UNIVERSIDADE ESTADUAL DO MARANHÃO
CENTRO DE CIÊNCIAS TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS
MESTRADO PROFISSIONAL EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

WESLEY BATISTA DOMINICES DE ARAUJO

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

São Luís
2014

WESLEY BATISTA DOMINICES DE ARAUJO

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

Dissertação apresentada ao Mestrado Profissional de Engenharia da Computação e Sistemas da Universidade Estadual do Maranhão, como parte dos requisitos para a obtenção do título de Mestre em Engenharia da Computação e Sistemas.

Orientador: Prof. Dr. Lúcio Flávio de A. Campos
Co-orientadora: Profª. Ma. Aline S. Furtado

São Luís
2014

Araujo, Wesley Batista Dominices de.

Método de detecção de câncer de ovário utilizando padrões proteômicos, análise de componentes independentes e máquina de vetores de suporte / Wesley Batista Dominices de Araujo.– São Luís, 2014.

73 f.: il.

Dissertação (Mestrado) – Curso de Engenharia da Computação e Sistemas, Universidade Estadual do Maranhão, 2014.

Orientador: Prof. Dr. Lúcio Flávio de Albuquerque Campos

1. Câncer de ovário. 2. Análise de componentes independentes. 3. Máxima relevância e mínima redundância. 4. Máquina de vetores de suporte. I. Título

CDU: 618.11-006

WESLEY BATISTA DOMINICES DE ARAUJO

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

Dissertação apresentada ao Mestrado Profissional de Engenharia da Computação e Sistemas da Universidade Estadual do Maranhão, como parte dos requisitos para a obtenção do título de Mestre em Engenharia da Computação e Sistemas.

Aprovada em 02 de Setembro de 2014.

BANCA EXAMINADORA

Prof. Dr. Lúcio Flávio de Albuquerque Campos (Orientador)
Departamento de Engenharia da Computação e Sistemas
Universidade Estadual do Maranhão - UEMA

Profª. Ma. Aline Santos Furtado (Co-orientadora)
Departamento de Enfermagem
Faculdade Pitágoras - FAMA

Prof. MSc. Henrique Mariano Costa do Amaral (Membro)
Departamento de Engenharia da Computação e Sistemas
Universidade Estadual do Maranhão - UEMA

Prof. MSc. Luís Cláudio de Oliveira Silva (Membro externo)
Coordenação do Curso de Bacharelado Interdisciplinar em Ciência e Tecnologia
Universidade Federal do Maranhão - UFMA

AGRADECIMENTOS

A Deus, pelo dom da vida e por estar sempre presente em minha vida.

A meu orientador, prof. Dr. Lúcio Flávio de Albuquerque Campos, pela paciência, dedicação, credibilidade e confiança destinadas a mim durante todo o período de desenvolvimento deste trabalho.

À minha co-orientadora, profª. Ma. Aline Santos Furtado, pelo acompanhamento e auxílio durante todas as etapas deste trabalho.

À minha esposa, Raísa Oliveira da Silva Araujo, pelo amor, incentivo, compreensão e paciência durante todo o período do mestrado.

A meus pais, João Batista Serra de Araujo e Maria do Socorro Dominices de Araujo, e à minha irmã, Débora, pelo afeto e assistência incondicional ao longo de uma vida, fazendo com que nunca desistisse dos meus sonhos e objetivos.

A meus amigos e professores da 2ª turma do Mestrado de Engenharia da Computação e Sistemas da UEMA, pela fraternidade.

A todos os que direta e indiretamente contribuíram para a realização deste trabalho.

“O coração do inteligente adquire o conhecimento, e o ouvido dos sábios busca a sabedoria”.

Provérbios 18:15

RESUMO

O câncer de ovário é um tipo de câncer de origem ginecológica mais difícil de ser diagnosticado, pois a maioria dos tumores malignos de ovário só se manifesta no estágio avançado da doença, diminuindo assim a chance de cura. Somente cerca de 20% dos cânceres de ovário são diagnosticados precocemente. A ultrassonografia transvaginal é o método propedêutico mais utilizado para o diagnóstico diferencial, mas ainda não é eficaz para o diagnóstico precoce. Este trabalho propõe um método CAD (*Computer-Aided Diagnosis*) para detectar precocemente o câncer de ovário com o objetivo de auxiliar outros métodos já existentes, utilizando a técnica de Análise de Componentes Independentes (ICA), para a extração de características dos sinais proteômicos através da utilização do algoritmo FastICA, a técnica de Máxima Relevância e Mínima Redundância (mRMR), para diminuição do custo computacional e redução da dimensionalidade da matriz de características, isto se dá através da seleção das características mais significantes dentre todas as extraídas pela técnica de ICA. Após a extração e seleção das características utilizou-se a Máquina de Vetores de Suporte (SVM), para classificar as amostras entre presença ou ausência de câncer. O método foi testado na base de dados de padrões proteômicos SELDI-TOF, que contêm 253 amostras em baixa resolução (15.154 pontos), sendo 162 de câncer e 91 benignos. A partir dos testes realizados, o melhor desempenho foi obtido com um vetor de 10 características, resultando em uma taxa de acerto média de 98,80%, com 95,65% de especificidade e 100% de sensibilidade.

Palavras-chave: Câncer de Ovário, Análise de Componentes Independentes, Máxima Relevância e Mínima Redundância, Máquina de Vetores de Suporte.

ABSTRACT

Ovarian cancer is a kind of cancer gynecologic origin more difficult to be diagnosed, because the majority of malignant ovarian tumors is manifested only in the advanced stage of the disease, thus decreasing the chance of cure. Only about 20% of ovarian cancers are diagnosed early. Transvaginal ultrasonography is the most widely used diagnostic method for differential diagnosis, but is not yet effective for the early diagnosis. This paper proposes a CAD (Computer-Aided Diagnosis) method for early detection of ovarian cancer with the aim of helping other existing methods, using the Independent Component Analysis (ICA) technique, for feature extraction of proteomic signals using the FastICA algorithm, the technique of Maximum Relevance and Minimum Redundancy (mRMR), to decrease the computational cost and reduced dimensionality of the features matriz, it is through the selection of the most significant features among all extracted by ICA. After extraction and selection of features used the Support Vector Machine (SVM), to classify the samples between the presence or absence of cancer. The method was tested in the database SELDI-TOF proteomic patterns, containing 253 samples at low resolution (15,154 points), being 162 cancer and 91 benign. Based on the tests, the best performance was obtained with a vector of 10 features, resulting in an average accuracy of 98.80%, with 95.65% specificity and 100% sensitivity.

Keywords: Ovarian Cancer, Independent Component Analysis, Maximum Relevance and Minimum Redundancy, Support Vector Machine.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

1-DE	Eletroforese unidimensional
2-DE	Eletroforese bidimensional
ACS	*American Cancer Society* (Sociedade Americana de Câncer)
AUC	*Area Under Curve* (Área sob a Curva)
BRCA	*Breast Cancer, early onset*
BSS	*Blind Source Separation* (Separação Cega de Fontes)
CA-125	*Cancer Antigen 125*
CAD	*Computer-Aided Diagnosis* (Diagnóstico Auxiliado por Computador)
DNA	*DeoxyriboNucleic Acid* (*Ácido DesoxirriboNucleico*)
ECG	*ElectroCardioGraphy*
ESI	*Electrospray Ionization*
FN	Falso Negativo
FP	Falso Positivo
ICA	*Independent Component Analysis* (Análise de Componentes Independentes)
IEF	IsoElectric Focusing (Focalização isoelétrica)
INCA	Instituto Nacional do Câncer
MALDI	*Matrix-Assisted Laser Desorption/Ionization*
MEG	*MagnetoEncephaloGraphy*
mRMR	Máxima Relevância e Mínima Redundância
pH	Power of Hydrogen
pI	ponto Isoelétrico
RBF	*Radial Basis Function* (Função de Base Radial)
ROC	*Receiver Operating Characteristic* (Característica de Operação do Receptor)
SDS-PAGE	Sodium Dodecyl Sulfate- PolyAcrylamide Gel Electrophoresis
SELDI-TOF MS	*Surface Enhanced Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry*
SVM	*Support Vector Machine* (Máquina de Vetores de Suporte)
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

O câncer de ovário é um tipo de câncer de origem ginecológica mais difícil de ser diagnosticado, pois a maioria dos tumores malignos de ovário só se manifesta no estágio avançado da doença, diminuindo a 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 2014].

Segundo o Instituto Nacional de Câncer (2014), a maioria dos tumores de ovário são carcinomas epiteliais (câncer que se inicia nas células da superfície do órgão), tumor em células do tecido conjuntivo (mantêm os ovários juntos para a produção dos hormônios femininos) ou tumor maligno de células germinativas (que dão origem aos espermatozoides e aos ovócitos).

A última estimativa mundial apontou que ocorreram 238 mil novos casos de câncer de ovário no ano de 2012, com um risco estimado de 6,1 casos a cada 100 mil mulheres. As mais altas taxas de incidência podem ser observadas na parte ocidental e norte da Europa e na América do Norte. A África apresenta as taxas de incidência mais baixas. Mesmo em países de alto risco para o desenvolvimento do câncer de ovário, as taxas de incidência permanecem estáveis [INCa 2014].

O número de mortes por câncer de ovário no Brasil em 2011 foi de 3.027, e a estimativa para o ano de 2014 será de 5.680 novos casos, com um risco estimado de 5,58 casos a cada 100 mil mulheres [INCa 2014]. Nos Estados Unidos a estimativa de novos casos para o ano de 2014 é de 21.980 com 14.270 mortes [ACS-statistics 2014]. As estimativas são feitas normalmente a cada biênio.

Segundo o Instituto Nacional de Câncer (2014), sem considerar os tumores de pele não melanoma, o câncer do ovário é o quinto mais incidente na região Centro-Oeste, com

um risco estimado de 6,96/100 mil. Nas regiões Sul (6,63/100 mil), Sudeste (6,58/100 mil) e Nordeste (4,03/100 mil), é o sétimo. Já na região Norte, é o oitavo mais frequente, com um risco estimado de 2,52/100 mil.

No Maranhão, a estimativa do câncer de ovário para o ano de 2014 é de 80 casos, distribuídos da seguinte forma: 40 casos na capital, São Luís, e 40 casos no interior do Estado. A Figura 1 ilustra a estimativa de câncer no Brasil para o ano de 2014, somente em mulheres [INCa 2014].

Localização Primária	Casos Novos	%
Mama feminina	57.120	20,8%
Cólon e Reto	17.530	6,4%
Colo do útero	15.590	5,7%
Traqueia, Brônquio e Pulmão	10.930	4,0%
Glândula Tireoide	8.050	2,9%
Estômago	7.520	2,7%
Corpo do útero	5.900	2,2%
Ovário	5.680	2,1%
Linfoma não-Hodgkin	4.850	1,8%
Leucemias	4.320	1,6%
Sistema Nervosos Central	4.130	1,5%
Cavidade Oral	4.010	1,5%
Pele Melanoma	2.930	1,1%
Esôfago	2.770	1,0%
Bexiga	2.190	0,8%
Linfoma de Hodgkin	880	0,3%
Laringe	770	0,3%
Todas as Neoplasias sem pele*	**190.520**	
Todas as Neoplasias	274.230	

Figura 1 - Estimativa de Câncer em mulheres para o ano de 2014
Fonte: adaptada de [INCa 2014]

O fator de risco mais importante para o desenvolvimento do câncer de ovário é a história familiar de câncer de mama ou ovariano. Mulheres que já desenvolveram câncer de mama e são portadoras de mutações nos genes BRCA1 e BRCA2 possuem um risco aumentado de desenvolver câncer de ovário [ACS-riskfactors 2014].

Infelizmente, a prevenção desse tipo de neoplasia é limitada pelo pouco conhecimento de suas causas, além da falta de disponibilidade de técnicas para o diagnóstico precoce. Não existem comprovações de que o rastreamento do câncer seja suficientemente

efetivo para a população. Geralmente, os diagnósticos são feitos de forma ocasional ou quando o tumor já apresenta sintomas que indicam uma doença mais avançada [INCa 2014].

Esquemas de diagnóstico auxiliado por computador têm sido propostos com o objetivo de auxiliar no diagnóstico precoce de câncer, indicando áreas suspeitas, bem como anormalidades mascaradas. Esses esquemas CAD (*Computer-Aided Diagnosis*) têm sido desenvolvidos por vários grupos de pesquisa, visando auxiliar também na detecção precoce do câncer de ovário.

O diagnóstico auxiliado por computador é aquele no qual o profissional da saúde usa os resultados de uma análise computadorizada de imagens ou sinais médicos como uma "segunda opinião" na detecção de lesões, características em sinais e na elaboração do diagnóstico. Os esquemas CAD têm representado uma importante ferramenta no auxílio ao diagnóstico médico em diversas aplicações.

Este trabalho propõe um método CAD para auxiliar ao diagnóstico do câncer de ovário, utilizando dados ou sinais proteômicos. Para a extração de características dos sinais proteômicos foi utilizada a técnica de Análise de Componentes Independentes (*Independent Component Analysis* – ICA), somada com o algoritmo de Máxima Relevância e Mínima Redundância (mRMR), para selecionar as características mais significativas e reduzir a dimensionalidade da matriz gerada. Após a seleção das características mais relevantes, estas foram classificadas utilizando a Máquina de Vetores de Suporte (*Support Vector Machine* - SVM) entre duas classes, câncer ou normal.

1.1 Organização do Trabalho

Este trabalho será composto de mais cinco Capítulos, conforme descrição sumária a seguir.

No Capítulo 2 será mostrada a revisão teórica de literatura necessária ao desenvolvimento do método proposto. Apresenta conceitos sobre o câncer de ovário, proteômica, a técnica de Análise de Componentes Independentes, o algoritmo de Máxima Relevância e Mínima Redundância, e a Máquina de Vetores de Suporte para classificação das amostras.

No Capítulo 3 serão apresentados o objetivo geral e os objetivos específicos do trabalho.

No Capítulo 4 serão descritos a metodologia proposta juntamente com os materiais utilizados em cada etapa deste trabalho.

No Capítulo 5 apresentam-se os resultados obtidos, discussões, análise das técnicas utilizadas e a validação do método pelas medidas de desempenho: Acurácia, Especificidade, Sensibilidade e Área sob a Curva ROC (AuC).

No Capítulo 6, as conclusões e discussões sobre o trabalho, mostrando a eficiência do método proposto e sugestões para trabalhos futuros.

2 REVISÃO TEÓRICA

2.1 O Câncer

O corpo é composto de trilhões de células vivas. As células normais do corpo crescem, se dividem para fazer novas células e morrem de uma forma ordenada. Durante os primeiros anos de vida de uma pessoa, as células normais se dividem mais rapidamente para permitir que o indivíduo cresça. Depois que a pessoa se torna um adulto, a maioria das células se divide apenas para substituir células desgastadas ou mortas, ou para reparar lesões. O câncer começa quando células em uma parte do corpo começam a crescer fora de controle. Existem muitos tipos de câncer, mas todos eles começam por causa do crescimento descontrolado de células anormais [ACS 2014].

Segundo a Sociedade Americana de Câncer (2014), o crescimento das células cancerosas é diferente do crescimento celular normal. Em vez de morrer, as células cancerosas continuam a crescer e a formar novas células anormais. As células cancerosas também podem invadir ou crescer em outros tecidos, algo que as células normais não podem fazer. Crescer fora de controle e invadir outros tecidos são as causas que fazem com que uma célula seja, de fato, cancerosa. A Figura 2 mostra o crescimento celular descontrolado gerando um câncer invasivo.

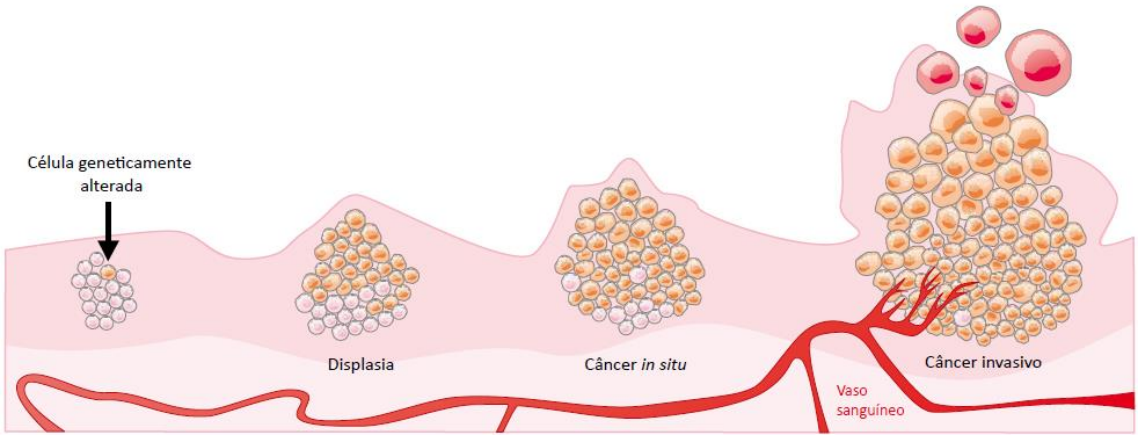

Figura 2 – Crescimento celular descontrolado
Fonte: [INCA 2014]

As células tornam-se cancerosas devido aos danos de DNA (*DeoxyriboNucleic Acid*). O DNA está em todas as células e direciona todas as suas ações. Em uma célula normal, quando o DNA é danificado, a célula ou repara o dano ou morre. Em células cancerosas, o DNA danificado não é reparado, mas a célula também não morre como deveria acontecer. Em vez disso, essa célula vai criar novas células que o corpo não precisa. Estas novas células terão o mesmo DNA danificado criado pela primeira célula [ACS 2014].

Na maioria dos casos, as células cancerosas formam um tumor. As células cancerosas muitas vezes migram para outras partes do corpo, onde elas começam a crescer e formar novos tumores que substituem o tecido normal. Este processo é chamado de metástase. Isso acontece quando as células cancerosas entram na corrente sanguínea ou nos vasos linfáticos do corpo [INCa 2014].

Nem todos os tumores são malignos. Os tumores que não são malignos são chamados de benignos. Os tumores benignos podem causar problemas, pois podem se tornar grandes e pressionar outros órgãos e tecidos saudáveis. Estes tumores normalmente não são fatais [ACS 2014]. A Figura 3 mostra as diferenças entre um tumor maligno e um tumor benigno.

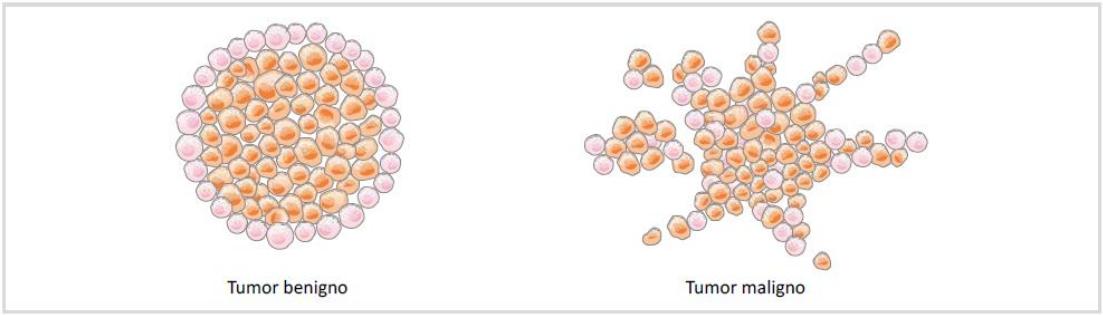

Figura 3 – Diferenças entre o tumor benigno e tumor maligno
Fonte: [INCA 2014]

2.2 O Ovário

Os ovários são glândulas reprodutivas encontradas apenas nas mulheres. Os ovários produzem ovos (óvulos) para a reprodução. Os ovos migram através das trompas de

Falópio para o útero, onde o óvulo fertilizado se implanta e se desenvolve em um feto. Os ovários também são a principal fonte dos hormônios femininos: estrogênio e progesterona [ACS 2014]. A Figura 4 mostra, de forma simplificada, o aparelho reprodutor feminino, onde se pode observar a localização dos ovários e que há dois ovários, direito e esquerdo, localizados na pelve, um em cada lado do útero.

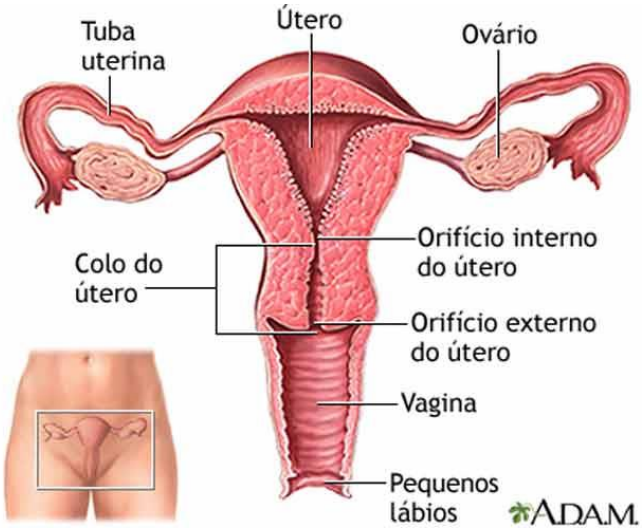

Figura 4 – Estrutura do aparelho reprodutor feminino
Fonte: [ADAM 2014]

No ovário são identificadas duas porções distintas: a **medula do ovário**, que é constituída por tecido conjuntivo frouxo, rico em vasos sanguíneos, célula hilares (intersticiais), e a **córtex do ovário**, rica em folículos ovarianos, corpo lúteo e células intersticiais. Na região cortical predominam os folículos ovarianos, os quais são formados por ovócitos envolvidos por células epiteliais (células foliculares ou da granulosa). A superfície do ovário é revestida por um epitélio (de origem celomática) que varia do tipo pavimentoso ao cilíndrico simples (denominado impropriamente de epitélio germinativo). Logo abaixo deste epitélio há uma camada de tecido conjuntivo denso, a túnica albugínea [UNIFESP 2014].

O estroma do ovário, entre as estruturas medulares e corticais, possui algumas células fusiformes, denominadas de células intersticiais ou de Leydig. Estas últimas respondem aos estímulos das gonadotrofinas e produzem hormônios sexuais,

principalmente androgênios. A hiperplasia destas células pode ser vista em afecções ovarianas relacionadas ao hiperandrogenismo (hirsutismo), como na síndrome dos ovários policísticos. Na região cortical, dependendo da fase e da idade, podem-se identificar: folículos ovarianos, corpo lúteo e corpos albicantes [UNIFESP 2014].

2.3 O Câncer de ovário

O câncer de ovário, como o próprio nome sugere, é o câncer que começa nos ovários. Os tumores de ovário são designados de acordo com o tipo de células onde o tumor se originou, e se é benigno ou canceroso. Existem vários tipos de câncer de ovário, dentre esses, três tem maior destaque [Oncoguia 2014], e serão descritos na subseção 2.3.1, a seguir.

2.3.1 Tipos de Câncer de Ovário

- **Câncer Epitelial de Ovário** – desenvolve-se a partir das células que cobrem a superfície externa do ovário. A maioria dos tumores epiteliais é benigna. Existem vários tipos de tumores benignos epiteliais, incluindo adenomas serosos, adenomas mucinosos e tumores de Brenner. Os tumores epiteliais malignos são carcinomas. Os subtipos do câncer epitelial de ovário incluem: tumores com baixo potencial de malignidade, tumores epiteliais malignos de ovário, carcinoma peritoneal primário e câncer das trompas de Falópio.

- **Câncer de Células Germinativas do Ovário –** desenvolve-se a partir das células que produzem os óvulos. A maioria dos tumores de células germinativas é benigna, embora alguns sejam cancerosos. Menos de 2% dos cânceres de ovário são tumores de células germinativas. Os subtipos do câncer de células germinativas de ovário incluem: teratoma, disgerminoma, tumor do seio endodérmico (tumor do saco vitelino) e coriocarcinoma.

- **Câncer Estromal do Ovário** - desenvolve-se a partir de células do tecido conjuntivo que mantêm os ovários juntos para a produção dos hormônios femininos estrogênio e progesterona. Os subtipos de câncer estromal de ovário incluem: tumores de

células granulosas, tumores da teca granulosa e tumores de células de Sertoli-Leydig. Estes tumores são bastante raros e geralmente são considerados cânceres de baixo grau, com cerca de 70% que se apresentam como a fase inicial da doença. Os tecomas e os fibromas são tumores benignos de estroma. Cerca de 1% dos cânceres de ovário são tumores do estroma ovariano. Mais da metade dos tumores estromais são encontrados em mulheres com mais de 50 anos, mas cerca de 5% dos tumores estromais ocorrem em mulheres jovens.

2.3.2 Fatores de risco para o câncer de ovário

Segundo a Sociedade Americana de Câncer (2014), um fator de risco é algo que muda a chance da paciente adquirir uma doença como o câncer. Cânceres diferentes têm diferentes fatores de risco. Por exemplo, a exposição desprotegida ao sol forte é um fator de risco para câncer de pele. O tabagismo é um fator de risco para uma série de tipos de câncer.

Os fatores de risco não dizem tudo. Ter um fator de risco, ou mesmo vários fatores de risco, não significa que a paciente terá a doença. E muitas pessoas que contraem a doença podem não ter tido quaisquer fatores de risco conhecidos. Mesmo se uma pessoa com câncer de ovário tem um fator de risco, é muito difícil saber o quanto esse fator de risco pode ter contribuído para o aparecimento do câncer.

Pesquisadores descobriram vários fatores específicos que mudam a probabilidade de desenvolver câncer epitelial de ovário de uma mulher. Esses fatores de risco encontrados em ACS-riskfactors (2014) não se aplicam a outros tipos menos comuns de câncer de ovário, como tumores de células germinativas e tumores estromais. Dentre os fatores de risco mais comuns para o câncer de ovário, destacam-se:

- **Idade** - O risco de desenvolver câncer de ovário aumenta quanto maior é a idade da paciente. O câncer de ovário é raro em mulheres com menos de 40 anos. A maioria

dos cânceres de ovário desenvolve-se após a menopausa. A metade de todos os cânceres de ovário é encontrada em mulheres 63 anos de idade ou mais.

- **Obesidade** - Vários estudos analisaram a relação entre obesidade e câncer de ovário. Em geral, nas mulheres obesas (aquelas com um índice de massa corporal de pelo menos 30%) têm um risco maior de desenvolver câncer de ovário.

- **Histórico reprodutivo** - As mulheres que já engravidaram têm um menor risco de ter câncer de ovário do que as mulheres que nunca engravidaram. O risco pode diminuir mais se a mulher engravidar mais de uma vez. A amamentação pode reduzir o risco ainda mais.

- **Controle de natalidade** - As mulheres que usam contraceptivos orais (também conhecidos como pílulas anticoncepcionais) têm um menor risco de ter câncer de ovário. A diminuição do risco somente é observada depois de 3 a 6 meses de utilização da pílula, e o risco diminui ainda mais quanto mais tempo os comprimidos forem utilizados.

- **Cirurgia ginecológica** - Laqueadura tubária pode reduzir a chance de desenvolver câncer de ovário em até 67%. A histerectomia (remoção do útero, sem retirar os ovários) também reduz o risco de ter câncer de ovário em aproximadamente 33%.

- **Drogas de Fertilidade** - pesquisadores descobriram que o uso do medicamento citrato de clomifeno fertilidade (Clomid ®) por mais de um ano pode aumentar o risco de desenvolvimento de tumores de ovário. O risco é maior em mulheres que não engravidam durante a utilização deste medicamento.

- **Terapia com estrogênio e hormônio** - Alguns estudos recentes sugeriram que as mulheres que utilizam estrogênios após a menopausa têm um risco aumentado de desenvolver câncer de ovário. O risco é maior em mulheres que tomam unicamente estrogênio (sem progesterona) por muitos anos (pelo menos cinco ou dez).

- **Histórico familiar** - O risco de câncer de ovário é aumentado se tiver casos de câncer na família da paciente. O risco fica maior, quanto mais pessoas da família da paciente tiveram câncer de ovário. Até 10% dos cânceres de ovário resultam de uma tendência hereditária para desenvolver a doença.

- **Histórico pessoal de câncer de mama** - Se uma mulher já teve câncer de mama, o risco de desenvolver câncer de ovário é aumentado. Alguns dos fatores de risco para o câncer de ovário também podem afetar o risco de câncer de mama. O risco de câncer de ovário após o câncer de mama é maior nas mulheres com histórico familiar de câncer de mama. Um forte histórico familiar de câncer de mama pode ser causado por uma mutação herdada nos genes BRCA1 ou BRCA2. Estas mutações também podem causar câncer de ovário.

2.3.3 Auxílio ao diagnóstico do câncer de ovário

Um grande esforço tem sido realizado para prover novos métodos e técnicas de sucesso para o auxílio ao diagnóstico de câncer de ovário. Dentre os mais usuais, estão a ultrassonografia transvaginal, tomografia computadorizada, ressonância magnética, biópsia do tumor e os marcadores tumorais.

A ultrassonografia transvaginal 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]. É um exame que usa ondas sonoras para visualizar o útero, trompas e ovários, colocando uma varinha de ultra-som na vagina. Ele pode ajudar a encontrar uma massa (tumor) no ovário, mas não pode realmente dizer se uma massa é câncer ou benigno.

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 como único método para o diagnóstico da doença [Instituto do Câncer 2014].

Baseado nos resultados do exame de sangue e da ultrassonografia transvaginal poderá ser indicada a biópsia (feita por laparoscopia ou laparotomia) do tecido ovariano. A biópsia do tumor é o método capaz de confirmar a existência de câncer de ovário.

2.4 O diagnóstico precoce do câncer de ovário

Somente cerca de 20% dos cânceres de ovário são encontrados em um estágio inicial. Quando isto acontece, cerca de 94% das pacientes vivem mais de 5 anos após o diagnóstico. Diversos estudos estão em andamento para aprender as melhores maneiras de encontrar o câncer de ovário em seu estágio inicial [ACS-detection 2014].

Durante um exame pélvico, o profissional de saúde verifica se o tamanho, forma e consistência dos ovários e útero estão adequados. Um exame pélvico pode ser útil, pois pode encontrar alguns tipos de câncer do sistema reprodutor em um estágio inicial, mas a maioria dos tumores ovarianos no estágio inicial é difícil ou impossível de ser sentido ou observado até mesmo para o examinador mais qualificado [ACS-detection 2014].

2.4.1 A Proteômica no auxílio ao diagnóstico

A ciência tem procurado marcadores moleculares que auxiliem no diagnóstico precoce e no tratamento de várias doenças humanas, incluindo o câncer. Vários estudos têm focado em alterações nos genes, seus transcritos e produtos proteicos envolvidos em processos celulares importantes.

Para identificar e entender essas alterações é fundamental conhecer o conjunto de proteínas codificadas pelo genoma. A partir deste entendimento surgiu o termo proteoma (proteínas de um genoma). O genoma representa a soma de todos os genes de uma pessoa

(características fixas). O proteoma é alterado conforme o estado de desenvolvimento do tecido ou sob as condições atuais da pessoa [ACS-detection 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%. O CA-125 é uma proteína encontrada no sangue. Em muitas mulheres com câncer do ovário, os níveis de CA-125 são elevados. Este teste pode ser útil para ajudar a orientar o tratamento e acompanhamento de mulheres que já estão com câncer de ovário, porque o nível do marcador tumoral CA-125 diminui, se o tratamento estiver funcionando.

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, chamada de proteômica [Wilkins et. al. 1996].

Segundo Barbosa et. al. (2012), as metodologias mais comumente utilizadas em dados proteômicos envolvem extração de proteínas da amostra, separação por eletroforese unidimensional (1-DE) ou bidimensional (2-DE) e/ou por cromatografia líquida, ionização, fragmentação, análise e detecção de peptídeos, e análise de dados. A Figura 5 mostra essas metodologias.

Para separação de proteínas por eletroforese uni (1-DE) e bidimensional (2-DE), as moléculas devem ser inicialmente isoladas de materiais biológicos, tais como tecidos e fluídos corporais.

Na eletroforese bidimensional usual, as proteínas são separadas em duas etapas consecutivas. Na primeira, denominada focalização isoelétrica (IEF), as moléculas migram

em gel de poliacrilamida com gradiente de pH imobilizado ou gerado por tampões anfoté ricos até atingirem um ponto (pH) no qual sua carga é igual a zero (ponto isoelétrico ou pI).

Na segunda etapa, as proteínas são submetidas a uma eletroforese com direção perpendicular a IEF em gel de poliacrilamida contendo dodecil sulfato de sódio (SDS-PAGE), e então separadas de acordo com sua massa molecular. Essa segunda etapa é similar a uma eletroforese 1-D, na qual as moléculas são diretamente aplicadas no gel SDS-PAGE e separadas de acordo com seu tamanho.

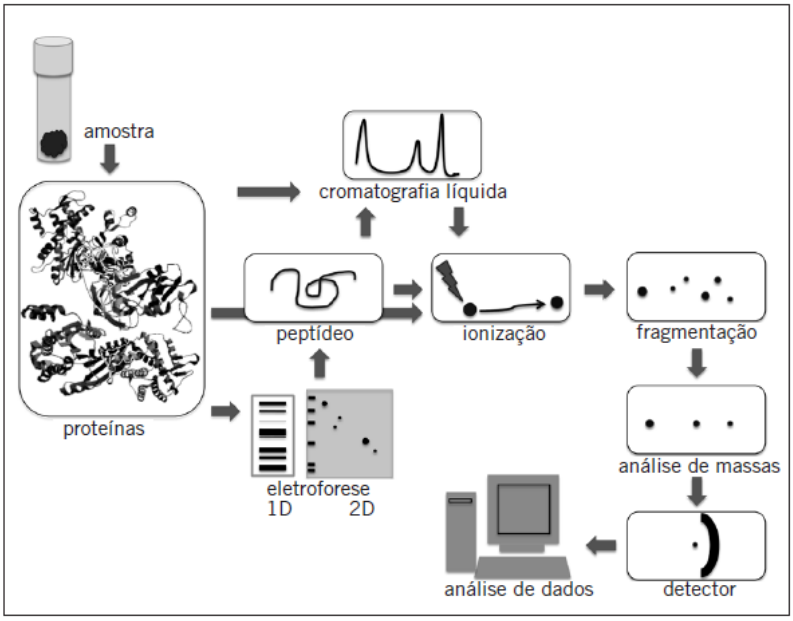

Figura 5 - Diferentes metodologias podem ser combinadas em estudos proteômicos
Fonte: [Barbosa et. al. 2012]

Atualmente, existem dois métodos principais de ionização utilizados em proteômica, o MALDI (*Matrix-Assisted Laser Desorption/Ionization*) e o ESI (*Electrospray Ionization*), o primeiro é empregado para amostras em estado sólido e o segundo para amostras em estado líquido. No método MALDI, os peptídeos são co-cristalizados com uma matriz orgânica, geralmente ácido α-ciano-4-hidroxicinaminico. Após bombardeamento por laser, a matriz sublima e seus íons transferem a carga para os analitos, resultando na formação de íons peptídicos [Karas, Bachmann and Hillenkamp 1985].

Uma variante do MALDI denominada SELDI (*Surface-Enhanced Laser Desorption/Ionization*) é geralmente empregada para análise do proteoma de baixo peso molecular e utiliza várias matrizes ou chips que exploram as características cromatográficas e biofísicas das diferentes proteínas. Esses chips podem apresentar superfícies hidrofóbicas, de troca iônica ou com íons metálicos imobilizados, ou mesmo anticorpos, receptores, enzimas e ligantes com alta afinidade por proteínas específicas. Assim, após a lavagem dos compostos não ligados, uma matriz é colocada sobre o chip e os espectros são obtidos por ionização com laser [Hutchens and Yip 1993].

Independentemente do método de ionização, a massa molecular dos íons é avaliada em um analisador após passagem por uma câmara de vácuo. Os tipos mais comuns de analisadores são o TOF (*Time-Of-Flight*), o quadrupolo e o *ion trap* [May et. al. 2011].

Nos analisadores TOF [Cotter 1994], os íons resultantes da primeira fase são acelerados por um potencial entre dois eletrodos e atravessam um tubo de vácuo com velocidade inversamente proporcional a sua massa. Quando os íons atingem o detector, o tempo decorrido entre a ionização e a detecção é utilizado para derivar o valor m/z. O valor m/z representa a relação entre a massa de um determinado íon e o número de cargas elementares que ele carrega. O detector converte o sinal da passagem do íon em sinal analógico, que é lido e interpretado por uma estação de trabalho. O resultado final é um gráfico de m/z versus intensidade (contagem de íons).

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].

Esta técnica utiliza um chip onde são depositadas as amostras de proteínas juntamente com um ácido, e independente das manipulações iniciais, apenas um subconjunto das proteínas na amostra se liga à superfície do chip. O analisador de massa utilizado usa a técnica *time-of-flight* (tempo de voo), sistema em que moléculas ionizadas aceleradas são lançadas em um tubo sob vácuo e sem campo elétrico para medida do seu tempo de voo até um detector. A Figura 6 mostra um espectrômetro de massa.

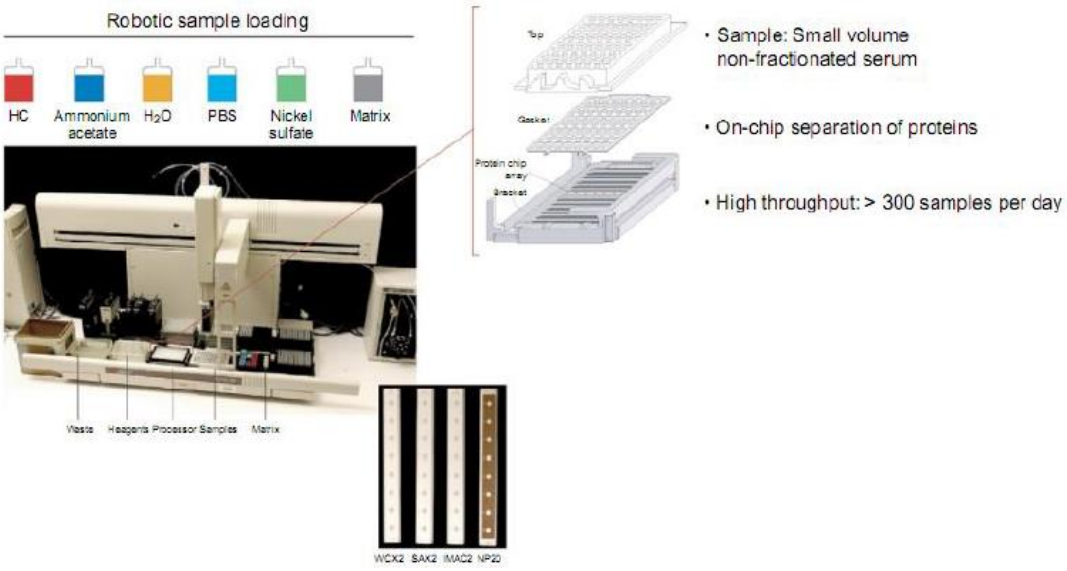

Figura 6 – Espectrômetro de massa utilizado para obtenção de padrões proteômicos
Fonte: adaptada de [Petricoin and Liotta 2004]

Nas últimas décadas, a comunidade científica também vem desenvolvendo técnicas de diagnóstico auxiliadas por computador (*Computer-Aided Diagnosis* - CAD) aplicadas ao câncer de ovário.

A eficácia dos biomarcadores SELDI-TOF MS combinados com as técnicas 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.

Petricoin et. al. (2002) combinaram algoritmos genéticos com mapas auto-organizáveis, para identificar câncer no ovário, obtendo sensibilidade de 100% (com intervalo de confiança de 95%) e especificidade de 95%.

Yu et. al. (2005) obtiveram 96,7% de sensibilidade e especificidade utilizando proteômica, ProteinChip *Software* 3.1 (Ciphergen) e SVM com *kernel* RBF.

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.

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.

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.

2.5 Análise de Componentes Independentes

A Análise de Componentes Independentes (*Independent Component Analysis* - ICA) é uma técnica que tem como objetivo a análise, ou separação, de fontes estatisticamente independentes a partir de um determinado modelo de mistura das fontes originais.

ICA é um método computacional desenvolvido inicialmente, para resolver problemas de Separação Cega de Fontes, do inglês *Blind Source Separation* (BSS) [Hyvärinen, Karhunen and Oja 2001]. O método é chamado de cego, pois muito pouco, ou nada, é conhecido sobre a matriz de mistura, e são feitas considerações sobre as fontes a serem estimadas.

Um exemplo clássico da aplicação da ICA é a separação da fala, conhecido como *cocktail-party*, onde três pessoas conversam em uma sala com três microfones posicionados em locais diferentes. Cada microfone grava uma amostra da conversa ao longo do tempo. Essas amostras são, portanto, misturas das falas de cada pessoa.

As técnicas de ICA propõem separar as três fontes partindo do simples pressuposto de que elas são estatisticamente independentes entre si. Mas a ordem das componentes independentes não pode ser estimada e a energia dos sinais não é a mesma da original, mas o som final é uma aproximação do original, praticamente o mesmo.

ICA 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].

2.5.1 Definições de ICA

Considera-se que um dado ou sinal proteômico $\boldsymbol{x}$, extraído de um espectrômetro de massa, 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:

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

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

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

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

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

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.

O objetivo deste modelo é permitir que se estime a matriz de mistura **A**, bem como a matriz de componentes independentes **S**, somente observando **X**.

A estimação das componentes é baseada nas seguintes condições:

- As componentes independentes são estatisticamente independentes;
- As componentes possuem distribuição não-gaussiana.

O modelo de ICA apresenta, no entanto, algumas ambiguidades no que diz respeito às componentes independentes:

- Não se podem determinar suas variâncias;
- Não se pode determinar sua ordem.

Tais ambiguidades se devem ao fato de **A** e **S** serem desconhecidas. Como consequência, não é possível determinar as energias ou as amplitudes dos sinais, nem tão pouco os sinais ou a ordem de s_n [Hyvärinen, Karhunen and Oja 2001].

2.5.2 Independência e Descorrelação

Duas variáveis são consideradas independentes quando o valor de uma não fornece informação acerca do valor da outra. Consideremos duas variáveis $\mathbf{x_1}$ e $\mathbf{x_2}$. Estas

variáveis são ditas independentes se, e somente se, $\mathbf{x_1}$ não fornecer nenhuma informação de $\mathbf{x_2}$ e vice-versa. Matematicamente, tem-se:

$$p(x_1, x_2) = p(x_1).p(x_2) \tag{2.6}$$

Usando outros termos, pode-se dizer que a probabilidade conjunta de $\mathbf{x_1}$ e $\mathbf{x_2}$ é igual ao produto das densidades marginais $p(x_1)$ e $p(x_2)$.

Duas variáveis $\mathbf{x_1}$ e $\mathbf{x_2}$ são descorrelacionadas se a sua covariância for igual a zero, ou seja:

$$cov_{x1,x2} = E[(x_1 - \mu_1)]E[(x_2 - \mu_2)] = 0 \tag{2.7}$$

Sendo $\boldsymbol{\mu_1}$ e $\boldsymbol{\mu_2}$ as médias das variáveis $\mathbf{x_1}$ e $\mathbf{x_2}$, respectivamente.

2.5.3 Estimação das Componentes Independentes

A estimação das componentes independentes $\boldsymbol{s_n}$ pode ser obtida através da matriz de mistura **A**, da seguinte forma:

$$\mathbf{S} = \mathbf{A^{-1}X} \tag{2.8}$$

Sendo a matriz **A** desconhecida, a ideia principal da análise de componentes independentes consiste em considerar que os sinais observáveis $\boldsymbol{x_n}$ estão relacionados com os sinais originais através de uma transformação linear. Assim, os sinais originais podem ser obtidos a partir de uma transformação inversa. Supondo, dessa forma, uma combinação linear de $\boldsymbol{x_i}$, de modo que:

$$y = b^T X \tag{2.9}$$

Sendo **X=AS**, pode-se escrever:

$$y = b^T \mathbf{AS} \tag{2.10}$$

Onde $\boldsymbol{b}$ deve ser determinado. A partir da equação (2.10), é possível observar que $\boldsymbol{y}$ é uma combinação linear se $\boldsymbol{s_i}$, com coeficientes dados por $q=b^T\mathbf{A}$. Sendo assim, obtêm-se:

$$y=q^T\mathbf{S} \tag{2.11}$$

Se $\boldsymbol{b}$ corresponde a uma das linhas da inversa de **A**, então $\boldsymbol{y}$ será uma das componentes independentes, e neste caso, apenas um dos elementos de $\boldsymbol{q}$ será igual a um, e todos os outros serão iguais a zero. No entanto, sendo **X** conhecido, $\boldsymbol{b}$ não pode ser determinado exatamente, porém pode-se estimar seu valor.

Uma forma de determinar $\boldsymbol{b}$ é variar os coeficientes em $\boldsymbol{q}$ e verificar como a distribuição de $y=q^T\mathbf{S}$ muda. Como pelo Teorema do Limite Central [Papoulis and Pillai 2002], a soma de duas variáveis aleatórias independentes é mais gaussiana que as variáveis originais, $\boldsymbol{y}$ é mais gaussiana que qualquer uma das $\boldsymbol{s_i}$ e menos gaussiana quando se iguala a umas das $\boldsymbol{s_i}$. Assim, apenas um elemento $\boldsymbol{q_i}$ de $\boldsymbol{q}$ é diferente de zero [Hyvärinen, Karhunen and Oja 2001]. Como, na prática, os valores de $\boldsymbol{q}$ são desconhecidos, e através da equação (2.9) e da equação (2.11) tem-se que:

$$b^T\mathbf{X} = q^T\mathbf{S} \tag{2.12}$$

Pode-se variar $\boldsymbol{b}$ e observar a distribuição de $b^T\mathbf{X}$. Dessa forma, pode-se tomar como $\boldsymbol{b}$ um vetor que maximiza a não-gaussianidade de $b^T\mathbf{X}$, sendo $q=\mathbf{A^TS}$, contendo apenas uma de suas componentes diferente de zero. Isso significa que $\boldsymbol{y}$ na equação (2.9) é igual a uma das componentes independentes, e a maximização da não-gaussianidade de $b^T\mathbf{X}$, permite encontrar uma das componentes.

2.5.4 Negentropia como medida de Não Gaussianidade

A entropia de uma variável aleatória está relacionada com a quantidade de informação que essa variável possui. Sendo $\boldsymbol{y}$ um vetor aleatório com função densidade de probabilidade *f(y)*, a sua entropia diferencial é dada por:

$$H(y) = -\int f(y) log f(y) dy \tag{2.13}$$

Sabendo-se que uma variável gaussiana tem a maior entropia dentre todas as variáveis aleatórias de igual variância [Hyvärinen, Karhunen and Oja 2001], [Papoulis and

Pillai 2002], tem-se que uma versão modificada da entropia diferencial pode ser usada como medida de não-gaussianidade. Tal medida é denominada negentropia, definida por:

$$J(\mathrm{y}) = \mathrm{H}(\mathrm{y}_{\mathrm{gauss}}) - \mathrm{H}(\mathrm{y}) \tag{2.14}$$

Sendo $\mathrm{y}_{\mathrm{gauss}}$ uma variável aleatória de mesma matriz de covariância que ***y***.

A negentropia é sempre não negativa, e pode assumir zero se, e somente se, ***y*** tem distribuição gaussiana e é invariante para transformações lineares inversíveis.

Apesar de permitir que se possa medir não-gaussianidade, a negentropia é de difícil estimação, sendo necessária sua estimação por aproximações através de momentos de alta ordem. Assim,

$$J(y) \approx \frac{1}{12} E\{y^3\}^2 + \frac{1}{48} kurt(y)^2 \tag{2.15}$$

Sendo $kurt(y)$ a *kurtosis* de ***y***, definida como o momento de quarta ordem da variável aleatória ***y***, definida por:

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2 \tag{2.16}$$

2.5.5 Algoritmo FastICA

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

Sob este modelo generativo de Análise de Componentes Independentes, a medida em ***X*** tenderá a ser mais Gaussiana que as componentes de origem ***S***. Assim, a fim de extrair as componentes independentes, busca-se uma matriz não misturada ***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:

$$J_{G(y)} = \left| E_y \{G(y)\} - E_v \{G(v)\} \right|^p \tag{2.17}$$

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

2.6 Seleção de 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 [Peng, Long and Ding 2005], [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.

Seja $\boldsymbol{v}$ um vetor de dados, com $\boldsymbol{n}$ amostras e $\boldsymbol{m}$ características, representado por $v = \{v_i, i = 1, \cdots m\}$ e seja $\boldsymbol{c}$ um vetor de classe (rótulo).

O problema de seleção de características é encontrar do conjunto de observação de dimensão $\boldsymbol{v}$, um subconjunto de $\boldsymbol{m}$ características que represente $\boldsymbol{c}$, de maneira satisfatória e efetiva.

Dada a condição de encontrar um mapeamento "ótimo", o algoritmo deve buscar a melhor forma de encontrar este subconjunto. As duas formas mais comuns são: classificando-as por algum critério e selecionando as $\boldsymbol{k}$ melhores características, a segunda é escolher um subconjunto mínimo dentro do conjunto de características sem afetar a precisão da classificação.

Sendo assim, na seleção de um subconjunto eficiente, os algoritmos podem automaticamente determinar o número de características, ou o operador pode estipular o tamanho do subconjunto de características.

A condição de caracterização significativa implica em um erro de classificação mínimo possível, que requer a máxima dependência estatística entre o subconjunto $\boldsymbol{m}$ selecionado, e o vetor de classe $\boldsymbol{c}$. Tal esquema é chamado de Máxima Dependência.

Em termos de informação mútua *I*, que vem da teoria da informação, a proposta de realizar a seleção de características para encontrar um vetor *v*, com *m* características $\{v_i\}$, que conjuntamente tenham a maior dependência possível com o vetor de classe $\boldsymbol{c}$, é dada por:

$$maxD(v,c), \quad D = I(\{v_i, i = 1, \cdots, m\}: c) \tag{2.18}$$

Entretanto, a equação (2.18) pode demandar um esforço computacional grande, quando os dados são multivariados, já que para estimar o vetor *v*, é necessário calcular inúmeras vezes a inversa da matriz de covariância. Por este motivo, o algoritmo para encontrar a máxima dependência entre variáveis é considerado de grande esforço computacional, ocasionando um custo computacional elevado.

Para diminuir o esforço computacional, utiliza-se uma técnica baseada em Máxima Relevância e Mínima Redundância (mRMR), que maximiza a informação mútua e minimiza a medida de redundância.

2.6.1 Máxima Relevância e Mínima Redundância

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

$$maxD(v,c), \quad D = \frac{1}{|v|}\sum_{v_i \epsilon\, d} I(v_i;c) \tag{2.19}$$

Sendo $\boldsymbol{v}$ um vetor de características, $\boldsymbol{c}$ o vetor de classe e $\boldsymbol{v_i}$ 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:

$$minR(v), \quad R = \frac{1}{|v|^2}\sum_{v_i,v_j \in v} I\left(v_i, v_j\right) \tag{2.20}$$

Os critérios descritos nas equações (2.19) e (2.20) 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:

$$max\phi(D,R), \phi = D - R \tag{2.21}$$

Com o vetor de características reduzido pela técnica de Máxima Relevância e Mínima Redundância, pode então ser feita a classificação das amostras. O que será mostrado na próxima Seção.

2.7 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. A Figura 7 ilustra hiperplanos de separação entre duas classes linearmente separáveis.

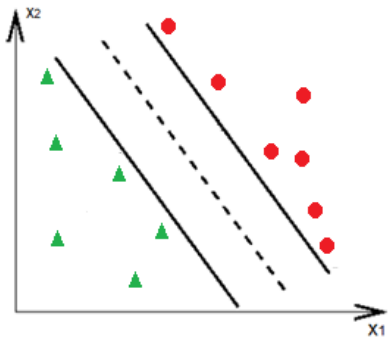

Figura 7 – Separação de duas classes, com o auxílio de vetores de suporte

O hiperplano ótimo (linha tracejada) separa as duas classes, com o auxílio dos vetores de suporte (linhas continuas) e mantém a maior distância possível com relação aos pontos da amostra.

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].

2.7.1 Definições da SVM

Para simplificar, será considerado o caso de classificação utilizando duas classes, usando um modelo linear descrito por:

$$y(z) = w^T z + \epsilon = 0 \tag{2.22}$$

Sendo $\boldsymbol{w}$ é um vetor de pesos ajustados, $\boldsymbol{\epsilon}$ é um viés e $\boldsymbol{z}$ é um vetor de treinamento de características, com seus respectivos rótulos $y_i \in Y$, em que $Y = \{-1, +1\}$. O modelo definido na equação (2.22) define um hiperplano ótimo, que classifica todos os vetores de treinamento, e $\boldsymbol{z}$ é dito linearmente separável, se é possível separar os dados das classes –1 e +1 por este hiperplano [Schölkopf and Smola 2002]. A Figura 8 ilustra um hiperplano ótimo para padrões linearmente separáveis.

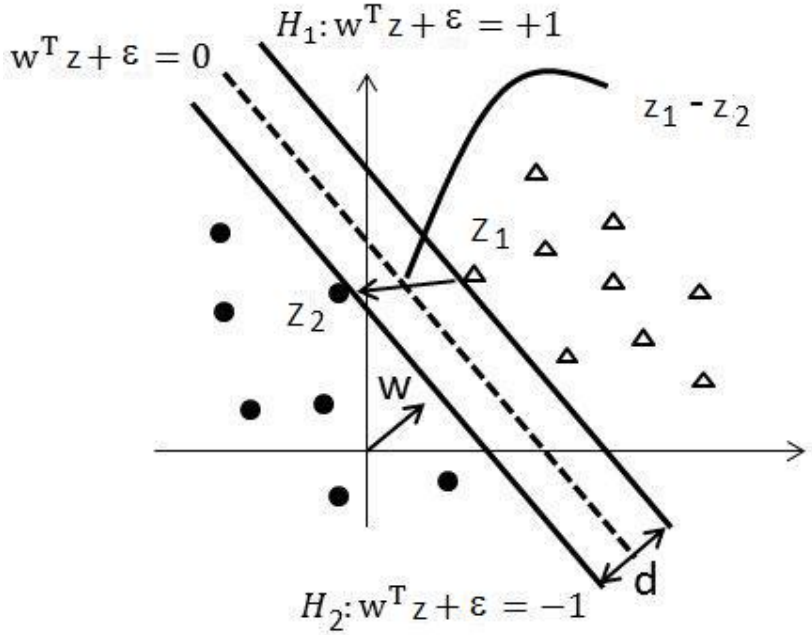

Figura 8 - Hiperplano ótimo, com dois vetores de suporte H_1 e H_2

Baseado no modelo apresentado, tem-se:

$$w^T z + \epsilon \geq 0, \text{ para y(z)=+1} \tag{2.23}$$

$$w^T z + \epsilon < 0, \text{ para y(z) = -1} \tag{2.24}$$

Sendo Z_1 um ponto pertencente ao hiperplano $H_1 = w^T z + \epsilon = +1$, e Z_2 um ponto pertencente ao hiperplano $H_2 = w^T z + \epsilon = -1$, conforme ilustrado anteriormente na Figura 8.

Ao se projetar $Z_1 - Z_2$ na direção de **W**, perpendicular ao hiperplano ótimo, é possível obter a distância (*dist*) entre os hiperplanos H_1 e H_2, dada pela equação a seguir:

$$dist = (z_1 - z_2)\left(\frac{W}{\|W\|}.\frac{(Z_1 - Z_2)}{\|Z_1 - Z_2\|}\right) \tag{2.25}$$

De acordo com as equações (2.23) e (2.24), tem-se que $H_1 = w^T z + \epsilon = +1$ e $H_2 = w^T z + \epsilon = -1$. A diferença entre as duas equações dão como resultado $W.(Z_1 - Z_2) = 2$. Substituindo o resultado encontrado na equação (2.25), tem-se:

$$\frac{2(Z_1 - Z_2)}{\|W\|.\|Z_1 - Z_2\|} \tag{2.26}$$

Tomando-se a norma da equação (2.26) acima, tem-se:

$$\frac{2}{\|W\|} \tag{2.27}$$

Esta é a distância ***d***, já ilustrada na Figura 8, entre os hiperplanos H_1 e H_2, paralelos ao hiperplano ótimo separado. Minimizando $\|w\|$, pode-se minimizar a margem de separação dos dados em relação ao $w^T z + \epsilon = 0$. Assim, tem-se o problema de otimização descrito por:

$$min_{w,\epsilon}\, \frac{1}{2}\|w\|^2 \tag{2.28}$$

Com a restrição $y_i(w^T z_i + \epsilon) - 1 \geq 0, \forall_i = 1,2,\ldots,n,$ que é imposta para assegurar que não haja dados de treinamento entre as margens de separação das classes. Este é um problema de otimização quadrática, cuja função objetiva é convexa e os pontos que satisfazem as restrições formam um conjunto convexo, logo possui um único mínimo global.

Para solucionar tal problema, aplica-se um operador Lagrangiano, capaz de englobar as restrições às funções objetivo, associadas aos multiplicadores de Lagrange, conforme a equação a seguir:

$$K(w, \epsilon, \rho) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \rho_i (y_i(w^T z_i + \epsilon) - 1) \tag{2.29}$$

Onde o operador Lagrangiano deve ser minimizado, implicando na minimização das variáveis α_i e na minimização de w e ϵ.

Igualando as derivadas de K em relação a $\boldsymbol{\epsilon}$, e **w** a zero, obtêm-se as condições abaixo:

$$w = \sum_{i=1}^{n} \rho_i y_i z_i \tag{2.30}$$

$$\sum_{i=1}^{n} \rho_i y_i = 0 \tag{2.31}$$

Substituindo as equações (2.30) e (2.31) na equação (2.29), tem-se o seguinte problema de otimização exposto nas equações abaixo:

$$\max_{\rho} \sum_{i=1}^{n} \rho_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_i \rho_j y_i y_j (z_i . z_j) \tag{2.32}$$

$$\text{Sujeito a} \begin{cases} \alpha_i \geq 0, \forall_i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases} \tag{2.33}$$

Sendo que $(z_i . z_j)$ corresponde ao produto interno entre z_i *e* z_j.

Como se trata de padrões não separáveis linearmente, não é possível construir um hiperplano ótimo de separação sem encontrar erros de classificação [Haykin 1999]. Para o caso de pontos de dados não separáveis, é introduzido um conjunto de variáveis escalares não

negativas, ε_i, na definição do hiperplano de separação, para resolver este problema, assim tem-se:

$$u_i(w^T z + \rho) \geq 1 - \varepsilon_i,\ i = 1, 2, \cdots, n \tag{2.34}$$

As variáveis ε_i são chamadas variáveis de folga e medem o desvio de um ponto de dado na condição ideal de separabilidade de padrões. Já *C* um parâmetro positivo especificado pelo usuário. Dessa forma, o problema agora é:

$$\min_{w,\varepsilon} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \varepsilon_i \tag{2.35}$$

$$\text{Sujeito a: } y_i(w.z_i + \epsilon) \geq 1 - \varepsilon_i,\ \ \varepsilon_i \geq 0,\ \forall_i = 1, \cdots, n \tag{2.36}$$

Novamente, trata-se de um problema de otimização quadrática, com as restrições lineares dadas pela equação (2.36). Aplicando o operador Lagrangiano, obtêm-se:

$$\max_{\rho} \sum_{i=1}^{n} \rho_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_i \rho_j y_i y_j \left(z_i . z_j\right) \tag{2.37}$$

$$\text{Sujeito a } \begin{cases} 0 \leq \rho_i \geq C, \forall_i = 1, \ldots, n \\ \sum_{i=1}^{n} \rho_i y_i = 0 \end{cases} \tag{2.38}$$

Os problemas não lineares de classificação são resolvidos mapeando o conjunto de treinamento, saindo de seu espaço de entrada para um novo espaço com maior dimensão, denominado espaço de características. Seja $\theta(z)$, uma função que mapeia o espaço de entrada sobre o espaço de características. Esta função é usada para mapear z_i e z_j para o espaço de características, antes da realização do produto interno entre eles:

$$k(z_i, y_i) = \theta(z_i).\theta\left(z_j\right) \tag{2.39}$$

Modificando desta forma, o problema de maximização proposto na equação (2.39), para:

$$\max_{\rho} \sum_{i=1}^{n} \rho_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_i \rho_j y_i y_j k\left(z_i . z_j\right) \tag{2.40}$$

Com as restrições da equação (2.40). A equação (2.39) é chamada de função de núcleo (*kernel)*.

Uma função de núcleo bastante utilizada e com excelentes resultados em SVM é a Função de Base Radial, do inglês *Radial Basis Function* – RBF, e é descrita por:

$$k(z_i . z_j) = e^{-\gamma \|z_i - z_j\|^2} \tag{2.41}$$

Sendo γ o parâmetro, que é definido a priori, pelo usuário.

3 OBJETIVOS

3.1 Objetivo geral

Propor um método de classificação de câncer de ovário utilizando padrões proteômicos para auxiliar no diagnóstico junto com outros métodos já existentes, tais como o marcador tumoral CA-125 e a ultrassonografia transvaginal.

3.2 Objetivos específicos

- Propor um método para auxílio ao diagnóstico precoce de câncer de ovário auxiliado por computador.
- Contribuir para redução de mortalidade de câncer de ovário por meio do auxílio ao diagnóstico precoce não invasivo.
- Avaliar as técnicas de Análise de Componentes Independentes, Máxima Relevância e Mínima Redundância, e a Máquina de Vetores de Suporte como classificador.
- Avaliar o sistema proposto através da sensibilidade, especificidade, acurácia e área sob a curva ROC.

4 MATERIAIS E MÉTODOS

Este Capítulo contém a descrição dos materiais e métodos utilizados para implementação da metodologia para o auxílio ao diagnóstico de câncer de ovário. O método consiste basicamente em: extrair as características 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 trabalho serão descritos nas subseções subsequentes. O diagrama em blocos do método proposto é mostrado na Figura 9.

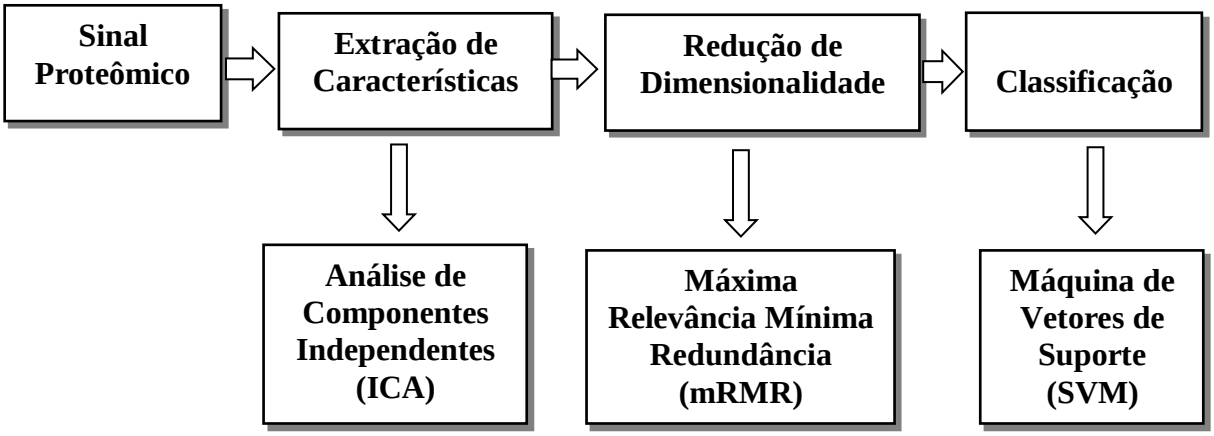

Figura 9 - Diagrama em blocos do método proposto

4.1 *Database*

Os dados adquiridos neste trabalho foram baseados em padrões proteômicos usando a técnica SELDI-TOF. 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 casos com diagnóstico maligno, 100 casos com diagnóstico normal

(controle) e 16 casos com diagnóstico benigno. A segunda base de dados consiste em 162 casos com diagnóstico maligno e 91 casos com diagnóstico normal, totalizando 253 casos. Cada caso consiste em 15.154 níveis de intensidade ou características diferentes. Ambas as bases de dados contêm amostras em baixa resolução. Para este trabalho foi utilizada a segunda base, que contêm a maior quantidade de amostras.

A Figura 10 ilustra um caso com várias amostras que foram extraídas através de um espectrômetro de massa e que foi posteriormente convertida em um sinal multinível através dos níveis de intensidade proteômicos encontrados nas amostras.

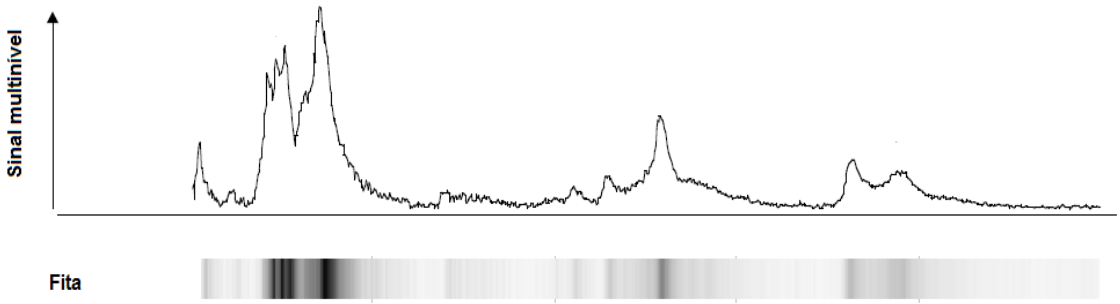

Figura 10 - Relação entre sinal proteômico e níveis de intensidade

Com os dados obtidos, pode-se então passar para a próxima fase deste trabalho, que é a extração de características.

4.2 Extração de Características

Para esta etapa de extração de característica será utilizada a técnica de Análise de Componentes Independentes, vista na Seção 2.5. Com o objetivo de obter a matriz X do modelo ICA, gerada pela união da matriz dos casos com câncer, de dimensões 162x15.154, com a matriz das amostras de casos normais, de dimensões 91x15.154, formando assim a matriz X de dimensões 253x15.154. As linhas da matrix X representam cada um dos casos observados na base de dados, e as colunas as características de cada caso.

A matriz X gerada servirá de entrada para o algoritmo FastICA [Marchini, Heaton and Ripley 2004], para que possam ser obtidas 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].

4.3 Seleção das características mais significantes

O objetivo desta etapa é selecionar as características que melhor represente os dados gerados a partir da etapa de extração de características, ou seja, as características mais significantes. Caso todos os dados gerados sejam utilizados como entrada para um classificador, o resultado poderá ser insatisfatório, com baixa acurácia e grande esforço computacional. Nesta etapa, será utilizada a técnica de seleção de características por Máxima Relevância e Mínima Redundância (mRMR), vista na Seção 2.6.1.

4.4 Classificação

Durante a etapa de classificação será utilizada a Máquina de Vetores de Suporte (SVM), que analisará a matriz de características, já reduzida através da técnica de mRMR, e será atribuída uma classificação, ou seja, vai rotular como câncer ou normal.

Para testar a confiabilidade do método e do classificador, será 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 a classificação usando o subconjunto restante. As fases de treino e teste são depois repetidas 10 vezes, permutando-se circularmente os subconjuntos. A acurácia final é calculada usando a média das acurácias de cada fase.

4.5 Avaliação do Método de Classificação

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, respectivamente, 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).

A Acurácia (*A*) é a taxa de acerto do classificador durante a fase de teste, e é definida por:

$$A = \frac{VP+VN}{VP+VN+FP+FN} \tag{2.42}$$

A Especificidade (*E*) é a proporção de verdadeiros negativos que são corretamente classificados pelo teste, e é definida por:

$$E = \frac{VN}{VN+FP} \tag{2.43}$$

A Sensibilidade (*S*) é a proporção de verdadeiros positivos que são corretamente classificados pelo teste, e é definida por:

$$S = \frac{VP}{VP+FN} \tag{2.44}$$

Área sob a Curva ROC (AuC) é uma forma de representar graficamente a relação entre a sensibilidade e especificidade. Os valores da sensibilidade são representados no eixo y e os valores de (1 – especificidade) no eixo x, no plano cartesiano. A confiabilidade do método é avaliada de acordo com a área abaixo da curva ROC, quanto mais próxima de 1 há uma relação entre sensibilidade e especificidade muito próxima de 100%.

5 RESULTADOS E DISCUSSÕES

Este Capítulo apresenta e discute os resultados obtidos nas abordagens utilizadas. 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. Foi criado um projeto com um arquivo principal (ANEXO A), que contêm as chamadas para as demais funções do trabalho. Vários testes foram realizados para avaliar o método proposto e serão mostrados nas Seções deste Capítulo.

5.1 Utilização da Base de Dados

Para este trabalho foi utilizada a segunda base de dados de *serum* SELDI MS [Seldi–MS 2002], mostrada na Seção 4.1, que possui maior número de amostras (*Ovarian Dataset* 8-7-02), totalizando 253 casos. Os arquivos estão no formato ".csv" (valores separados por vírgula), separados em dois diretórios: *Control*, com 91 arquivos e *Ovarian Cancer*, com 162 arquivos.

Em cada arquivo há duas informações: m/z (relação entre a massa de um determinado íon e o número de cargas elementares que ele carrega) e intensidade (intensidade do sinal do íon), estas informações estão separadas por vírgula. A Tabela 1 mostra um exemplo dos valores m/z que foram utilizados neste trabalho, com suas respectivas intensidades, para uma parte de um caso com câncer (*Ovarian Cancer* daf-0601).

Tabela 1 – Valores m/z com as respectivas intensidades

m/z	Intensidade
-0,000078602611	4,100553
0,000000217736	4,1206637
0,000096021472	4,0361991
0,000366013820	4,1246858
0,000810194770	4,0261438
0,001428564300	3,9457014
0,002221122500	3,8793363
0,003187869300	3,9859226
0,004328804700	4,0160885

No diretório *Control*, os arquivos têm a nomenclatura "*Control* daf-0181.csv", "*Control* daf-0182.csv", e assim sucessivamente. Na Tabela 2 são mostrados dois exemplos com os dez primeiros níveis de intensidade dos 15.154 pontos de cada arquivo com diagnóstico Normal (grupo controle).

Tabela 2 – Dez amostras de dois casos do grupo Controle (Normal)

***Control* daf-0181**	***Control* daf-0182**
M/Z,*Intensity*	M/Z,*Intensity*
-7.8602611e-005,4.1689291	-7.8602611e-005,4.164907
2.1773576e-007,4.13273	2.1773576e-007,4.1106083
9.6021472e-005,4.0965309	9.6021472e-005,4.0502765
0.00036601382,4.1548517	0.00036601382,4.1126194
0.00081019477,4.0542986	0.00081019477,4.0201106
0.0014285643,3.9718451	0.0014285643,3.9457014
0.0022211225,3.9195576	0.0022211225,3.9034691
0.0031878693,4	0.0031878693,4
0.0043288047,4.034188	0.0043288047,3.9939668
0.0056439287,4.0402212	0.0056439287,4.0281549

No diretório *Ovarian Cancer* os arquivos têm a nomenclatura "*Ovarian Cancer* daf-0601.csv", "*Ovarian Cancer* daf-0602.csv", e assim sucessivamente. Na Tabela 3 são mostrados dois exemplos com as dez primeiras amostras das 15.154 amostras de cada arquivo com diagnóstico de Câncer.

Tabela 3 – Dez amostras de dois casos com Câncer

***Ovarian Cancer* daf-0601**	***Ovarian Cancer* daf-0602**
M/Z,*Intensity*	M/Z,*Intensity*
-7.8602611e-005,4.100553	-7.8602611e-005,4.1045752
2.1773576e-007,4.1206637	2.1773576e-007,4.1106083
9.6021472e-005,4.0361991	9.6021472e-005,4.0542986
0.00036601382,4.1246858	0.00036601382,4.1206637
0.00081019477,4.0261438	0.00081019477,4.0221217
0.0014285643,3.9457014	0.0014285643,3.9537456
0.0022211225,3.8793363	0.0022211225,3.8833585
0.0031878693,3.9859226	0.0031878693,3.9698341
0.0043288047,4.0160885	0.0043288047,4.0180995
0.0056439287,4.0040221	0.0056439287,4.0100553

A partir dos dados originais, foi criada uma função no Matlab para ler cada arquivo e criar um *dataset* consolidado para os casos de câncer (ANEXO B) e um *dataset*

para os casos normais (ANEXO C), cada *dataset* contém os valores das intensidades de cada caso.

A Tabela 4 mostra uma parte do *dataset* de câncer com 5 casos dos 162, e com 6 amostras das 15.154 possíveis.

Tabela 4 – Intensidades consolidadas de 6 amostras de 5 casos com câncer

Caso	1ª amostra	2ª amostra	3ª amostra	4ª amostra	5ª amostra	6ª amostra
daf-0601	4,100553	4,1206637	4,0361991	4,1246858	4,0261438	3,9457014
daf-0602	4,1045752	4,1106083	4,0542986	4,1206637	4,0221217	3,9537456
daf-0603	4,0904977	4,1367521	4,0744093	4,1106083	4,0281549	3,9557567
daf-0604	4,1588738	4,1447964	4,0764203	4,1226747	4,0623429	3,9457014
daf-0605	4,1588738	4,0864756	4,0482655	4,1528406	4,0100553	3,9939668

A Tabela 5 mostra uma parte do *dataset* de casos normais com 5 casos dos 91, e com 6 amostras das 15.154 possíveis.

Tabela 5 – Intensidades consolidadas de 6 amostras de 5 casos normais

Caso	1ª amostra	2ª amostra	3ª amostra	4ª amostra	5ª amostra	6ª amostra
daf-0181	4,1689291	4,13273	4,0965309	4,1548517	4,0542986	3,9718451
daf-0182	4,164907	4,1106083	4,0502765	4,1126194	4,0201106	3,9457014
daf-0183	4,1689291	4,0784314	4,034188	4,1568627	4,0542986	3,9356461
daf-0184	4,164907	4,1126194	4,0462544	4,0844646	4,0040221	3,9758673
daf-0185	4,0723982	4,1106083	4,0281549	4,1126194	4,0301659	3,933635

5.2 Extração de Características

Para a Extração de Características foi utilizado o algoritmo *FastICA*, visto na Seção 2.5.5. A fim de se obter a matrix X, foi criada uma função (ANEXO D) que teve como objetivo fazer a união dos casos com câncer e casos normais. Portanto a matriz X criada tem dimensões (253x15.154), onde 253 é quantidade de casos da base de dados, e 15.154 é a quantidade de características por cada caso.

A matriz **X** foi passada como parâmetro para a função que contém o algoritmo *FastICA* (ANEXO E), e assim obteve-se as funções de base da matriz **A**, de dimensão 253x253, que contém as características de cada amostra. A Tabela 6 mostra uma parte da matriz **A**, com oito amostras e cinco características por amostra.

Tabela 6 – Matriz A (parcial) gerada pelo algoritmo FastICA

Amostras	Características				
1ª amostra	1,004886196	0,628938195	0,005186679	-0,93111101	-0,378550101
2ª amostra	0,798078621	0,293951586	-0,117804476	-1,024489223	-0,349740406
3ª amostra	1,085683242	0,382702841	-0,21438754	-1,103242816	-0,267920135
4ª amostra	1,225946425	0,385058947	-0,142486437	-1,216353246	-0,397192848
5ª amostra	1,172074919	0,401461135	-0,156670951	-1,164665919	-0,409047817
6ª amostra	1,069085326	0,346334549	0,010742867	-1,117032982	-0,403216429
7ª amostra	0,944657619	0,536006482	-0,137790179	-0,937357256	-0,390492232
8ª amostra	1,078704237	0,291422991	-0,000336939	-0,906811819	-0,391242357

5.3 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, pois percebeu-se que a partir não havia alterações significativas nos resultados do classificador.

Cada vetor gerado (ANEXO F) foi passado e testado com o classificador de Máquina de Vetores de Suporte (SVM), a fim de encontrar o vetor de melhor desempenho. A Tabela 7 mostra uma parte da matriz de características mais significantes, com dez amostras e cinco características por amostra, gerada pela técnica de mRMR.

Tabela 7 – Matriz de características mais significantes (parcial) gerada pela mRMR

Amostras	Características mais significantes				
1ª amostra	0,20584484	-0,40272994	0,46221081	0,338554941	0,422678294
2ª amostra	0,255166026	-0,56840057	0,102439095	0,28119548	0,682249085
3ª amostra	0,328797947	-0,40523844	0,32516486	0,197551299	0,437463449
4ª amostra	0,360476199	-0,51945597	0,564986064	0,494627569	0,651765187
5ª amostra	0,330629238	-0,3980311	0,351143532	0,48207504	0,434464209
6ª amostra	0,285169521	-0,43399617	0,367767455	0,576012086	0,372107243
7ª amostra	0,138456045	-0,38048075	0,362037458	0,382373615	0,473707087
8ª amostra	0,18802192	-0,37957872	0,410903582	0,515324031	0,693300246
9ª amostra	0,295223894	-0,49311045	0,448460742	0,566933745	0,998884572
10ª amostra	0,239449306	-0,62558286	0,435696191	0,291872007	0,519407843

5.4 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 (ANEXO G).

A Tabela 8 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 8 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 8 - 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 casos com câncer, todos foram classificados 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 área sob a curva ROC (AuC) obtida foi de 0,978, o que demonstra que o classificador atingiu um bom desempenho, pois se aproximou-se de 1, conforme visto na Seção 4.5. A Figura 11 mostra a curva ROC para o vetor de 10 características.

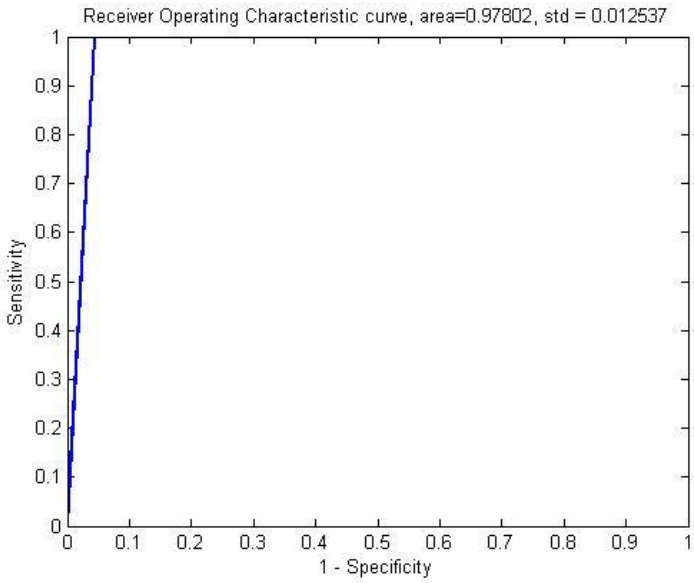

Figura 11 – Área sob a curva ROC para o vetor de 10 características

Outra informação relevante do método proposto é o tempo de processamento das funções e algoritmos que foi de 2,80 segundos, considerando somente o vetor de melhor desempenho (10 características).

6 CONCLUSÃO

Neste trabalho, foi desenvolvido um método de detecção de câncer de ovário, utilizando padrões proteômicos, Análise de Componentes Independentes para extração de características, Máquina de Vetores de Suporte para a classificação das amostras entre câncer ou normalidade.

A extração de características utilizando ICA demonstrou ser efetiva em relação aos sinais extraídos do espectrômetro de massa pela técnica SELDI-TOF, com a utilização do algoritmo FastICA.

A técnica de mRMR mostrou que a redução da dimensionalidade não afetou negativamente os resultados e ainda diminuiu o esforço computacional, mesmo excluindo algumas características.

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 no Capítulo 5 demonstraram que o método proposto alcançou um bom desempenho em relação a outros trabalhos encontrados na literatura. 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.

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 alguns anos poderá se tornar mais acessível, e até que isso seja concretizado deve se ter um *software* completo para auxiliar no diagnóstico de câncer.

O trabalho aqui descrito foi submetido, sob forma de artigo, para o XIV Workshop de Informática Médica, Brasília-DF, no XXXIV Congresso da Sociedade Brasileira de Computação - CSBC 2014, e foi aceito para a publicação, e posteriormente apresentado à comunidade acadêmica, em 29/07/2014.

Com as técnicas apresentadas com o método proposto neste trabalho obtiveram-se bons resultados. Desta maneira, algumas propostas e técnicas poderão ser implementadas, como trabalhos futuros, tais como: testes em bases de dados mais complexas, verificação da efetividade do método proposto em outros tipos de câncer, comparar com outros métodos de redução de dimensionalidade, e Desenvolvimento de um *software* que possa ser testado em hospitais e clínicas, auxiliando na redução da mortalidade para este e outros tipos de câncer.

REFERÊNCIAS

ACS (2014) “Ovarian Cancer Key Statistics”. American Cancer Society [Online]. Ovarian Cancer. Disponível em: http://www.cancer.org/cancer/ovariancancer/detailedguide/ovarian-cancer-what-is-cancer, acesso em 20/06/2014.

ACS-detection (2014) “Ovarian Cancer Detection”. American Cancer Society [Online]. Ovarian Cancer. Disponível em: http://www.cancer.org/cancer/ovariancancer/detailedguide/ovarian-cancer-detection, acesso em 20/06/2014.

ACS-statistics (2014) “Ovarian Cancer Key Statistics”. American Cancer Society [Online]. Ovarian Cancer. Disponível em: http://www.cancer.org/cancer/ovariancancer/detailedguide/ovarian-cancer-key-statistics, acesso em 20/06/2014.

ACS-riskfactors (2014) “Ovarian Cancer Risk Factors”. American Cancer Society [Online]. Ovarian Cancer. Disponível em: http://www.cancer.org/cancer/ovariancancer/detailedguide/ovarian-cancer-risk-factors, acesso em 05/07/2014.

ADAM (2014) “Ovarian Cancer”. ADAM Images [Online]. Disponível em: http://www.adamimages.com/Illustration/SearchResult/1/ovarian, acesso em 05/07/2014.

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.

Barbosa, E. B., Vidotto, A., Polachini, G. M., Henrique, T., Marqui, A. B. T. and Tajara, E. H. (2012) “Proteômica: metodologias e aplicações no estudo de doenças humanas”. Elsevier Editora Ltda.

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

Cotter, R. J. (1994). Time-of-flight mass spectrometry. Columbus, OH: American Chemical Society. ISBN 0-8412-3474-4.

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.

Haykin, S. (1999) "Neural Networks: A comprehensive foundation". 2. ed. [S.l.]: Prentice Hall.

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

Hutchens, T. W. and Yip, T. T. (1993) "New desorption strategies for the mass spectrometric analysis of macromolecules." Rapid Commun Mass Spectrom 7: 576-580.

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

Hyvärinen, A., Karhunen, J. and Oja, E. (2001) "Independent Component Analysis". New York: Wiley, 2001.

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.

INCa (2014) [Online] Instituto Nacional do Câncer. Disponível em: http://www2.inca.gov.br/wps/wcm/connect/0129ba0041fbbc01aa4fee936e134226/Apresentacao+Estimativa+2014_final+corrigido+tireoide.pdf?MOD=AJPERES&CACHEID=0129ba0041fbbc01aa4fee936e134226, acesso em 14/06/2014.

Instituto do Câncer (2014) "Câncer de Ovário". Disponível em: http://www.institutodocancer.com.br/php/index.php?link=5&sub=9.

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.

Karas, M., Bachmann, D. and Hillenkamp, F. (1985) "Influence of the Wavelength in High-Irradiance Ultraviolet Laser Desorption Mass Spectrometry of Organic Molecules". Analytical Chemistry 57 (14): 2935–9. doi:10.1021/ac00291a042

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.

May, C., Brosseron, F., Chartowski, P., Schumbrutzki, C., Schoenebeck, B. and Marcus, K. (2011) "Instruments and methods in proteomics". Methods Mol Biol.;696:3-26.

Oncoguia (2012) "Exames de Imagem para o Diagnóstico do Câncer de Ovário". Disponível em: http://www.oncoguia.org.br/conteudo/exames-de-imagem-para-o-diagnostico-do-cancer-de ovario/1785/229/, acesso em 10/06/2014.

Oncoguia (2014) "Tipos de Câncer de Ovário". Disponível em http://www.oncoguia.org.br/conteudo/tipos-de-cancer-de-ovario/1783/228/, acesso em 20/06/2014.

Papoulis, A. and Pillai, S. U. (2002) "Probability, Random Variables and Sthocastic Processes". 4. ed. New York: McGraw-Hill.

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.

Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone,C., Fishman, D. A., Kohn, E. C. and Liotta, L. A. (2002) "Use of proteomic patterns in serum to identify ovarian cancer", *The Lancet*, vol 359, pp 572-577

Petricoin, E. F. and Liotta, L. A. (2004) "Proteomic approaches in cancer risk and response assessment". Elsevier Editora Ltda.

Seldi–MS (2002) DATABASE. Disponível em: http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp, Acesso em Outubro, 2013.

Schölkopf, B. and Smola, A. "Learning with kernels", MIT Press, Cambridge, MA, 2002.

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.

UNIFESP (2014) "Histologia da Ovário" [Online]. Escola Paulista de Medicina. Departamento de Histologia e Biologia Estrutural. Disponível em: http://www.unifesp.br/dmorfo/histologia/ensino/ovario/histologia.htm.

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.

Yu, J. K., Zheng, S., Tang, Y. and Li, L. (2005). “An integrated approach utilizing proteomics and bioinformatics to detect ovarian cancer”, *Jornal of Zhejiang University SCIENCE*, vol. 6, no.4, pp. 227-231, July.

ANEXO A

```
% Universidade Estadual do Maranhão
% Mestrado em Engenharia da Computação e Sistemas
% Aluno: Wesley Batista Dominices de Araujo
% Orientador: Prof. Dr. Lúcio Flávio de A. Campos
% Co-orientadora: Prof. Enfª. Ma. Aline S. Furtado

% OvarianCancer.m

clear;
clc;
close all;

ds1_criado = 0;
gerou_X1 = 0;
gerou_fastica1 = 0;
gerou_mrmr1 = 0;
op = 0;
med = 0;
quantDS = 1;
while op < 3

op = menu('Escolha uma opção','Executar tudo','SAIR');

switch_expr = op;
 switch switch_expr
   case 1, % EXECUTAR TODAS AS ETAPAS
      ds1_criado = 0;
      gerou_X1 = 0;
      gerou_fastica1 = 0;
      gerou_mrmr1 = 0;

      ds = switch_expr;
      %chama a função/script para gerar a matriz com os dados de câncer
      %salva as informações no diretório database com o nome câncer
      [tamDsCa] = gera_db_cancer(ds);

      %chama a função/script para gerar a matriz com os dados de controle
      %salva as informações no diretório database com o nome control
      gera_db_control(ds);

      ds1_criado = 1;

      %inicia o tempo de execução do sistema
      tstart = tic;
      %chama a função/script para gerar a matriz consolidada com os dados gerais
      %salva as informações no diretório database com o nome cancer

      [X,tamX] = carrega_dbs1(ds); %X = [cancer;control];
```

```
        fprintf('salvando a matriz X.....\n\n');
        save('matrizes/X1','X');

        gerou_X1 = 1;
        fprintf('dataset control gerado com sucesso! \n');
        fprintf('dataset cancer gerado com sucesso! \n');
        fprintf('datasets carregados e gerada a matriz X!\n');

%%------------------ FastICA ----------------------%%
        fprintf('Chama o FastICA\n\n');
        [Y,W,A] = aapomagro(X,tamX);
        save('matrizes/A1','A');

        fprintf('FastICA concluído com sucesso\n');
        gerou_fastica1 = 1;

%%------------------ mRMR ----------------------%%
        fprintf('chamando o mRMR\n');

        % GERAR MATRIZ DE RÓTULOS AUTOMATICAMENTE
        f=ones(tamX,1);
        for i=tamDsCa+1:tamX,f(i,1)=-1;end
        save('matrizes/f1','f');

        for k=5:5:150 % nº de características

            [fea] =  mrmr_mid_d(A,f,k);
            gerou_mrmr1 = 1;
            [fs] = feature_matrix(A,fea);
            save('matrizes/fs1','fs');

%%------------------ SVM ----------------------%%
          fprintf('Fazendo a Classificação SVM\n');
          [acu,esp,sen,VP,FP,VN,FN] = fold_sens_espec_cross_validation(fs,f);

          % tempo total gasto para executar todo o programa
          tempototal = toc(tstart);
          fprintf('Tempo Total gasto para o algoritmo: %0.2f  \n \n',tempototal);

           med = med + 1;
           res1(med,1) = [k];
           res1(med,2) = [acu];
           res1(med,3) = [esp];
           res1(med,4) = [sen];
           res1(med,5) = [VP];
           res1(med,6) = [FP];
           res1(med,7) = [VN];
           res1(med,8) = [FN];

           save('resultados/res1','k','acu','esp','sen','VP','FP','VN','FN');
```

```
        save('resultados/res-consolidados1','res1');

        fprintf('CLASSIFICAÇÃO para %d características CONCLUÍDA \n \n',k);
    end

    fprintf('RESULTADOS DATASET 1:\n');

    res1
    fprintf('MELHOR RESULTADO ALCANÇADO PARA O DATASET:\n');
    fprintf('ACURÁCIA: %f\n',max(res1(:,2)));
    fprintf('ESPECIFICIDADE: %f\n',max(res1(:,3)));
    fprintf('SENSIBILIDADE: %f\n',max(res1(:,4)));
    fprintf('VP: %d\n',max(res1(:,5)));
    fprintf('FP: %d\n',min(res1(:,6)));
    fprintf('VN: %d\n',max(res1(:,7)));
    fprintf('FN: %d\n',min(res1(:,8)));


  case 2,
      break;
 end
end
```

ANEXO B

```
% gera_db_cancer.m

function [tamDsCa] = gera_db_cancer(ds)

   close all;

   fprintf('gerando dataset %d de cancer.....\n',ds);
   path = strcat('','ds',int2str(ds),'_cancer/','*.csv','');
   a = dir(path);
   cancer = [];
   for i = 1:length(a)

      temp = importdata(a(i).name);
      cancer = [cancer, temp.data(:,2)];

   end

   cancer=cancer';

   fprintf('salvando dataset %d cancer no diretório databases.....\n',ds);
   save(strcat('databases/cancer',int2str(ds)),'cancer');
   [tamDsCa] = size(cancer,1);

end
```

ANEXO C

```
% gera_db_control.m

function gera_db_control(ds)

   close all;

   fprintf('gerando dataset %d de controle.....\n',ds);
   path = strcat('','ds',int2str(ds),'_control/','*.csv','');
   a = dir(path);
   control = [];
   for i = 1:length(a)

      temp = importdata(a(i).name);
      control = [control, temp.data(:,2)];

   end

   control=control';

   fprintf('salvando dataset %d de controle no diretório databases.....\n',ds);
   save(strcat('databases/control',int2str(ds)),'control');
end
```

ANEXO D

```
% carrega_dbs1.m

function [X,tamX] = carrega_dbs1(ds)

   close all;
   clc;

   %fprintf('carregando dataset %d de cancer.....\n',ds);
   load databases/cancer1;

   %fprintf('carregando dataset %d de controle.....\n',ds);
   load databases/control1;

   %fprintf('gerando a matriz X consolidada do dataset %d.....\n',ds);
   X=[cancer;control];

   %fprintf('Matriz X do dataset %d gerada com sucesso...\n\n',ds);

   tamX = size(X,1);

end
```

ANEXO E

```
% aapomagro.m
function [Y,W,A]= aapomagro(x,pcomponents,no)

% FastICA - PIB

% Originally written by the Finish (Aapo) team;
% Modified by the PIB team in Aug. 21, 2007

fprintf ('Removing mean...\n');

%--------------------------------------------------------------------
%   Meanize
%   Removes the mean of X
%--------------------------------------------------------------------

[nn,M]=size(x);
if nn>M,
  x=x';
  [nn,M]=size(x);
end
X=double(x)-mean(x')'*ones([1,M]);     % Remove the mean.
X1=X;

%---- Meanize end ------

% Calculate the eigenvalues and eigenvectors of covariance matrix.
fprintf ('Calculating covariance...\n');
covarianceMatrix = X*X'/size(X,2);
[E, D] = eig(covarianceMatrix);
% Sort the eigenvalues and select subset, and whiten

%--------------------------------------------------------------------
%                   PCA beggins
%--------------------------------------------------------------------
[dummy,order] = sort(diag(-D));
E = E(:,order(1:pcomponents));
d = diag(D);
d = real(d.^(-0.5));
D = diag(d(order(1:pcomponents)));
X = D*E'*X;

whiteningMatrix = D*E';
dewhiteningMatrix = E*D^(-1);
%--------------------------------------------------------------------
%                   PCA end
%--------------------------------------------------------------------

N = size(X,2);
```

```
B = randn(size(X,1),pcomponents);
B = B*real((B'*B)^(-0.5));              % orthogonalize

W1=randn(size(B' * whiteningMatrix));
W=rand(size(B' * whiteningMatrix));

iter=0;
while abs(norm(W)-norm(W1'))>1e-50,
  iter = iter+1;
  fprintf('(%d)',iter);

  % This is tanh but faster than matlabs own version
  hypTan = 1 - 2./(exp(2*(X'*B))+1);

  % This is the fixed-point step
  B = X*hypTan/N - ones(size(B,1),1)*mean(1-hypTan.^2).*B;

  B = B*real((B'*B)^(-0.5));
  W1=W;
  W = B' * whiteningMatrix;
end
Y=W*X1;
A = dewhiteningMatrix * B;

fprintf(' Done!\n');
```

ANEXO F

```
function [fea] = mrmr_mid_d(d, f, K)
% function [fea] = mrmr_mid_d(d, f, K)
% MID scheme according to MRMR
% By Hanchuan Peng
% April 16, 2003

bdisp=0;
nd = size(d,2);
nc = size(d,1);
t1=cputime;
for i=1:nd,
  t(i) = mutualinfo(d(:,i), f);
end;
fprintf('calculate the marginal dmi costs %5.1fs.\n', cputime-t1);
[tmp, idxs] = sort(-t);
fea_base = idxs(1:K);
fea(1) = idxs(1);
KMAX = min(1000,nd); %500
idxleft = idxs(2:KMAX);
k=1;
if bdisp==1,
fprintf('k=1 cost_time=(N/A) cur_fea=%d #left_cand=%d\n', ...
     fea(k), length(idxleft));
end;

for k=2:K,
  t1=cputime;
  ncand = length(idxleft);
  curlastfea = length(fea);
  for i=1:ncand,
    t_mi(i) = mutualinfo(d(:,idxleft(i)), f);
    mi_array(idxleft(i),curlastfea) = getmultimi(d(:,fea(curlastfea)), d(:,idxleft(i)));
    c_mi(i) = mean(mi_array(idxleft(i), :));
  end;
  [tmp, fea(k)] = max(t_mi(1:ncand) - c_mi(1:ncand));
  tmpidx = fea(k); fea(k) = idxleft(tmpidx); idxleft(tmpidx) = [];
  if bdisp==1,
  fprintf('k=%d cost_time=%5.4f cur_fea=%d #left_cand=%d\n', ...
    k, cputime-t1, fea(k), length(idxleft));
  end;
end;
return;

%====================================
function c = getmultimi(da, dt)
for i=1:size(da,2),
  c(i) = mutualinfo(da(:,i), dt);
end;
```

ANEXO G

```
%feature matrix.m
function [fs] = feature_matrix(d,fea)
   [a1 b1]=size(fea);
   fs=[];
   for i=1:b1
      c1=d(:,fea(i));
      fs=[fs,c1];
   end
end
```

ANEXO H

```
% fold_sens_espec_cross_validation.m

function [acuracia,especificidade,sensibilidade,VP,FP,VN,FN] = fold_sens_espec_cross_validation(fs,f)
 A_10 = fs;
 At = f;
 ker.ker=2;ker.gamma=0;
 tp=[];
 tn=[];
 fp=[];
 fn=[];
 SENS=[];
 ESPEC=[];
 TP = 0;
 TN = 0;
 FP = 0;
 FN = 0;
 ACC=[];
 TMP=[];
 PRED=[];

 % quantidade de divisões
 fold = 10;

 CVO = cvpartition(At,'k',fold);
 err = zeros(CVO.NumTestSets,1);
 for i = 1:CVO.NumTestSets
     trIdx = CVO.training(i);
     teIdx = CVO.test(i);
     [ acc, predict_label ] = svmPrediction( A_10(trIdx,:), At(trIdx,:), A_10(teIdx,:) , At(teIdx,:), ker );
      ACC=[ACC;acc];
      temp = At(teIdx,:);
      for m = 1:length(temp)
          if (predict_label(m) == temp(m)) & (temp(m) == 1)
            TP = TP + 1;
          end
          if (predict_label(m) == temp(m)) & (temp(m) == -1)
              TN = TN + 1;
          end
          if (predict_label(m) ~= temp(m)) & (temp(m) == 1)
               FN = FN + 1;
          end
          if (predict_label(m) ~= temp(m)) & (temp(m) ==-1)
               FP = FP + 1;
          end
       end
```

```
        sens = 100*(TP / (TP+FN));
        espec = 100*(TN / (FP+TN));
        tp=[tp;TP];
        tn=[tn;TN];
        fp=[fp;FP];
        fn=[fn;FN];
        SENS=[SENS;sens];
        ESPEC=[ESPEC;espec];
        TMP=[TMP;temp];
        PRED=[PRED;predict_label];

        err(i) = sum(~strcmp(predict_label,At(teIdx)));
    end
    cvErr = sum(err)/sum(CVO.TestSize);

    acuracia=sum(ACC)/fold;
    sensibilidade=sum(SENS)/fold;
    especificidade=sum(ESPEC)/fold;

    VP=max(tp);
    FP=max(fp);
    VN=max(tn);
    FN=max(fn);

end
```