.

Metz C. E. (1986). " ROC methodology in radiologic imaging". Invest. Radiol, v. 21(9), p. 720-33.

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 na área de Ciência da Computação, (Programa de Pós-Graduação em Engenharia de Eletricidade), Universidade Federal do Maranhão, São Luís.

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

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

Vapnik. N, "Statistical Learning Theory". New York: John Wiley & Sons, 1998.

Webb, Campbell O. "Exploring the phylogenetic structure of ecological communities: an example for rain forest trees." The American Naturalist 156.2 (2000): 145-155.

WEKA - Machine Learning Group at the University of Waikato. Disponível em:<http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em 26 de fev. de 2016