Método de Detecção de Câncer de Ovário Utilizando Padrões Proteômicos, Análise de Componentes Independentes e Máquina de Vetores de Suporte

Wesley B. Dominices de Araujo[1], Lúcio F. A. Campos[1], Aline S. Furtado[2]

[1]Departamento de Engenharia da Computação – Universidade Estadual do Maranhão
Cidade Universitária Paulo VI – Tiririca, São Luís – MA – Brasil

[2]Departamento de Enfermagem – Faculdade Pitágoras-FAMA
Av. São Luís Rei de França – Turu, São Luís – MA – Brasil

wesleydominices@gmail.com,lucio@engcomp.uema.br,alinesanf@yahoo.com.br

Abstract. *It is proposed a CAD method to detect ovarian cancer, using Independent Component Analysis, the technique of Maximum Relevance and Minimum Redundancy, to reduce dimensionality and the computational cost, and Support Vector Machine, for classification of samples between presence or absence of cancer. The method was tested with a proteomic patterns set from SELDI-TOF database, and best performance was achieved with 10 features vector, resulting 98.80% of accuracy, with 95.65% of specificity and 100% of sensitivity.*

Resumo. *É proposto um método CAD para detectar câncer de ovário, utilizando Análise de Componentes Independentes, a técnica de Máxima Relevância e Mínima Redundância, para redução da dimensionalidade e custo computacional, e Máquina de Vetores de Suporte, para classificar as amostras entre presença ou ausência de câncer. O método foi testado com a base de dados de padrões proteômicos SELDI-TOF, e o melhor desempenho foi obtido com um vetor de 10 características, resultando em uma acurácia de 98,80%, com 95,65% de especificidade e 100% de sensibilidade.*

1. Introdução

O câncer de ovário é um tipo de câncer de origem ginecológica mais difícil de ser diagnosticado e o de menor chance de cura. É uma neoplasia de baixa incidência, mas de alta mortalidade. Apenas 25% dos casos diagnosticados são tratados e as baixas taxas de sucesso nos tratamentos estão associadas ao diagnóstico tardio da doença [INCa 2013].

Um grande esforço tem sido realizado para prover novos métodos e técnicas de sucesso para o auxílio no diagnóstico de câncer de ovário. Dentre os mais usuais, estão a ultrassonografia transvaginal e os marcadores tumorais. A ultrassonografia é o método propedêutico mais solicitado para o diagnóstico diferencial de tumores pélvicos e tem elevada precisão para a determinação de presença, tamanho, localização e característica destes tumores [Oncoguia 2012].

Um biomarcador ou marcador tumoral é em geral uma substância detectada no exame de sangue e que aumenta na presença de tumores malignos. Esses marcadores

são muito úteis para o acompanhamento da paciente com câncer de ovário, porém pouco confiáveis para o diagnóstico [Instituto do Câncer 2014]. O marcador tumoral CA-125 [Bast et. al. 1981] tem sido utilizado como metodologia de diagnóstico precoce, alcançando acurácia de 50% a 60% em pacientes ainda no estágio inicial da doença, aumentando assim a taxa de sucesso do diagnóstico precoce em aproximadamente 10%.

Para identificar e entender a interação entre os marcadores tumorais com patologias em humanos é importante que em paralelo com os dados clínicos sejam também obtidas informações sobre o conjunto de proteínas e de padrões codificados expressos pelo genoma (proteoma) entre tecidos e fluidos corporais normais e/ou alterados [Wilkins et. al. 1996].

Um dos métodos mais utilizados para obtenção de padrões proteômicos de forma precisa é o espectrômetro de massa, que é baseado na tecnologia de dessorção e ionização no tempo (*Surface Enhanced Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry* - SELDI-TOF MS) e têm mostrado resultados promissores nos últimos anos [Donald 2006], [Yang 2005].

Nas últimas décadas, a comunidade científica vem desenvolvendo técnicas de CAD (*Computer-Aided Diagnosis*) aplicadas ao câncer de ovário. A eficácia dos biomarcadores SELDI-TOF MS combinados com métodos de CAD têm mostrado sucesso no diagnóstico precoce de vários tipos de câncer, tais como câncer de ovário, câncer de próstata [Donald 2006], câncer colorretal [Yu 2004], câncer de pulmão [Yang 2005] entre outros.

Thakur et. al. (2011) utilizaram algoritmos de seleção de características e redes neurais *feed forward* para classificar 216 amostras entre câncer ou não câncer de ovário. Segundo os autores, o método proposto obteve 98% de sensibilidade e 96% de especificidade.

Whelean et. al. (2006) utilizaram análise quimiométrica para classificar 48 amostras com câncer e 46 amostras sem câncer (grupo controle). A técnica conseguiu alcançar uma taxa média de 100% de sensibilidade e especificidade.

Arieshanti et. al. (2013) combinaram as técnicas de clusterização *one-pass* e k-vizinhos mais próximos para discriminar 121 amostras do grupo câncer e 95 amostras do grupo controle. Obtiveram acurácia, sensibilidade e especificidade de 97,8%, 97,9% e 97,7%, respectivamente.

Neste trabalho, foi utilizada junto com os padrões proteômicos, a técnica de Análise de Componentes Independentes (*Independent Component Analysis*– ICA) para extração de características, somada com o algoritmo de Máxima Relevância e Mínima Redundância (mRMR), para selecionar as características mais significativas e obter o melhor conjunto destas, e posteriormente a classificação final utilizando a Máquina de Vetores de Suporte (*Support Vector Machine* - SVM).

Este trabalho foi divido da seguinte forma. Na seção 2 serão mostrados os métodos para extração e classificação das amostras. Na seção 3, os resultados e discussões, e finalmente na seção 4, as conclusões e considerações finais.

2. Método Proposto

O diagrama em blocos do método proposto é mostrado na Figura 1. O método consiste basicamente em: extrair as características significantes do sinal proteômico utilizando Análise de Componentes Independentes (ICA), realizar a redução de dimensionalidade utilizando a técnica de Máxima Relevância e Mínima Redundância (mRMR) e posteriormente, a classificação final feita através da Máquina de Vetores de Suporte (SVM). Todos os métodos abordados neste artigo serão descritos nas subseções subsequentes.

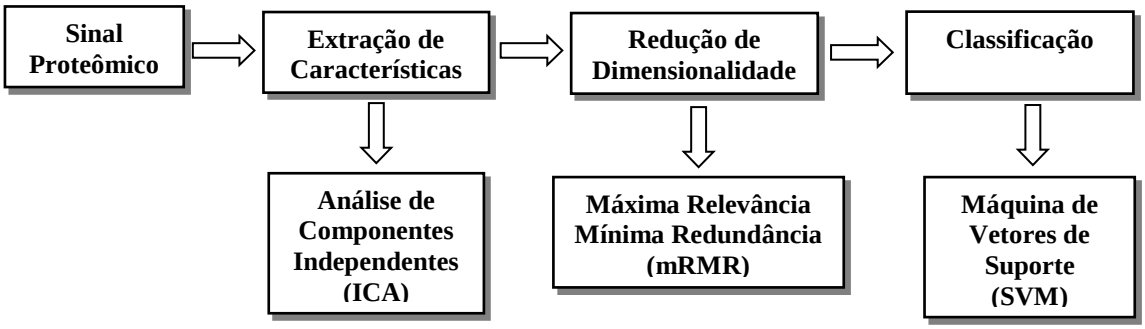

Figura 1. Diagrama em blocos do método proposto

2.1. Aquisição de Dados

Os dados foram baseados em padrões proteômicos usando a técnica SELDI-TOF, que por sua vez é um padrão de informação mais preciso para auxiliar no diagnóstico de câncer de ovário. Cada amostra adquirida possui 15.154 pontos [Seldi–MS 2002].

A Figura 2 ilustra uma amostra que foi extraída através de um espectrômetro de massa e que foi posteriormente convertida em um sinal multinível através dos níveis de intensidade proteômicos encontrados na amostra.

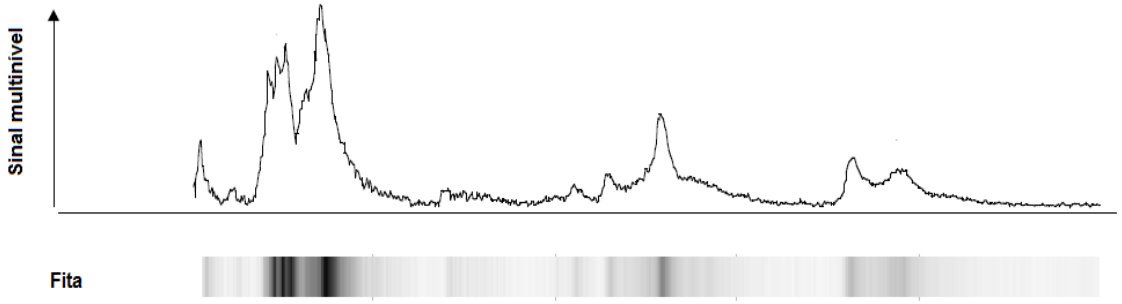

Figura 2. Relação entre sinal proteômico e níveis de intensidade

2.2. Extração de Características pela Análise de Componentes Independentes

Na etapa de extração de características, foi utilizada a Análise de Componentes Independentes (ICA), que é aplicada em diferentes situações, tais como processamento de sinais em reconhecimento de padrões em ECG e MEG [Vigário 1997] e câncer de mama [Campos, Costa and Barros 2008], [Costa, Campos and Barros 2011].

Considera-se que um dado ou sinal proteômico $\boldsymbol{x}$ extraído pode ser expresso como uma combinação linear de funções de bases $a_1, a_2, \cdots, a_n$, ponderadas por seus

coeficientes independentes $s_1, s_2, \cdots, s_n$ mútua e estatisticamente entre si [Hyvärinen and Oja 1997], tais que:

$$\boldsymbol{x_i = a_{i1}.s_1 + \cdots + a_{in}.s_n} \quad \text{para todo} \quad \mathbf{i = 1, \ldots, n} \tag{2.1}$$

Sendo:

- $\boldsymbol{x_i}$ = Sinal aleatório.
- $\boldsymbol{a_{ij}}$ = Coeficiente de mistura.
- $\boldsymbol{s_n}$ = Componente independente aleatório.

Onde cada $\boldsymbol{a_{ij}}$ é um coeficiente real. Define-se $\boldsymbol{X}$, $\boldsymbol{A}$ e $\boldsymbol{S}$ como:

$$\boldsymbol{X} = [\, \boldsymbol{x_1}\ \boldsymbol{x_2}\ \boldsymbol{x_n}\,]^{\boldsymbol{T}} \tag{2.2}$$

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a_{11}} & \cdots & \boldsymbol{a_{1n}} \\ \vdots & & \vdots \\ \boldsymbol{a_{n1}} & \cdots & \boldsymbol{a_{nn}} \end{bmatrix} \tag{2.3}$$

$$\boldsymbol{S} = [\, \boldsymbol{s_1}\ \boldsymbol{s_2}\ \boldsymbol{s_n}\,]^{\boldsymbol{T}} \tag{2.4}$$

Usando as equações (2.2), (2.3) e (2.4) para reescrever a equação (2.1), tem-se:

$$\boldsymbol{X = A.S} \tag{2.5}$$

O modelo apresentado na equação (2.5) é chamado de Análise de Componentes Independentes, que descreve como os dados são gerados a partir do processo de mistura com as componentes independentes.

2.2.1. Algoritmo *FastICA*

A matriz de dados $\boldsymbol{X}$ é considerada uma combinação linear das componentes não-gaussianas (independentes), tais que, $\boldsymbol{X=A.S}$, sendo que as colunas de $\boldsymbol{S}$ contêm as componentes independentes e $\boldsymbol{A}$ é a matriz de mistura. Em suma, ICA tenta "desmisturar" os dados, estimando uma matriz não misturada **W**, sendo $\boldsymbol{X.W = S}$.

Sob este modelo generativo de Análise de Componentes Independentes, a medida em $\boldsymbol{X}$ tenderá a ser mais Gaussiana que as componentes de origem $\boldsymbol{S}$. Assim, a fim de extrair as componentes independentes, busca-se uma matriz não misturada $\boldsymbol{W}$, que maximiza a não-gaussianidade das fontes. No Algoritmo *FastICA*, a não-gaussianidade é medida usando aproximações para negentropia (J), que são mais robustas do que as medidas de curtose e possuem um custo computacional menor [Marchini, Heaton and Ripley 2004]. A aproximação assume a seguinte forma:

$$J_{G(y)} = \left| E_y\{G(y)\} - E_v\{G(v)\} \right|^p \tag{2.6}$$

Sendo v uma variável aleatória gaussiana normalizada, y é assumido normalizado e com variância unitária, e o expoente p = 1, 2 tipicamente (A notação J_G não pode ser confundida com a notação da entropia negativa, J).

2.3. Seleção das características mais significantes

Identificar as características mais importantes dentre um vetor de características observado é uma das tarefas mais críticas encontradas em sistemas de reconhecimento de padrões. Tal tarefa é considerada de essencial importância para diminuir o erro de

classificação e o custo computacional [Webb 1999], [Jain, Duin and Mao 2000], [Kwak and Choi 2002], [Iannarilli and Rubin 2003].

As características irrelevantes podem ser removidas sem comprometer o resultado da classificação, pois neste contexto, são consideradas redundantes, ou seja, implicam na presença de outra característica com a mesma funcionalidade, e não trazem nenhuma informação nova ao vetor de características.

Neste trabalho, foi utilizado um critério, baseado em Máxima Relevância e Mínima Redundância (mRMR), que maximiza a informação mútua e minimiza a medida de redundância.

Uma das formas de selecionar características é através da Máxima Relevância descrita por:

$$maxD(v,c), \quad D = \frac{1}{|v|}\sum_{v_i \epsilon\, d} I(v_i; c) \tag{2.7}$$

Sendo v um vetor de características, c o vetor de classe e v_i uma característica individual.

É provável que as características selecionadas de acordo com o critério descrito anteriormente tenham muita redundância, ou seja, a dependência entre estas características pode ser grande. Para resolver tal problema, aplica-se em conjunto, a condição de Mínima Redundância, que seleciona mutuamente apenas as características mutuamente exclusivas [Ding and Peng 2003], tem-se, portanto:

$$minR(v), \;\; R = \frac{1}{|v|^2}\sum_{v_i, v_j \in v} I\left(v_i, v_j\right) \tag{2.8}$$

Os critérios descritos nas equações 2.7 e 2.8 são chamados conjuntamente de Máxima Relevância e Mínima Redundância (mRMR) [Peng, Long and Ding 2005].

Pode-se definir o operador $\phi(D,R)$ para combinar D e R, para em seguida otimizá-los simultaneamente, obtendo assim:

$$max\phi(D,R), \phi = D - R \tag{2.9}$$

2.4. Classificação

Como última fase, foi realizada a classificação das amostras através da Máquina de Vetores de Suporte (SVM). Isto foi feito analisando o vetor de características já reduzido através da técnica mRMR, rotulando-as em normal (grupo controle) ou anormal (câncer).

Para aumentar a confiabilidade do resultado e avaliar a capacidade de generalização do classificador e do método proposto, foi utilizada a técnica estatística de validação cruzada *10-fold cross-validation* [Kohavi 1995], onde o conjunto de dados é dividido igualmente em 10 subconjuntos. O treino efetua-se concatenando 9 subconjuntos e classifica-se o subconjunto restante. As fases de treino e teste são depois repetidas 10 vezes, permutando-se circularmente os subconjuntos. A taxa de acerto ou acurácia final é calculada usando a média das acurácias de cada fase, dando mais confiabilidade ao resultado final.

2.4.1. Máquina de Vetores de Suporte

A Máquina de Vetores de Suporte (SVM) é um método de aprendizagem supervisionada, capaz de classificar a partir de ***n*** indivíduos observados pertencentes a diversos subgrupos, a que classe um indivíduo que deve ser classificado pertence [Vapnik 1998].

A ideia da SVM é construir um hiperplano como superfície de decisão, de tal forma que a margem de separação entre as classes seja máxima possível. O objetivo do treinamento através da SVM é a obtenção de hiperplanos que dividam as amostras de tal maneira que sejam otimizados os limites de generalização.

As SVMs são consideradas sistemas de aprendizagem que utilizam um espaço de hipóteses de funções lineares em um espaço de muitas dimensões. Em casos em que o conjunto de amostras é composto por duas classes separáveis, um classificador SVM é capaz de encontrar um hiperplano baseado em um conjunto de pontos, denominados vetores de suporte, o qual maximiza a margem de separação entre as classes. Mesmo quando as duas classes não são separáveis, a SVM é capaz de encontrar um hiperplano através do uso de conceitos pertencentes à teoria da otimização [Ding and Peng 2003].

2.5. Métricas e Desempenho

Em processamento de sinais biomédicos e reconhecimento de padrões, a metodologia de desempenho usual é medida calculando algumas medidas estatísticas sobre o resultado dos testes [Bushberg et. al. 2001]. Os resultados da classificação dos testes podem ser divididos em: Verdadeiro Positivo (VP), Falso Positivo (FP), Verdadeiro Negativo (VN) e Falso Negativo (FN).

Sendo VP e VN o número de amostras que são corretamente identificadas como positiva ou negativa pelo classificador, FP e FN representam o número de amostras correspondentes aos casos que são erroneamente classificados como positivo ou negativo, respectivamente.

Tais números são utilizados para gerar medidas capazes de quantificar o desempenho da metodologia, para avaliar o quão este é eficiente e se os objetivos foram alcançados. As medidas de desempenho utilizadas neste trabalho foram: Acurácia, Especificidade, Sensibilidade e Área sob a Curva ROC (AuC).

3. Resultados e Discussões

O método proposto foi implementado usando a linguagem MatLab, R2013a, em conjunto com um Processador AMD Phenom II X4 B95 3.00 Ghz e 4 GB de memória RAM. Esta seção apresenta e discute os resultados obtidos nas abordagens utilizadas.

3.1. Bases de Dados

As bases de dados de serum SELDI MS são extensamente usadas em pesquisa para identificar padrões proteômicos e distinguir casos de câncer ou não câncer de ovário. As bases de dados são públicas, gratuitas, e podem ser adquiridas em [Seldi–MS 2002].

Para o câncer de ovário estão disponíveis duas bases de dados. A primeira base de dados consiste em 100 amostras com diagnóstico maligno, 100 amostras com diagnóstico normal (controle) e 16 casos com diagnóstico benigno. A segunda base de dados consiste em 162 amostras com diagnóstico maligno e 91 com diagnóstico normal,

totalizando 253 amostras. Cada amostra consiste em 15.154 níveis de intensidade ou características diferentes, conforme ilustrado na Figura 2.

Para este trabalho foi utilizada a base de dados maior, com 253 amostras, e em baixa resolução (15.154 pontos).

3.2. Extração de Características

A matriz **X** do modelo ICA foi gerada pela união da matriz dos casos com câncer, de dimensão 162x15.154, com a matriz das amostras de casos normais, de dimensão 91x15.154, formando a matriz X de dimensão 253x15.154.

A matriz **X** gerada serviu de entrada para o algoritmo *FastICA,* para que pudesse ser obtida as funções de base da matriz **A**, de dimensão 253x253, que contém as características de cada amostra. Cada linha da matriz **A** corresponde a uma amostra, e cada coluna corresponde a uma característica, ou seja, um parâmetro de entrada para o classificador [Christoyianni, Koutras and Kokkinakis 2002], [Campos, Barros and Silva 2007]. Usando o algoritmo *FastICA* e a matriz **X**, foi obtida a matriz de funções bases **A**, que contém as características de cada amostra.

3.3. Seleção das Características mais Significantes

Os testes para a redução do vetor de características de cada amostra foram realizados incrementando, de cinco em cinco, o número de características selecionadas pela técnica de Máxima Relevância e Mínima Redundância (mRMR), até o limite de cento e cinquenta, sendo que cada vetor gerado foi testado com o classificador de Máquina de Vetores de Suporte (SVM), a fim de encontrar o vetor de melhor desempenho.

3.4. Classificação

Na classificação foi utilizada a SVM com núcleo baseado em RBF (*Radial-Basis Function*), com a configuração padrão dos parâmetros, sem otimização dos mesmos. As amostras foram divididas em 10 subconjuntos, com o objetivo de realizar o teste de validação cruzada *10-fold cross-validation.*

A Tabela 1 mostra a média dos indicadores obtidos através do método *10-fold cross-validation* para 5, 10 e 15 características, vetores que obtiveram melhor desempenho durante o período de testes do classificador. Baseado nos resultados da Tabela 1 verifica-se que com somente 10 características das 253 possíveis o método obteve 98,80% de acurácia, 95,65% de especificidade e 100% de sensibilidade.

Tabela 1. Desempenho do classificador para cada vetor de características

Características	VP	FP	VN	FN	Acurácia (%)	Especificidade (%)	Sensibilidade (%)	AuC
5	162	8	83	0	96,83	95,30	100	0,961
10	**162**	**3**	**88**	**0**	**98,80**	**95,65**	**100**	**0,978**
15	162	5	86	0	98,02	96,62	100	0,972

Considerando o vetor de dez características, observou-se também, que das 162 amostras com câncer, todas foram classificadas corretamente (VP), logo não houve nenhum caso de câncer que foi classificado como normal (FN). Dos 91 casos com diagnóstico normal, somente em 3 casos (FP) houve erro de classificação, diagnosticando-os como câncer. A AuC obtida foi de 0,978, o que demonstra que o

classificador atingiu um bom desempenho, pois se aproximou-se de 1. O tempo de processamento do algoritmo, considerando somente o vetor de melhor desempenho, foi de apenas 2,80 segundos.

4. Conclusões e Considerações Finais

A extração de características utilizando ICA demonstrou ser efetiva em relação aos sinais extraídos do espectrômetro de massa pela técnica SELDI-TOF. A técnica de mRMR mostrou que a redução da dimensionalidade não afetou negativamente os resultados e ainda diminuiu o esforço computacional. A classificação com SVM, para dados não-lineares com duas classes, alcançou um excelente resultado e com um custo computacional bem pequeno.

Os resultados apresentados na Seção 3 demonstraram que o método proposto alcançou um bom desempenho. Com um vetor de apenas 10 características, o método obteve uma acurácia média de 98,80%, com especificidade de 95,65% e sensibilidade de 100%, em um estudo que utilizou 253 amostras com baixa resolução.

Para efeito de comparação com outros estudos relacionados, pode-se dizer que o método proposto alcançou resultados similares aos encontrados na literatura.

Em Thakur et. al. (2011) foram utilizadas 216 amostras, e a combinação das técnicas de seleção de características e redes neurais *feed foward*, alcançando 98% de sensibilidade e 96% de especificidade. Entretanto, as Redes Neurais como classificadores ainda não são dotadas de algoritmos de treinamento capazes de maximizar a capacidade de generalização [Hornik, Stinchombe and White 1989], [Cerqueira et. al. 2001].

Whelean et. al. (2006) utilizaram técnicas inovadoras, como análise quimiométrica. Entretanto, utilizaram apenas 94 das 256 amostras disponíveis na base de dados, conseguindo dessa forma 100% de sensibilidade e especificidade.

Arieshanti et. al. (2013) combinaram técnicas de clusterização aplicadas à base de dados com 216 amostras de alta resolução (370.000 pontos), obtendo acurácia, sensibilidade, especificidade de 97,8%, 97,9% e 97,7%, respectivamente.

Baseado no desempenho do método e no comparativo com outros encontrados na literatura, o trabalho encoraja o teste em bases de dados mais complexas, para ser obtido um modelo de conhecimento melhor do problema e posteriormente o desenvolvimento de um *software* que possa ser testado em hospitais e clínicas, auxiliando na redução da mortalidade para este e outros tipos de câncer. O custo do exame para retirar a amostra do *serum* ainda é elevado, por isso vai demorar um pouco a tornar-se popular, mas com o avanço da tecnologia, em algumas décadas poderá se tornar mais acessível, e até que isso seja concretizado deve se ter um *software* completo para auxiliar no diagnóstico de câncer.

Referências

Arieshanti, I., Purwananto, Y. and Tjandrasa, H. (2013) “Ovarian Cancer Identification using One-Pass Clustering and k-Nearest Neighbors”. TELKOMNIKA, Vol.11, No.4, December, pp. 797~802.

Bast, R. C., Feeney, M., Lazarus, H., Nadler, L. M., Colvin, R. B. and Knapp, R. C. (1981) "Reactivity of a monoclonal antibody with human ovarian carcinoma". *J. Clin. Invest*, November.

Bushberg, J. T., Seibert, A. J., Leidholdt, E. M. and Boone, J. M. (2001) "The Essential Physics of Medical Imaging", second ed., Lippincott Williams & Wilkins, Philadelphia, PA.

Campos, L. F. A., Barros, A. K. and Silva, A. C. (2007) "Independent Component Analysis and Neural Networks Applied for Classification of Malignant, Benign and Normal Tissue in Digital Mammography", In: Special Issue - Methods of Information in Medicine, v. 46, p. 212-215.

Campos, L. F. A., Costa, D. D. and Barros, A. K. (2008) "Segmentation of breast cancer in Digital Mammograms using texture features and independent component analysis". Proceedings of the BICS. Brain Inspired Cognitive Systems, Brazil.

Cerqueira, E. O., Andrade, J. C., Poppi, R. J. and Mello, C. (2001) "Determinação de constituintes químicos em madeira de eucalipto por pi-cg/em e calibração multivariada: comparação entre redes neurais artificiais e Máquinas de Vetor suporte". Quim. Nova, v. 24, p. 864.

Christoyianni, I., Koutras, A., Kokkinakis, G. (2002) "Computer aided diagnosis of breast cancer in digitized mammograms", In:Comp. Med. Imag. & Graph., 26:309-319.

Costa, D. D., Campos, L. F. A. and Barros, A. K. (2011) "Classification of breast tissue in mammograms using efficient coding". In *BioMedical Engineering OnLine*. Disponível em: http://www.biomedical-engineering-online.com/content/10/1/55.

Ding, C. and Peng, H. (2003) "Minimum Redundancy Feature Selection from Microarray Gene Expression Data", Proc. Second IEEE Computational Systems Bioinformatics Conf., p. 523-528, August.

Donald, D. (2006) "Bagged super wavelets reduction for boosted prostate cancer classification of seldi-tof mass spectral serum profiles", Chemometrics and intelligent Laboratory Systems, vol 82, no. 1, p. 2- 7, January.

INCa (2013). Instituto Nacional do Câncer [Online]. Disponível em: http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/ovario, acesso em 08/10/2013.

Hornik, K., Stinchombe, M. and White, H. (1989) "Multilayer feedforward networks are universal approximators". Neural Netw, v. 2, p. 359-366, Marth.

Hyvärinen, A. and Oja, E. (1997) "A fast fixed-point algorithm for independent component analysis", In: Neural Computation, 9(7):1483-1492.

Iannarilli, F. J. and Rubin, P. A. (2003) "Feature Selection for Multiclass Discrimination via Mixed-Integer Linear Programming", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, no. 6, pp. 779-783.

Jain, A. K., Duin, R. P. W. and Mao, J. (2000) "Statistical Pattern Recognition: A Review", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4-37.

Kohavi, R. (1995) “A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection”. In: International joint Conference on artificial intelligence, v. 14, p. 1137-1145.

Kwak, N. and Choi, C. H. (2002) “Input Feature Selection by Mutual Information Based on Parzen Window”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1667-1671.

Marchini, J. L., Heaton, C. and Ripley, B. D. (2004) “FastICA algorithms to perform ICA and Projection Pursuit”. Disponível em: http://www.stats.ox.ac.uk/~marchini/software.html.

Oncoguia (2012). “Exames de Imagem para o Diagnóstico do Câncer de Ovário”, http://www.oncoguia.org.br/conteudo/exames-de-imagem-para-o-diagnostico-do-cancer-de-ovario/1785/229/.

Instituto do Câncer (2014). “Câncer de Ovário”. Disponível em:

http://www.institutodocancer.com.br/php/index.php?link=5&sub=9.

Peng, H., Long, F. and Ding, C. (2005) “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy”. IEEE Transactions on Pattern Analisys and Machine Inteligence. Vol. 27, August.

Seldi–MS (2002) DATABASE. Disponível em:

http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp, Acesso em October, 2013.

Thakur, A., Mishra, V. and Jain, S. K. (2011) “Feed Forward Artificial Neural Network: Tool for Early Detection of Ovarian Cancer”. Scientia Pharmaceutica. Open Access. Avaliable at: http://dx.doi.org/10.3797/scipharm.1105-11.

Vapnik, V. N. (1998) “Statistical Learning Theory”. John Wiley and Sons.

Vigário, R. (1997) “Extraction of ocular artifacts form ecg using independent components analysis”, Eletroenceph. Clin. Neurophysiol., 103 (3) : 395-404.

Webb, A. (1999) “Statistical Pattern Recognition”. Arnold.

Whelean, O. P., Earll, M. E., Johansson, E., Toft, M. and Eriksson, L. (2006) “Detection of ovarian cancer using chemometric analysis of proteomic profiles”. Chemometrics and Intelligent Laboratory Systems 84, 82–87.

Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, R. D., Humphery-Smith, I, Hochstrasser, D. F. (1996) “Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it”. Biotechnol Genet Eng; 13:19-50.

Yang, S. Y. (2005) “Application of serum SELDI proteomic patterns in diagnosis of lung cancer”, BMC cancer, vol. 83, no. 5, September.

Yu, J. K. (2004) “An integrated approach to the detection of colorectal cancer utilizing proteomics and bioinformatics”, *World J. Gastroenterol*, vol. 21, no.10, pp. 3127-3131, October.