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

ELZENIR DE ARAUJO MONTES

Aplicação de um Método Computacional no Diagnóstico Precoce do Câncer de Próstata Baseado em Reconhecimento de Padrões Proteômicos.

São Luís
2017

ELZENIR DE ARAUJO MONTES

Aplicação de um Método Computacional no Diagnóstico Precoce do Câncer de Próstata Baseado em Reconhecimento de Padrões Proteômicos.

Dissertação apresentada ao programa do 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 Sistema.

Orientador: Prof. Dr. Lúcio Flávio de A. Campos

São Luís
2017

BANCA EXAMINADORA

Prof. Dr. LÚCIO FLÁVIO DE ALBUQUERQUE CAMPOS
Universidade Estadual do Maranhão – UEMA
Orientador

Prof. Msc. Antônio Fernando Lavareda Jacob Junior
Universidade Estadual do Maranhão – UEMA
Membro

Prof. Msc. Aline Santos Furtado Campos
Universidade Federal do Maranhão – UFMA
Membro

AGRADECIMENTOS

A Deus, pelo dom da vida e me permitir gozar dela com saúde e disposição para seguir em frente em meio a tantas dificuldades.

Aos meus pais, por todos os esforços e sacrifícios feito ao longo de minha vida.

Ao meu orientador, profº Dr. Lúcio Flávio de Albuquerque Campos, pela ajuda, dedicação, ensinamentos e principalmente pela paciência durante todo o período de desenvolvimento deste trabalho.

Aos meus amigos da turma 2014.2 do Mestrado em Engenharia da Computação e Sistemas da UEMA, pelos diversos momentos de descontração, apoio e confraternização.

Aos amigos de trabalho, que contribuíram para o fim desta etapa, em especial Samara Araújo, pela paciência, correções e apontamentos, e Conceição Alencar pelo apoio e momentos de laser.

Aos professores, coordenação e secretaria do Mestrado em Engenharia da Computação e Sistema da UEMA, pela ajuda e paciência em diversas ocasiões.

A Jardiel Nunes Almeida e José dos Nascimento Linhares pela ajuda e apoio no decorrer do curso.

A FAPEMA pelo apoio financeiro neste período de curso.

RESUMO

O câncer de próstata na sua fase inicial possui uma evolução bastante sigilosa, tanto que a maioria dos pacientes não apresentam sintomas, e quando os mesmo aparecem são confundidos com o crescimento benigno da próstata. Neste trabalho propõe-se a aplicação de um conjunto de técnicas computacionais para compor um novo método de diagnóstico precoce do câncer de próstata, baseado em reconhecimento de padrões proteômicos. O método possui basicamente três etapas. A primeira etapa é realizada pela técnica de Análise de Componentes Independentes (ICA) através do algoritmo FastICA, com o objetivo de extrair as características dos sinais proteômicos. A segunda etapa objetivando diminuir o conjunto de características, e com isso o custo computacional utilizou-se a técnica Máxima Relevância e Mínima Redundância (mRMR). Na terceira etapa utilizou-se dois classificadores de modo a comparar os resultados entre eles e decidir pelo melhor conjunto de técnicas a serem empregadas no diagnóstico precoce do câncer de próstata, Máquina de Vetores de Suporte (SVM) e a Análise Discriminante Linear (LDA). Assim, os resultados obtidos com conjunto de técnicas (ICA $\Longrightarrow$ mRMR $\Longrightarrow$ SVM) foram satisfatório, mas foi fazendo uso do conjunto (ICA $\Longrightarrow$ mRMR $\Longrightarrow$ LDA) que os melhores resultados foram alcançados a partir de um vetor de 77 características, o classificador LDA obteve uma ótima resposta na fase de classificação, obtendo acurácia, especificidade e sensibilidade respectivamente de 100%, 100% e 100%.

Palavras Chaves: Sinais Proteômico, Análise de Componentes Independentes; Máxima Relevância e Mínima Redundância, Máquina de Vetores de Suporte, Análise Discriminante Linear.

ABSTRACT

Prostate câncer in its early stages has a fairly steady evolution, so much so that most patients do not show symptoms, and when they appear they are confused with benign prostate growth. This work proposes the application of a set of computational techniques to compose a new method of early diagnosis of prostate cancer, based on recognition of proteomic patterns. The method has basically three steps. The first step is performed by the Independent Component Analysis (ICA) technique through the FastICA algorithm, in order to extract the characteristics of the proteomic signals. The second step aimed at reducing the set of characteristics, and with this the computational cost was used the technique Maximum Relevance and Minimum Redundancy (mRMR). In the third step, two classifiers were used to compare the results between them and to decide on the best set of techniques to be used in the early diagnosis of prostate cancer, Supporting Vector Machine (SVM) and Linear Discriminant Analysis (LDA) . Thus, the results obtained with set of techniques (ICA $\Longrightarrow$ mRMR $\Longrightarrow$ SVM) were satisfactory, but it was making use of the set (ICA $\Longrightarrow$ mRMR $\Longrightarrow$ LDA) that the best results were achieved From a vector of 77 characteristics, the LDA classifier obtained an excellent response in the classification phase, obtaining accuracy, specificity and sensitivity respectively of 100 %, 100 % and 100 %.

Keywords: Proteomic Signals, Independent Component Analysis; Maximum Relevance and Minimum Redundancy, Supporting Vector Machine, Linear Discriminant Analysis

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

O câncer de próstata é a neoplasia maligna visceral mais comum no homem excetuando se os tumores cutâneos e, a incidência tende a crescer nas próximas décadas com o aumento da expectativa de vida. O risco de desenvolvimento da doença durante a vida é de 17,6% para homens brancos e de 20,6% para homens negros. Aproximadamente 543 mil casos novos são diagnosticados por ano no mundo [1].

Foram esperados, para 2012, de acordo com a última estimativa mundial, cerca de 1 milhão de casos novos para essa neoplasia. Aproximadamente 70% dos casos diagnosticados ocorrem em regiões mais desenvolvidas [2]. O alto índice de incidência dessa patologia em regiões bem desenvolvidas pode se justificar de certo modo [2] pelas práticas de rastreamento por meio dos teste do antígeno prostático específico (PSA).

O Brasil vem passando, nas últimas décadas, por alterações de contexto social,econômico e, consequentemente, de saúde. O aumento da expectativa de vida, a melhoria e a evolução dos métodos diagnósticos podem explicar o crescimento das taxas de incidência ao longo dos anos no país. Além disso, a melhoria da qualidade dos sistemas de informação do país e a ocorrência de sobrediagnóstico, em função da disseminação do rastreamento do câncer de próstata com PSA e toque retal, também influenciam na magnitude da doença [2].

O diagnóstico, tratamento e morbimortalidade do câncer de próstata tiveram sua história modificada após a introdução da dosagem do PSA (Antígeno Prostático Específico) na prática clínica. Na década de 1990 houve uma espécie de "explosão"no diagnóstico do câncer de próstata, sendo atualmente o câncer mais prevalente na população [1].

Esse câncer é o segundo tipo em mortalidade no Brasil. Em 2012, foram registrados 13.354 óbitos pela doença, o que representa 13,1% dos óbitos por câncer em homens [3]. As altas taxas de incidência e mortalidade do câncer de próstata provocam um debate mundial entre diversos segmentos da sociedade a respeito das estratégias de controle, que continuam sendo um desafio e um campo de interesses conflitantes, na medida em que não é possível determinar com precisão qual será a evolução dos casos detectados no rastreamento.

As estimativas que se tem para o país no ano de 2016 são preocupantes. Estimam-se 61.200 casos novos de câncer de próstata para o Brasil em 2016. Esses valores correspondem a um risco estimado de 61,82 casos novos a cada 100 mil homens [2].

No Maranhão, a estimativa do câncer de próstata para o ano de 2016 é de 1050 casos, distribuídos da seguinte forma: 210 casos na capital, São Luís, e 840 no interior do estado [2].

A figura 1 ilustra a estimativa do câncer de próstata no Brasil para o ano de 2016, somente em homens [2].

Localização Primária	Casos	%	
Próstata	61.200	28,6%	Homens
Traqueia, Brônquio e Pulmão	17.330	8,1%	
Cólon e Reto	16.660	7,8%	
Estômago	12.920	6,0%	
Cavidade Oral	11.140	5,2%	
Esôfago	7.950	3,7%	
Bexiga	7.200	3,4%	
Laringe	6.360	3,0%	
Leucemias	5.540	2,6%	
Sistema Nervoso Central	5.440	2,5%	

Figura 1: Estimativa de Câncer em homens para o ano 2016.
Fonte:[2] .

É fundamental que o monitoramento da morbimortalidade por câncer incorpore-se na rotina da gestão da saúde de modo a tonar-se instrumento essencial para o estabelecimento de ações de prevenção e controle do câncer e de seus fatores de risco. Esse monitoramento engloba a supervisão e a avaliação de programas, como ações necessárias para o conhecimento da situação e do impacto no perfil de morbimortalidade da população, bem como a manutenção de um sistema de vigilância com informações oportunas e de qualidade que subsidie análises epidemiológicas para as tomadas de decisões [2].

Embora essa doença seja passível de detecção precoce, muitas vezes deixa de ser identificada e, quando ocorre o diagnóstico, já se encontra em estágios avançados, comprometendo o seu prognóstico [4]. De acordo com [5], um em cada seis homens com idade acima de 45 anos pode ter a doença sem que conheça o diagnóstico. A alta frequência, que faz do câncer de próstata um problema de saúde pública, aliada à possibilidade de detecção, através de procedimentos relativamente simples, deveria fazer desta doença uma prioridade na atenção à saúde masculina [5].

Os maiores fatores de risco identificados para o câncer de próstata são: idade, história familiar de câncer e etnia/cor da pele. Entretanto, a idade é o único fator de risco bem estabelecido para o desenvolvimento do câncer de próstata. A maioria dos cânceres de próstata é diagnosticada em homens acima dos 65 anos, sendo que somente menos de 1% é diagnosticado em homens abaixo dos 50 anos. Com o aumento da expectativa de vida mundial, é esperado que o número de casos novos de câncer de próstata aumente cerca de 60% [2].

A hereditariedade também é considerado como fator de risco pois se existir casos na

família de câncer, como pai ou irmão, diagnosticados de forma previa com esta patologia aumenta-se em duas a três vezes o risco de desenvolver essa neoplasia (aproximadamente 11 vezes) se o diagnóstico do pai ou do irmão tiver ocorrido antes dos 40 anos [2].

Tem-se ainda como fator de risco a etnia/cor da pele, pois o câncer de próstata é 1,6 vezes mais comum em homens negros quando comparados aos homens brancos. Apesar disso, é possível que essa diferença entre negros e brancos se dê em função do estilo de vida ou dos fatores associados à detecção da doença. [2].

O câncer de próstata na sua fase inicial tem evolução silenciosa. Muitos pacientes não apresentam nenhum sintoma ou, quando apresentam, são semelhantes aos do crescimento benigno da próstata (dificuldade de urinar, necessidade de urinar mais vezes durante o dia ou a noite). Na fase avançada, pode provocar dor óssea, sintomas urinários ou, quando mais grave, infecção generalizada ou insuficiência renal [6]. O que a grava a situação do homem quando apresenta esta doença pois, segundo [7], a presença de homens nos serviços de atenção primária à saúde é menor do que a das mulheres. Há autores que associam esse fato à própria socialização dos homens, em que o cuidado não é visto como uma prática masculina.

Achados no exame clínico (toque retal) combinados com o resultado da dosagem do antígeno prostático específico (PSA, na sigla em inglês) no sangue podem sugerir a existência da doença. Nesses casos, é indicada a ultrassonografia pélvica (ou prostática transretal, se disponível). O resultado da ultrassonografia, por sua vez, poderá mostrar a necessidade de biópsia prostática transretal. O diagnóstico de certeza do câncer é feito pelo estudo histopatológico do tecido obtido pela biópsia da próstata [6]. A própria biópsia prostática por si já é um procedimento invasivo e com índice de complicações importante, principalmente nos pacientes diabéticos e com história de infecção urinária de repetição e prostatites [8].

A escolha do tratamento ideal para cada paciente depende de alguns fatores. Existem várias opções para o tratamento do câncer de próstata, que devem visar não somente ao controle oncológico da doença, mas também a manutenção da qualidade de vida [8]. A escolha do tratamento mais adequado deve ser individualizada e definida após discutir os riscos e benefícios do tratamento com o seu médico [6].

Para doença localizada, cirurgia, radioterapia e até mesmo observação vigilante (em algumas situações especiais) podem ser oferecidos. Para doença localmente avançada, radioterapia ou cirurgia em combinação com tratamento hormonal têm sido utilizados. Para doença metastática (quando o tumor original já se espalhou para outras partes do corpo), o tratamento de eleição é a terapia hormonal[9].

O presente trabalho propõe um método de diagnóstico auxiliado por computador, também conhecido pela sigla CAD (computer-aided diagnosis) para o diagnóstico precoce do câncer de próstata, baseado em reconhecimento de padrões proteômico.

O Reconhecimento de Padrões tem como objetivos atribuir um padrão a um con-

junto desconhecido de objetos (clustering) ou identificar um padrão como membro de um conjunto conhecido de classes (classificação).

Um método baseado em reconhecimento de padrões engloba três etapas principais e neste trabalho elas são realizadas da seguinte forma: primeira etapa é a extração de características dos sinais proteômicos através da Análise de Componentes Independentes (ICA), em seguida é utilizado o algoritmo de Máxima Relevância e Mínima Redundância (mRMR), para selecionar as características mais relevantes e reduzir a dimensionalidade da matriz de características gerada, assim diminui o custo computacional e o tempo de processamento. Na terceira etapa é feita a classificação dos pacientes em dois grupos, controle (não portador da patologia do câncer de próstata) e o grupo ativo (portador da patologia câncer de próstata) essa classificação é feita pelos classificadores Máquina de Vetores de Suporte (SVM) e Análise Discriminante Linear (LDA).

1.1 Organização do Trabalho

Este trabalho está divido em seis capítulos, que serão apresentados da seguinte forma.

O capítulo 2 aborda uma revisão teórica da literatura necessária para desenvolvimento do método proposto, apresenta alguns conceitos sobre câncer de próstata, reconhecimento de padrões, proteômica, a técnica Analise de Componentes Independentes, o algorítimo Máxima Relevância e Mínima Redundância, e a Máquina de Vetores de Suporte.

O capítulo 3 apresenta o objetivo geral e os objetivos específicos do presente trabalho.

O capítulo 4 descreve-se sobre a base de dados utilizada e a metodologia proposta dividida em: extração e seleção de características e a classificação nos grupos controle ou ativo, juntamente com os materiais utilizados em cada etapa deste trabalho .

O capítulo 5 apresenta os resultados obtidos, discussões, análise das técnicas utilizadas em cada etapa do trabalho e a avaliação do método pelas medidas mais utilizadas para descrever um sistema de diagnóstico: Acurácia, Sensibilidade, Especificidade.

O capítulo 6 apresenta as conclusões sobre o trabalho, mostrando a eficiência do método proposto e sugestões para futuros trabalhos.

Este trabalho deu origem a dois artigos, o primeiro que foi escrito apresentando apenas o conjunto de técnicas (ICA, mRMR, SVM) e seus resultados e o mesmo foi submetido à "Revista de Ciências da Computação"(RCC) que respondeu com um aceite do trabalho e solicitando umas pequenas alterações. O segundo artigo trata do trabalho completo usando todos os conjuntos de técnicas juntamente com os resultados obtidos e foi submetido à Revista de Sistemas e Computação (RSC) e ainda esta em processo de avaliação.

2 Fundamentação Teórica

2.1 O câncer

Câncer é o nome dado a um conjunto de mais de 100 doenças que têm em comum o crescimento desordenado (maligno) de células que invadem os tecidos e órgãos, podendo espalhar-se (metástase) para outras regiões do corpo [9].

Segundo [10], O crescimento das células cancerosas é diferente do crescimento das células normais. As células cancerosas, em vez de morrerem, continuam crescendo incontrolavelmente, formando outras novas células anormais, que se dividem de forma rápida, agressiva e incontrolável, espalhando-se para outras regiões do corpo acarretando transtornos funcionais.

O envelhecimento traz mudanças nas células que aumentam a sua suscetibilidade à transformação maligna. Isso, somado ao fato de as células das pessoas idosas terem sido expostas por mais tempo aos diferentes fatores de risco para câncer, explica em parte o porquê de o câncer ser mais frequente nesses indivíduos.Os fatores de risco ambientais de câncer são denominados cancerígenos ou carcinógenos. Esses fatores atuam alterando a estrutura genética (DNA) das células [9]. A figura 2 mostra o crescimento celular descontrolado das células gerando um câncer invasivo.

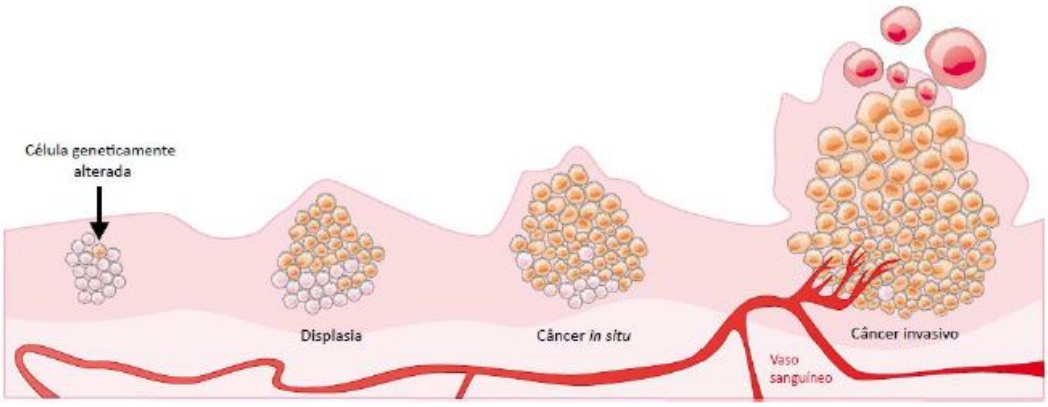


Figura 2: Crescimento Celular descontrolado.
Fonte:[10] .

Nem todos os tumores são câncer. Os tumores que não são cancerosos são denominados benignos. Os tumores benignos podem causar problemas, como crescerem em demasia e pressionarem outros órgãos e tecidos saudáveis. Mas eles não podem invadir outros tecidos e órgãos. Dessa forma, eles não podem se espalhar para outras partes do corpo (metástase)[11].

As neoplasias malignas ou tumores malignos manifestam um maior grau de autonomia e são capazes de invadir tecidos vizinhos e provocar metástases, podendo ser resistentes ao tratamento e causar a morte. A Figura 3 ilustra a diferença entre tumor benigno e

maligno [10]

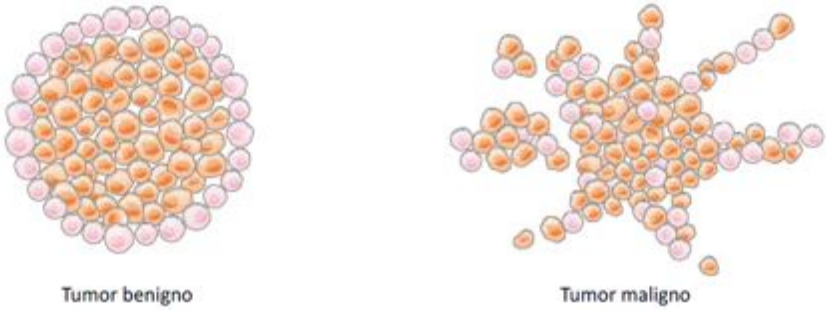


Figura 3: Diferença entre o tumor maligno e benigno.
Fonte:[10] .

Os fatores de risco de câncer podem ser encontrados no meio ambiente ou podem ser herdados. A maioria dos casos de câncer (80%) está relacionada ao meio ambiente, no qual encontramos um grande número de fatores de risco. Entende-se por ambiente o meio em geral (água, terra e ar), o ambiente ocupacional (indústrias químicas e afins) o ambiente de consumo (alimentos, medicamentos) o ambiente social e cultural (estilo e hábitos de vida) [9].

Já está comprovado que uma dieta rica em frutas, verduras, legumes, grãos e cereais integrais, e com menos gordura, principalmente as de origem animal, ajuda a diminuir o risco de câncer, como também de outras doenças crônicas não-transmissíveis. Nesse sentido, outros hábitos saudáveis também são recomendados, como fazer, no mínimo, 30 minutos diários de atividade física, manter o peso adequado à altura, diminuir o consumo de álcool e não fumar [6].

2.2 A próstata

A próstata é uma glândula que só o homem possui e que se localiza na parte baixa do abdômen. Ela é um órgão muito pequeno, tem a forma de maçã e se situa logo abaixo da bexiga e à frente do reto. A próstata envolve a porção inicial da uretra, tubo pelo qual a urina armazenada na bexiga é eliminada.A próstata produz parte do sêmen, líquido espesso que contém os espermatozoides, liberado durante o ato sexual [6]. A figura 4 mostra a localização da próstata no homem.

2.3 O câncer de próstata

Como todos os outros tecidos e órgãos, a próstata é composta por células, que normalmente se dividem e se reproduzem de forma ordenada e controlada, no entanto, quando ocorre uma disfunção celular que altere este processo de divisão e reprodução, produz-se um excesso de tecido, que dá origem ao tumor, podendo este ser classificado como benigno,

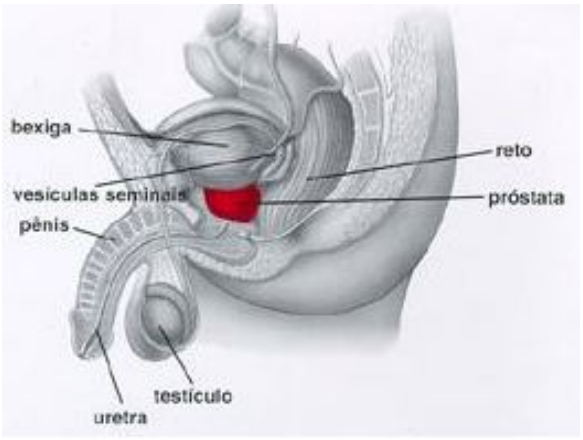


Figura 4: Localização da próstata no homem.
Fonte:[6] .

chamado de hiperplasia prostática benigna (HPB), e o maligno, denominado câncer de próstata (CP), podendo este último sugir associado ou não ao crescimento benigno [12].

O Câncer de próstata se origina, na maior parte das vezes, na zona periférica da próstata e por este motivo raramente causa sintomas precocemente. O screening do câncer de próstata tem o objetivo de possibilitar a detecção precoce e a redução da mortalidade desta doença, porém há uma preocupação cada vez maior com a morbidade associada às diversas formas de tratamento [8].Costuma ser assintomático até fases avançadas. A evolução dos pacientes é relativamente imprevisível, com casos de rápida disseminação, antes mesmo de surgirem os sintomas locais, e casos de evolução lenta e indolente (as células tumorais levam em média 15 anos para crescerem $1cm^3$) [1].

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 são garantias da presença de uma doença e nem mesmo é possível predizer a influência de um determinado fator de risco na presença da doença, como o câncer por exemplo.

Os maiores fatores de risco identificados para o câncer de próstata são: idade, história familiar de câncer e etnia/cor da pele. Entretanto, a idade é o único fator de risco bem estabelecido para o desenvolvimento do câncer de próstata [2].

- histórico familiar, se um parente de primeiro grau tem a doença, o risco é, no mínimo, duas vezes maior do indivíduo ter câncer de próstata. Se dois ou mais indivíduos da mesma família são afetados, o risco aumenta em 5 a 11 vezes [8]

- Etnia, o câncer de próstata parece estar associado ao estilo de vida ocidental, sendo encontrada uma menor incidência na população asiática. A mortalidade relacionada

ao câncer é 2,4 vezes maior na população afro–americana quando comparados à raça branca [8].

- A idade é considerada como principal fator de risco, pois cerca de 65% dos casos de câncer de próstata são diagnosticados em pacientes com idade superior a 65 anos, sendo apenas 0,1% dos casos diagnosticados antes dos 50 anos de idade [8].

Com o aumento da expectativa de vida mundial, é esperado que o número de casos novos de câncer de próstata aumente cerca de 60%. Com relação à história familiar, aproximadamente 25% dos casos diagnosticados apresentam história familiar de câncer de próstata.Dieta e nutrição também são fatores importantes na etiologia do câncer de próstata. O excesso de peso corporal, assim como uma dieta com carne vermelha em demasia, apresenta aumento no risco de desenvolver esse tipo de câncer [2].

2.3.1 Diagnóstico do câncer de próstata

Quanto ao diagnóstico precoce do câncer de próstata de acordo com a Organização Mundial de Saúde (OMS) tem-se duas diferentes estratégias: uma destinada ao diagnóstico em pessoas que apresentam sinais iniciais da doença (diagnóstico precoce) e outra voltada para pessoas sem nenhum sintoma e aparentemente saudáveis (rastreamento) [6].

A utilização do PSA no rastreamento do câncer de próstata reduz a mortalidade por este câncer em cerca de 20%. A maioria das Sociedades de Urologia e das entidades de saúde preconiza o rastreamento do câncer de próstata através da associação do toque retal com a dosagem sérica do PSA. A chance do indivíduo com toque retal alterado ter câncer de próstata aumenta conforme o valor do PSA [8].

O PSA é uma glicoproteína sérica produzida pelo epitélio de revestimento dos ácinos glandulares, normalmente encontrado no interior do lúmen dos ductos prostáticos. Age na liquefação do líquido seminal. Sua elevação pode estar associada à transformação maligna do epitélio prostático [13].

Uma taxa elevada de PSA no sangue não garante a presença do câncer de próstata, pois este alto índice pode esta associado a uma outra patologia. Dentre as principais patologias prostáticas que resultam em elevação do PSA, citamos: a hiperplasia prostática benigna (HPB), a prostatite e o câncer de próstata[1].

Na presença de carcinoma, os níveis séricos de PSA sofrem influência da quantidade de tumor e da sua diferenciação histológica. Na maioria dos pacientes com PSA acima de 4ng/ml, recomendam-se testes adicionais para a confirmação diagnóstica de malignidade e estadiamento [13].

Variações dos valores de referência do PSA sérico podem ocorrer com a idade. A Tabela 1 apresenta as faixas de valores normais para o PSA, de acordo com a faixa etária.

Achados no exame clínico (toque retal) combinados com o resultado da dosagem do antígeno prostático específico (PSA, na sigla em inglês) no sangue podem sugerir

Tabela 1: Faixa de valores normais para PSA.

Faixa etária	PSA sérico (ng/ml)
40 – 49	0 – 2,5
50 – 59	0 – 3,5
60 – 69	0 – 4,5
70 – 79	0 – 6,5

Fonte [13]

a existência da doença. Nesses casos, é indicada a ultrassonografia pélvica (ou prostática transretal, se disponível) [6].

A ultrassonografia transretal (USGTR) tem o objetivo principal de orientar a biópsia de próstata em pacientes com suspeita de câncer de próstata, porém, em casos selecionados, pode auxiliar na detecção de eventuais lesões prostáticas, levando à indicação de biópsias. Somente 37,5% das lesões malignas da próstata são detectáveis na USGTR e 50% das lesões não palpáveis, com dimensões superiores a 1,0 cm em seu maior diâmetro, não são visíveis à USGTR . O valor preditivo positivo para o uso das diversas combinações de exames diagnósticos em população de rastreamento varia de 20% a 80% [8].

O resultado da ultrassonografia, por sua vez, poderá mostrar a necessidade de biópsia prostática transretal. O diagnóstico de certeza do câncer é feito pelo estudo histopatológico do tecido obtido pela biópsia da próstata. O relatório anatomopatológico deve fornecer a graduação histológica do sistema de Gleason, cujo objetivo é informar sobre a provável taxa de crescimento do tumor e sua tendência à disseminação, além de ajudar na determinação do melhor tratamento para o paciente [6]

A própria biópsia prostática por si já é um procedimento invasivo e com índice de complicações importante, principalmente nos pacientes diabéticos e com história de infecção urinária de repetição e prostatites. Mesmo apresentando provas de coagulação normais, o sangramento é a principal complicação da biópsia prostática. De 27% a 63% dos pacientes experimentam episódios de hematúria que pode persistir por até sete dias; destes, cerca de 0,7% apresentam retenção de coágulos. O sangramento retal é uma complicação comum, sendo habitualmente controlado durante a realização do exame [8].

Então o rastreamento como forma de detecção precoce do câncer de próstata pode não ser a forma mais indicada devido o alto risco de falsos positivos que podem levar a uma biópsia, e com isso causar várias sequelas a pacientes saudáveis. A U.S. Preventive Service Task Force (USPSTF) dos EUA, em sua revisão de 2012 das recomendações de 2008, fez recomendação contrária à realização rotineira do PSA para o rastreamento do câncer de próstata (recomendação grau D). Segundo a revisão, essa prática deve ser desencorajada, pois há moderada ou alta certeza de que os danos associados ao rastreamento do câncer de próstata superam seus possíveis benefícios, os quais seriam no máximo muito pequenos. Esta recomendação é direcionada à população masculina dos EUA, independentemente da faixa etária [14].

Tabela 2: Recomendações internacionais sobre o rastreamento populacional do câncer de próstata.

Local	Órgão	Ano	Recomendação
Brasil	INCA/ Secretaria de Atenção á Saúde/Ministério da Saúde	2013	Não recomenda a organização de programas de rastreamento populacional para o câncer de próstata. Homens que demandam espontaneamente a realização de exames de rastreamento devem ser informados sobre os riscos e possíveis benefícios associados a essa prática
Estados Unidos	United States Preventive Services Task Force (USPSTF)	2012	Não recomenda o rastreamento do câncer de próstata com PSA
Reino Unido	National Screening Committee in the United Kingdom	2010	Não recomenda programa de rastreamento populacional
Austrália	Ministério da Saúde	2010	Não recomenda programa de rastreamento populacional. Recomenda que pacientes que demandem por exames de rastreamento recebam informação apropriada para tomada de decisão

Fonte:[3]

A American Joint Committee on Cancer (AJCC) e a União Internacional de Controle do Câncer (UICC) utilizam o sistema de classificação TNM como uma ferramenta para os médicos estadiarem diferentes tipos de câncer com base em determinadas normas. Ele é atualizado a cada 6 a 8 anos para incluir os avanços na compreensão de uma doença como o câncer. No sistema TNM, a cada tipo de câncer é atribuída uma letra ou número para descrever o tumor, linfonodos e metástases[11].

A classificação do câncer de próstata segue o sistema TNM de 2002 Tabela 3. A forma mais utilizada para estadiar histologicamente o ade nocarcinoma de próstata é o escore de Gleason. O sistema é graduado de 2 a 10 de acordo com o grau de diferenciação celular, sendo 2 o menos agressivo e 10 o mais agressivo [8].

Tabela 3: Classificação do Tumor Nodo Metástase (TNM) para câncer de Próstata.

T	Tumor primário
Tx	O tumor primário não pode ser avaliado.
T0	Não há evidência de tumor primário
T1	Tumor não diagnosticado clinicamente, não palpável ou visível por meio de exame de imagem T1a Achado histológico incidental em 5% ou menos tecido ressacado T1b Achado histológico incidental em mais que 5% de tecido ressacado T1c Tumor identificado por biópsia de agulha
T2	Tumor limitado à próstata 1 T2a Tumor envolve metade de um dos lobos ou menos T2b Tumor envolve mais do que a metade de um dos lobos, mais não os dois lobos T2c Tumor envolve os dois lobos
T3	Tumor se extender através da capsula prostática 2 T3a Extensão extracapsular (unilateral ou bilateral) T3b Tumor invade vesícula seminal
T4	Tumor está fixo ou invade outras estruturas adjacentes, que não as vesículas seminais: colo vesical, sfincter externo, reto, músculos elevadores.
N	Linfonodos Regionais
Nx N0 N1	Os linfonodos regionais não podem ser avaliados Ausência de metástase em linfonodo regional Metástase em linfonodo regional
M	Metástase à distância
M0	Ausência de metástase à distância
M1	Metástase à distância
	M1a Linfonodo não regionais M1b Ossos M1c Outras localizações

Fonte: [8]

O estadiamento clínico é de fundamental importância para a melhor definição das possibilidades terapêuticas. A chance de metástases ósseas aumenta quando o PSA é

$> 20ng/mL$ e na presença de tumores moderadamente diferenciados ou indiferenciados. Os pacientes, podem ser classificados em baixo, médio e alto risco de acordo com o valor de PSA e dados da biópsia, inclusive o escore de Gleason. A Tabela 4 de acordo com o valor de PSA e dados da biópsia, inclusive o escore de Gleason [8].

Tabela 4: Classificação de acordo com o grupo de risco dos pacientes com diagnósticos de câncer de próstata.

Grupo de Risco	Estadiamento Clinico	PSA(ng/mL)	Escore de Gleason	Critérios de biopsia
Baixo	T1a ou T1c	< 10	2–5	Unilateral ou menor $< 50\%$ da biópsia
Intermediário	T1b ou T2a	< 10	6 ou $4 + 3 = 7$	Bilateral
Alto	T2b ou T3	$10 - 20$	$4 + 3 = 7$	$> 50\%$ de envolvimento na biópsia
Muito alto	T4	> 20	$8 - 10$	Invasão linfovascular ou diferenciação neuroendócrina

Fonte [8]

2.3.2 Tratamento do câncer de próstata

Atualmente, as áreas técnicas das secretarias de saúde e o próprio Ministério da Saúde têm sido pressionados a incorporar cada vez mais tecnologias, como métodos de detecção precoce, diagnóstico e tratamento de todos os tipos de câncer. Entretanto, o Ministério da Saúde segue a tendência internacional de ampliar o uso de novas tecnologias ou das já existentes para outro fim terapêutico somente mediante evidências científicas bem estabelecidas [3].

Pacientes com câncer de próstata de baixo e intermediário risco são aqueles com doença localizada na próstata, baixo PSA e Gleason menor ou igual a 7. A primeira opção de tratamento para pacientes com expectativa de vida superior a cinco anos e que não tenham contraindicação cirúrgica é a prostatectomia radical. A depender do risco, pode-se optar por realizar, concomitantemente, a linfadenectomia pélvica. Nessa cirurgia, são retiradas por inteiro a próstata e as vesículas seminais. Os principais efeitos colaterais da prostatectomia radical, seja ela por via retropúbica aberta, perineal, laparoscópica ou robótica, são a disfunção erétil e a incontinência urinária [8].

Para doença localizada, cirurgia, radioterapia e até mesmo observação vigilante (em algumas situações especiais) podem ser oferecidos. Para doença localmente avançada, radioterapia ou cirurgia em combinação com tratamento hormonal têm sido utilizados. Para doença metastática (quando o tumor original já se espalhou para outras partes do corpo), o tratamento de eleição é a terapia hormonal [9].

A vigilância ativa tem sido empregada em alguns países nos casos de pacientes de baixo risco e baixo volume tumoral, ou que não sejam candidatos a nenhum outro tipo de tratamento. No nosso país, essa é uma prática atualmente reservada a casos muitos bem selecionados, principalmente quando o paciente não é candidato a nenhum outro tratamento ou quando ele deseja não se expor aos riscos inerentes às diversas formas de [8].

Pacientes de alto e muito alto risco devem receber tratamento mais agressivo, visto o maior potencial metastático da doença nestes casos. Mas, de uma maneira geral, seguem o mesmo princípio: controle oncológico mantendo a melhor qualidade de vida possível [8].

Assim, os tratamentos são indicados conforme risco do paciente. Pacientes de alto e muito alto risco devem receber tratamento mais agressivo, enquanto que os de baixo risco podem ser indicados uma observação vigilante e um tratamento hormonal. A Tabela 5 mostra as recomendações para cada tipo de paciente.

Tabela 5: Opções terapêuticas para tratamento do câncer de próstata de acordo com estadiamento e expectativa de vida. VA = vigilância ativa; PRR = prostatectomia radical; HT = Hormonioterapia; RT = Radio externa; TS = Terapia sitemática; TI = Terapia em investigação.

Risco	Expectativa de vida (anos)	Opções terapêuticas
Baixo	0 – 5	VA, HT
	5 – 10	VA, RT, HT
	> 10	PR, RT, VA
Intermediário	0 – 5	VA, RT, HT
	5 – 10	RT, HT, PR
	> 10	PR, RT, Ht
Alto	0 – 5	VA, RT
	5 – 10	RT, HT, PR
	> 10	RT, PR+RT+HT, HT
Muito alto	0 – 5	VA, RT
	5 – 10	HT, RT+HT, TS
	> 10	TS, TI

Fonte: adaptada de [8]

- **Doença metastática**

  Quando há doença metastática à época do diagnóstico, não é mais recomendado o tratamento local da doença através de prostatectomia radical ou radioterapêutica, sendo indicado o tratamento sistêmico da doença através de castração cirúrgica ou medicamentosa (bloqueio hormonal). Há algumas opções de castração medicamentosa, como os análogos do GnRH e os antiandrogênicos.

- **Hormônio refratário**

  Os pacientes em tratamento com bloqueio hormonal podem experimentar longos períodos de remissão de doença, porém após um período variável de tempo pode ocorrer a progressão da doença, mesmo com níveis baixos de testosterona, caracterizando a fase denominada de, escape hormonal, estes pacientes apresentam elevação do PSA ou alguma evidência clínica de progressão da doença, com testosterona baixa.

- **Quimioterapia**

  A quimioterapia sistêmica no câncer de próstata é reservada para pacientes com quadro avançado, principalmente aqueles pacientes que já não mais respondem às opções terapêuticas da hormonioterapia e possuem doença metastática dolorosa.

2.4 Biomarcadores e a Proteômica

As proteínas desempenham a maior parte das funções fisiológicas das células, constituindo também importantes alvos farmacológicos e biomarcadores de doenças. A pesquisa qualitativa, quantitativa e a elucidação estrutural destas moléculas são fundamentais para a compreensão do funcionamento dos sistemas biológicos, bem como na aplicação destas para o desenvolvimento de novos métodos diagnóstico [16].

Proteoma designa o conjunto de proteínas que estão sendo expressas por uma célula, tecido ou organismo em um determinado momento. De forma distinta, a análise proteômica consiste no estudo do proteoma utilizando técnicas de separação e identificação,tais como eletroforese, cromatografia, espectrometria de massas e bioinformática [17].

O estudo do proteoma nos permite identificar as proteínas que estão sendo expressas em um determinado momento, quantificá-las e observar suas modificações. Dessa maneira, a análise proteômica fornece informações mais abrangentes e que não podem ser inferidas a partir das informações obtidas através da análise genômica. Este tipo de estudo envolve etapas como: extração e tratamento da amostra, separação das proteínas e/ou peptídeos, espectrometria de massas e análise dos dados usando ferramentas de bioinformática [16].

O início da proteômica foi marcado pela caracterização de perfis proteicos, passando, posteriormente, a focar outros aspectos como a quantificação de proteínas, as interações entre proteínas e as modificações pós-tradicionais [15].

Um dos objetivos da proteômica é caracterizar estados de "sistemas biológicos"através de alterações no perfil de expressão de proteínas. Existem diversos métodos para obtenção de perfis proteômicos ou comparação de expressão de níveis de proteínas (ex. marcação por isótopos, comparação entre intensidade de íons, comparação entre o número de contagens espectrais [20] na fase de identificação de proteínas a maior parte das técnicas utilizadas fazem uso da espectrometria de massas.

Após a extração de proteínas, o resultado é uma mistura complexa que deve ser resolvida em frações simples de proteínas individuais ou em uma mistura simples de proteínas para identificação Figura 5. Na análise *bottom-up*, de misturas complexas, são utilizados géis desnaturantes (2D) ou cromatografia líquida [15].

A mistura complexa de proteínas resultantes da extração de proteínas totais pode ser resolvida por meio de HPLC (high performace liquid cromatography) convencional ou do sistema MudPIT, e/ou ainda utilizando géis. Uma vez obtida uma mistura simples, esta é submetida ao sequenciamento para identificação das proteínas. Os espectros de massa/carga obtidos podem ainda ser utilizados na quantificação dos níveis de expressão de proteínas de interesse e na identificação de modificações pós-traducionais (MPT's). O sequenciamento MS/MS pode ainda ser destinado à identificação em larga escala de proteínas que apresentam interações moleculares, previamente isoladas em experimentos conduzidos *in silico*, *in vitro* ou *in vivo* (interatômica) [15].

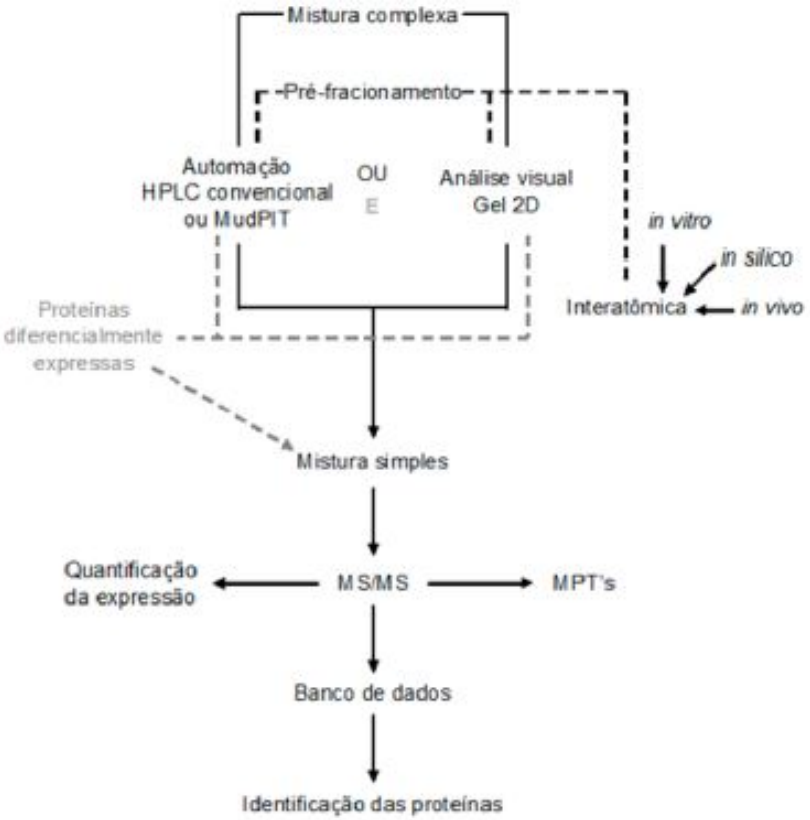


Figura 5: Fluxo experimental de um estudo de proteômica.
Fonte:[15]

Na busca de marcadores moleculares que auxiliem no diagnóstico precoce e no tratamento de várias doenças humanas, incluindo câncer, muitos estudos têm focado em alterações nos genes, seus transcritos e produtos proteicos envolvidos em processos celulares importantes[19].

Modificações genéticas que promovam a ausência de algumas proteínas ou defeitos em sua estrutura (afetando sua função), podem acarretar doenças ou ser marcadores destas, como nos casos da fenilcetonúria (doença causada pela atividade reduzida ou inexistente da fenilalanina hidroxilase)(ARN, 2014) e a anemia falciforme (causada pela substituição de um resíduo de glutamato por uma valina na posição seis das cadeias betas de globina, promovendo uma mudança importante na estrutura terciária da proteína e da morfologia da célula [16].

Muitas das técnicas empregadas em proteômica têm como foco a identificação de biomarcadores, mas são limitadas nas aplicações médicas diretas. Outras têm potencial para automatização e utilização na rotina clínica com propósitos diagnósticos e permitem a análise de muitos tipos de amostras e de alterações no padrão de expressão proteica associadas a uma doença [19].

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 obti-

das 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 [23].

Diferentes metodologias figura 6 podem ser combinadas em estudos proteômicos. As metodologias mais comumente utilizadas envolvem extração de proteínas da amostra, separação por eletroforese uni (1-DE) ou bidimensional (2-DE) e/ou por cromatografia líquida, ionização, fragmentação, análise e detecção de peptídeos e análise de dados [19].

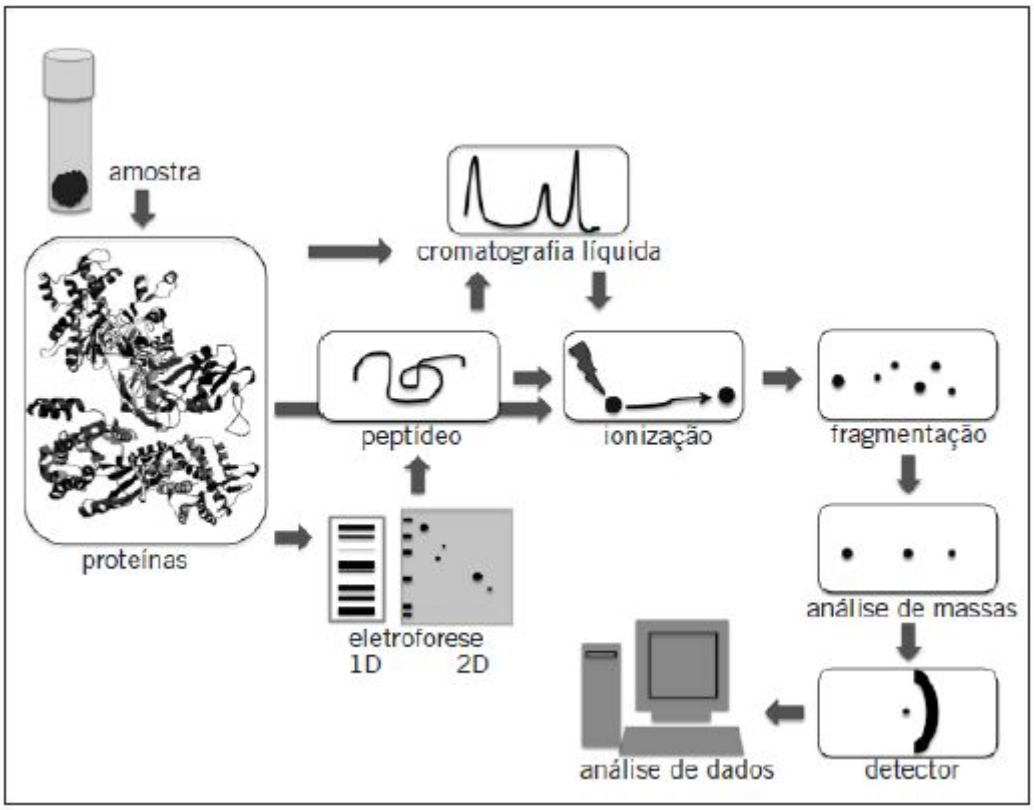


Figura 6: Diferentes metodologias podem ser empregadas em estudos proteômicos. Fonte:[19]

Na etapa da separação de proteínas a eletroforese, especialmente a eletroforese bidimensional (eletroforese 2D), é um método comumente utilizada para analise de proteínas. Esta técnica consiste na aplicação de um campo elétrico para separação destas macromoléculas, primeiramente, de acordo com o ponto isoelétrico e, posteriormente, por volume molecular [16].

As principais limitações associadas à eletroforese bidimensional são a sua baixa reprodutibilidade e o seu pequeno poder de automação. A reprodutibilidade pode ser aumentada definindo-se condições ótimas para a eletroforese, mas a automação do processo só é possível com relação à análise de géis [15].

Atualmente, existem dois métodos principais de ionização utilizados em proteômica, o MALDI (Matrix-Assisted Laser Desorption/Ionization) e o ESI (Electrospray Ionization),

o primeiro empregado para amostras em estado sólido e o segundo para amostras em estado líquido (Figura 2). No método MALDI, os peptídeos são cocristalizados com uma matriz orgânica, geralmente ácido α-ciano-4-hidroxicinamínico. Após bombardeamento por laser, a matriz sublima e seus íons transferem a carga para os analitos, resultando na formação de íons peptídicos. 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 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), comumente referido como espectro MS [19].

Em [16], tem-se uma comparação entre as duas técnicas de ionização MALDI e ESI-MS, e diz que ambas são eficientes para proteômica e podem ser aplicadas a analitos em concentrações mais baixas que picomols. Uma das maiores diferenças entre as técnicas é o estado em que os analitos são introduzidos no ionizador. Apesar da técnica de eletrosprayser capaz de reproduzir melhor os dados que a técnica de MALDI, deve-se atentar para o fato de que a abundância relativa dos íons gerados no ESI e MALDI não é uma representação real da concentração da amostra. Consequentemente, um padrão interno deve ser utilizado, preferencialmente um análogo isotópico do seu analito, para propósitos de quantificação.

Uma das limitações do sistema MALDI-TOF é a dificuldade de detecção de proteínas de baixo peso molecular que geram, por causa dessa característica, poucos peptídeos. O sistema também não é capaz de detectar mais de um componente de uma mistura. Para melhorar o desempenho, os analisadores TOF podem ser combinados com analisadores quadrupolos (Qs), que apresentam um conjunto de quatro eletrodos em bastão e funcionam como filtros de massas. Entre esses eletrodos, um campo elétrico assegura que somente íons de uma determinada razão m/z sigam a trajetória ao detector enquanto os demais são desviados [19].

O analisador de massas é a parte do instrumento que realiza a separação dos íons gerados através da sua razão massa/carga. Assim como os processos de ionização, tem-se um enorme leque de opções que podem realizar esta separação [16]. Independentemente do método de ionização, a massa molecular dos íons é avaliada em um analisador após passagem por uma câmara de vácuo [24]. Três diferentes princípios podem ser aplicados para alcançar a separação das massas: separação baseada no tempo de voo (TOF MS), separação através de campos elétricos gerados por hastes metálicas (quadrupolo MS), ou a separação pela ejeção seletiva de íons de um campo de aprisionamento de íons tridimensional (aprisionamento de íons ou transformada de Fourier cíclotron íon MS). Para analises estruturais, como o sequenciamento de peptídeos, duas etapas de MS podem ser realizadas em sequência (espectrometria de massas em sequência ou MS/MS), empregando o mesmo princípio de separação duas vezes ou através da combinação de dois princípios

de separação [25].

O analisador de tempo de voo separa os íons baseando-se na velocidade dentro do tubo de voo. Teoricamente, os íons são formados no mesmo lugar ao mesmo tempo na fonte de íons e então acelerados a um potencial fixo para dentro do "OF drift tube" – tubo de desvio TOF. Uma vez que todos os íons formados com a mesmacarga apresentam a mesma energia potencial elétrica quando expostos ao campo, pode-se inferir a razão massa carga dos mesmos, pois esta energia será convertida em energia cinética que é uma função, entre outras coisas, da massa da molécula; consequentemente, íons com menor valor de m/z alcançarão uma maior velocidade que os íons de maior m/z [16].

Elementos de um espectrômetro

Os espectros de massas basicamente possuem uma mesma estrutura e são constituídos principalmente pelos seguintes itens:

1. Sistema de alto vácuo. Para isso utiliza-se uma série de bombas (ex. bombas turbo moleculares, etc.);

2. Sistema de entrada de amostras (amostras líquidas podem entrar por capilares ou amostras sólidas podem ser colocadas em placas específicas);

3. Fonte de ionização onde as amostras são convertidos em íons e transferidos para a fase gasosa;

4. Analisador de massas. Pode ser de vários tipos TOF, quadrupolo, quadrupolo-ion trap, etc;

5. Detector.

Quando o objetivo do estudo compreende a produção de mapas proteômicos, a espectrometria de massas surge imperativa, permitindo o processamento de centenas de amostras em uma única análise. A identificação de proteínas por meio da espectrometria de massas depende da digestão proteolítica que produz uma coleção de peptídeos que são ionizados por eletronebulização ou por dessorção a laser auxiliada por matriz [15].

Atualmente, é inconcebível analisar qualquer processo biológico ou elaborar a criação de novos painéis de proteínas marcadoras para diagnóstico (biomarcadores) sem considerar o uso da proteômica. Porém, os dados gerados por esta tecnologia são numerosos e de interpretação difícil; isto implica a necessidade do desenvolvimento de algoritmos especializados [28].

A perspectiva oferecida pela proteômica tem-se utilizado na pesquisa de diferentes áreas da medicina, incluído a biomedicina. Estas pesquisas poderiam se classificar de distintas formas: em função do tipo de amostra empregada, da doença ou do tipo de doenças que abordam, da técnica ou das técnicas utilizadas, do uso ou da aplicação. As pesquisa clinicas com proteômica se baseiam em torno ao tipo de amostra utilizada.

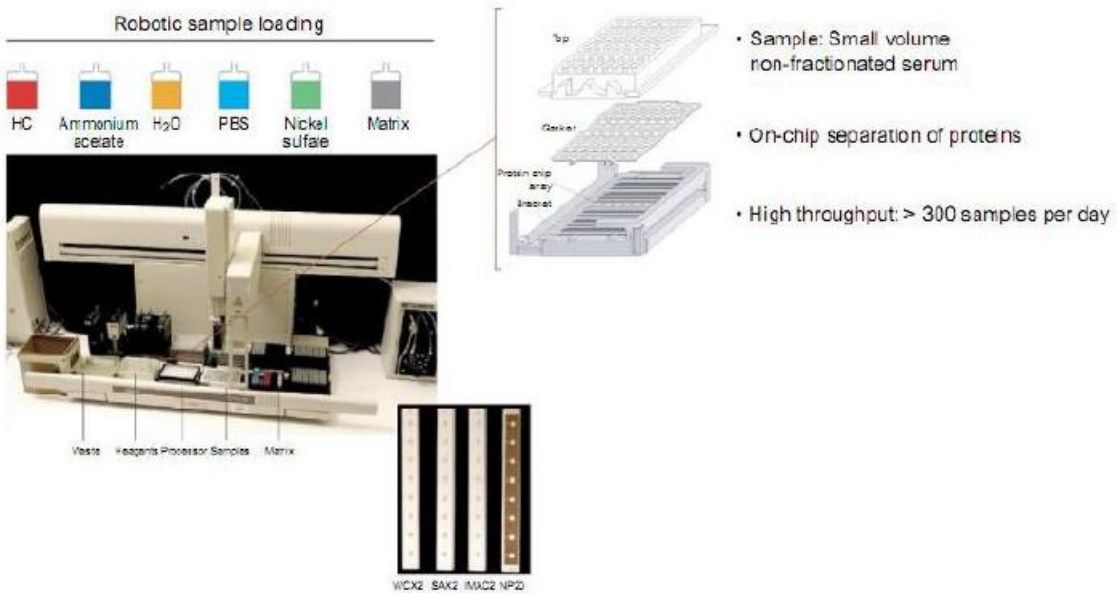


Figura 7: Espectrômetro de massa utilizado na obtenção na obtenção de padrões proteômicos.

Fonte:[19]

Nesta base de pesquisas destaca-se as pesquisas com linhagens celulares, com tecidos e proteômica de fluidos [26].

Nos últimos anos, muitas questões biológicas importantes têm sido respondidas pela proteômica e centenas de biomarcadores candidatos foram descobertos. Entretanto, poucos desses marcadores têm ultrapassado a fase de identificação. Sua aplicação com sucesso na prática clínica dependerá de plataformas sensíveis, desenvolvimento de painéis de proteínas e estudos colaborativos que incluam médicos, epidemiologistas, biologistas moleculares e bioinformatas, com uma questão clínica relevante e com parâmetros bem definidos de recrutamento e caracterização de pacientes e amostras [19].

2.5 Aprendizado de Máquinas

A grande variedade de dados coletados e armazenados em um ritmo cada vez mais acelerado, tem demonstrado a necessidade do surgimento de uma nova geração de teorias da computação, bem como de ferramentas que auxiliem os seres humanos a extraírem informações úteis deste crescente volume de dados [36].

A Mineração de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, embutidos nos dados. Sendo estes padrões definidos como modelos extraídos dos dados, determinados por um número de características comuns a eles [37]. Um processo de aprendizagem inclui a aquisição de novas formas de conhecimento: o desenvolvimento motor e a habilidade cognitiva (através de instruções ou prática), a organização do novo conhecimento (representações efetivas) e as descobertas

de novos fatos e teorias através da observação e experimentação. Desde o início da era dos computadores, tem sido realizadas pesquisas para implantar algumas destas capacidades em computadores. Resolver este problema tem sido o maior desafio para os pesquisadores de inteligência artificial (IA). O estudo e a modelagem de processos de aprendizagem em computadores e suas múltiplas manifestações constituem o objetivo principal do estudo de aprendizado de máquinas [47]. O aprendizado de máquina nada mais é um aprendizado por experiência, que conforme a tarefa é executada, o problema aprende a melhor maneira de resolver. Além de estruturar o conhecimento existente, para levar a um entendimento do aprendizado [39].

Diversos sistemas de AM utilizam operações de generalização e especialização para criar hipóteses a partir de exemplos. Em especial, os algoritmos capazes de representar a hipótese do conceito a ser aprendido utilizando como linguagem de representação regras de decisão [44, 45]. As técnicas de aprendizados de máquinas empregam um princípio de inferência denominado indução, no qual é possível obter conclusões genéricas a partir de um conjunto particular de exemplos. Estas técnicas de aprendizados indutivos podem ser divididas em dois principais tipos: os supervisionados e os não supervisionados [38], no presente trabalho abordamos o aprendizado supervisionado.

Especialmente em problemas de predição de ações, a aplicação de algoritmos de aprendizados de máquinas ainda tem se concentrado em dois algoritmos principais e suas modificações: Máquina de Vetores de Suporte (SVM) e Redes Neurais (NN). De um modo geral os trabalhos propõem diferentes configurações para os algoritmos, junção de técnicas e pré e pós-processamento [47].

2.5.1 Aprendizado Supervisionado

O aprendizado supervisionado uma das fundamentais, no qual presume-se a existência de um tutor conhecedor de um conjunto de exemplos de pares entrada-saída, supostamente, relevantes. No aprendizado supervisionado é fornecida uma referência do objeto a ser alcançado, isto é, um treinamento com conhecimento do ambiente, onde este treinamento é um conjunto de exemplos com entradas e uma saída esperada. Em contraste, há também o aprendizado não supervisionado, caracterizado principalmente pela ausência desse tutor, podendo, de acordo com [35], ser subdividido em aprendizado auto-organizado e aprendizado por reforço.

No aprendizado supervisionado o conjunto de exemplos fornecidos, de um modo geral é descrito por um vetor de valores de características, ou atributos, e o rótulo da classe associada. O objetivo do algoritmo de indução é construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados, ou seja, exemplos que não tenham ainda o rótulo da classe. Para rótulos de classe discretos, esse problema é conhecido como classificação e para valores contínuos como regressão.

No aprendizado supervisionado é fornecido ao sistema de aprendizado um conjunto de

exemplos $E = E_1, E_2, .., E_N$, sendo que cada exemplo $E_i \epsilon E$ possui um rótulo associado. Os rótulos ou classes representam o fenômeno de interesse sobre o qual se deseja fazer previsões. Um pouco mais formalmente, pode-se dizer que cada exemplo $E_i \ \epsilon \ E$ é uma tupla

$$E_i = (\vec{x}_i, y_i) \tag{1}$$

na qual $\vec{x}_i$ é um vetor de valores que representam as características, ou atributos, do exemplo E_i, e y_i é o valor da classe desse exemplo. O objetivo do aprendizado supervisionado é induzir um mapeamento geral dos vetores $\vec{x}$ para valores y. Portanto, o sistema de aprendizado deve construir modelo, $y = f(\vec{x})$, de uma função desconhecida, f também chamada de função conceito, que permite predizer valores y para exemplos previamente não vistos. Entretanto, o número de exemplos utilizados para a criação do modelo não é, na maioria dos casos, suficiente para caracterizar completamente essa função f. Na realidade, os sistemas de aprendizados são capazes de induzir uma função h que aproxima f, ou seja, $h(\vec{x}) \approx f(\vec{x})$. Nesse caso, h é chamada de hipótese sobre a função conceito f [43].

Segundo Mitchell (1997), o erro de um aprendizado supervisionado pode ser calculado como a diferença ente a saída desejada e a saída gerada, de acordo com a equação 2:

$$e_k = d_k - y_k \tag{2}$$

onde k representa o estímulo, e o sinal de erro e d a saída desejada apresentada durante o treinamento, y = saída real algoritmo após a apresentação do estimulo de entrada.

2.5.2 Aprendizado não-supervisionado

Como já comentado este tipo de aprendizado se caracteriza principalmente pela ausência do tutor. Aqui não se utiliza referências, ou seja, não ocorre o treinamento com o conhecimento do ambiente. O algoritmo de aprendizado aprende a representar (ou agrupar) as entradas submetidas, segundos medidas de similaridade [38]. O indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos ou clusters [46]. O mesmo pode ser subdivido em aprendizado auto-organizado e aprendizado por reforço. No aprendizado auto-organizado, ajustam-se os parâmetros livres de um modelo, usualmente com base em regras competitivas, de modo a otimizar uma dada medida de qualidade. O aprendizado por reforço é realizado por meio de um processo contínuo de interação entre a máquina e seu ambiente, buscando atingir objetivos específicos descritos por meio de um índice de desempenho [35]. As

técnicas de aprendizado não supervisionado são mais utilizadas quando o entendimento dos dados é feito através de padrões ou tendências [42].

2.6 O uso de reconhecimento de padrões na proteômica

É praticamente impérvio estudar algum tipo de amostra sem utilizar a proteômica. Todavia, os resultados gerados por ela são muitos e suas interpretações são difíceis; por conseguinte, algoritmos especializados são desenvolvidos a fim de possibilitar as análises. Algoritmos estes, que utilizam técnicas de reconhecimento de padrões probabilísticos e inteligência artificial para extrair dos resultados, informações necessárias para um determinado estudo [30].

O Reconhecimento de Padrões, ou Pattern Recognition, é uma subárea da Inteligência Artificial que trata da classificação e descrição de objetos [30]. Entende-se por padrão as propriedades que possibilitam o agrupamento de objetos semelhantes dentro de uma determinada classe ou categoria, mediante a interpretação de dados de entrada, que permitam a extração das características relevantes desses objetos [29].

Define-se reconhecimento de padrões como sendo um procedimento em que se busca a identificação de certas estruturas conhecidas e sua posterior classificação dentro de categorias, de modo que o grau de associação seja maior entre estruturas de mesma categoria e menor entre as categorias de estruturas diferentes. Diz ainda que os dados de entrada são medidos por sensores e selecionados segundo o conteúdo de informações relevantes para a decisão, e passam por um processo de redução de sua dimensionalidade para que possam ser usados pelo classificador, que o designará à classe que melhor o represente [31].

O reconhecimento de padrões é a área da ciência que qualifica e classifica objetos e formas com um determinado número de classes, a partir da análise de suas características. Atualmente é bastante comum sistemas computacionais possuírem processamento de objetos e análise da informação produzida por este processamento. O objetivo da extração de características de padrões é obter sistemas cada vez mais automáticos, capazes de abstrair, reconhecer e classificar objetos por meio do uso de uma ferramenta comum chamada de padrão, que capta semelhanças ou diferenças entre os elementos analisados e caracteriza o objeto [32].

Um sistema para reconhecimento de padrões engloba três grandes etapas: representação dos dados de entrada e sua mensuração, extração das características e finalmente identificação e classificação do objeto em estudo. A primeira etapa refere-se à representação dos dados de entrada que podem ser mensurados a partir do objeto a ser estudado. Essa mensuração deverá descrever padrões característicos do objeto, possibilitando a sua posterior classificação numa determinada classe. A segunda etapa consiste na extração de características intrínsecas e atributos do objeto e consequente redução da

dimensionalidade do vetor padrão. É a fase da extração das características. A escolha das características é de fundamental importância para um bom desempenho do classificador. A terceira etapa em reconhecimento de padrões envolve a determinação de procedimentos que possibilitem a identificação e classificação do objeto em uma classe de objetos [31].

2.7 Análise de Componentes Independentes (ICA)

Nesta subsecção, será discutido os conceitos básicos da técnica Análise de Componentes Independentes (ICA), sua definição como modelo estatístico de variáveis e algumas de suas restrições que levam este modelo a ser um estimador.

As origens do ICA vêm dos trabalhos de Darmois [50] na década de 50 e Kagan et al. [51] na década de 70, caracterizando variáveis aleatórias em estruturas lineares. Os primeiros trabalhos em análise de componentes independentes foram desenvolvidos por Jutten & Herault [52] na década de 80, e Comon [53] que desenvolveu uma teoria matemática sobre a separação cega de fontes, na década de 90, onde formalizou e desenvolveu toda a teoria básica da analise de componentes independentes. Desde o seu surgimento até os dias atuais, diversas aplicações do ICA têm sido propostas nos mais diversos problemas e têm-se obtidos bons resultados.

A análise de componentes independentes, do inglês *Independent Component Analysis* (ICA) é um método computacional desenvolvido inicialmente, para resolver problemas de Separação Cega de Fontes, do inglês *Blind Source Separation* (BSS) [49]. A presente técnica é capaz de determinar características embutidas em um conjunto de sinais. O ICA define um modelo gerador para os dados observados, que são denominados misturas de variáveis ocultas e desconhecidas. As variáveis latentes são assumidas mutuamente independentes, e são chamadas de componentes independentes ou fontes dos dados observados [35]. O principal objetivo do ICA é determinar uma representação linear de dados não gaussianos, minimizando a dependência estatística entre eles de forma que os componentes resultantes sejam estatisticamente independentes. A ICA se diferencia das demais técnicas existentes, justamente pelo fator de trabalhar com as componentes que são ao mesmo tempo não gaussianas e estatisticamente independentes, informalmente, isto equivale a afirma que os sinais-fontes não dependem um do outro do ponto de vista estatístico.

2.7.1 Definições

Considere que seja observado n misturas lineares de um sinal, modeladas como combinações lineares de n funções bases 3:

$$x_i = a_{i1}s_1 + a_{i2}s_2 + ... + a_{in}s_n, \text{ para todo } i = 1, ..., n \tag{3}$$

onde, temos que

- x_i=sinal aleatório
- s_i=Componentes independentes aleatórias
- a_i=coeficientes de mistura

e que cada sinal x_i, assim como cada componente s_i seja uma variável aleatória. Em análise funcional e suas aplicações, um espaço de funções pode ser interpretado como um espaço vetorial de dimensão infinita, cujo os vetores bases são funções e não vetores. De modo geral quer dizer que cada função no espaço pode ser representada como combinação linear das funções base [55].

Como uma combinação linear, o modelo pode ser representado por 4:

$$x_i = \sum_{i=1}^{n} a_i s_n \tag{4}$$

Na notação matricial podemos reescrever a equação 4 da seguinte forma:

$$\mathbf{X=A.S} \tag{5}$$

O modelo estatístico apresentado em 5 é chamado de modelo de análise de componentes independentes. É preciso estimar tanto a matriz de componentes independentes S quanto a matriz de funções base A, que também é desconhecida, pois tudo que se observa são as amostras do sinal X.

Por tanto é necessário fazer algumas suposições, tão gerais quanto possível. Para isso, considera que [49]:

- As componentes s_n são estatisticamente independentes;
- Elas possuem distribuição não-gaussiana;
- Para facilitar a aplicação do modelo, a matriz A é quadrada

O ICA é um modelo gerativo, isto é, descreve como os dados observados são gerados através do processo de mistura das componentes s_i.

Objetivando a ilustração do modelo ICA, considere duas variáveis aleatórias x_1 e x_2, pela definição dada temos que:

$$x_1 = a_{11}s_1 + a_{12}s_2 \tag{6}$$

$$x_2 = a_{21} + a_{22}s_2 \tag{7}$$

Assim, os x_i são as misturas observadas, os a_{ij} são os coeficientes da matriz de mistura e s_i são componentes independentes, de modo que:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} s_{11} & s_{12} & \dots & s_{13} \\ s_{21} & s_{22} & \dots & s_{23} \end{pmatrix} \tag{8}$$

O problema na estimativa do modelo ICA é estimar as componentes independentes, dado que os coeficientes de misturas e as componentes independentes não se tem conhecimento a cerca das mesmas, isto é, estimar uma matriz de separação W composta por vetores linha w_i, onde $i = 1, ..., n$ tal que:

$$s = Wx \tag{9}$$

Como a matriz A é desconhecida, não se pode detreminar uma matriz W tal a equação 9 seja satisfeita, mas pode se determinar uma matriz W^* de modo que:

$$y = W^*x \tag{10}$$

onde $\|s - y\| = min.$

O modelo ICA apresenta algumas ambiguidades em relação às componentes independentes. São elas:

- Não se pode determinar suas variâncias (energias);
- Não se pode determinar a ordem que serão estimadas;

Estas ambiguidades se devem ao fato de **A** e **S** serem desconhecidas. Como resultado, não é possível determinar as energias ou as amplitudes dos sinais, tão pouco os sinais ou ordem de s_n [49].

2.7.2 As restrições do modelo ICA

1. As componentes independentes são estatisticamente independentes.

   Duas ou mais variáveis aleatórias são ditas independentes se a informação contida nos valores de qualquer uma delas não fornece nenhuma informação a acerca dos valores de qualquer outra. Sejam x_1 e x_2 duas variáveis aleatórias, estas variáveis são denominadas independentes se, e somente se, x_1 não fornece nenhuma informação de x_2 e vice-versa. A independência estatística pode ser definida formalmente por meio das funções de probabilidade (fdp) das variáveis aleatórias. Considere $p(x_1, x_1)$ a função densidade de probabilidade conjunta (fdp) das variáveis aleatórias x_1 e x_2 e $p_1(x_i)$ a função densidade de probabilidade marginal de x_i, ou seja, a função densidade de probabilidade de y_i quando somente esta é considerada. É dito que y_1 e y_2 são estatisticamente independentes se e somente se a função densidade de probabilidade conjunta for fatorável da seguinte maneira:

$$p(x_1, x_2) = p(x_1)p(x_2) \tag{11}$$

Em outros termos, pode-se dizer que a probabilidade conjunta x_1 e x_2 é igual ao produto das densidades marginais $p(x_1)$ e $p(x_2)$.

Uma propriedade importante da independência de variáveis aleatórias é que, dado duas funções h_1 e h_2, sempre terá que:

$$Eh_1(y_1)h_2(y_2) = Eh_1(y_1)Eh_2(y_2) \tag{12}$$

Duas variáveis x_1 e x_2 são descorrelacionada se a sua covariância for igual a zero, ou seja:

$$cov_{x_1,x_2} = E[(x_1 - \mu_1)]E[(x_2 - \mu_2)] = 0 \tag{13}$$

Sendo μ_1 e μ_2 as médias das variáveis x_1 e x_2, respectivamente.

2. As componentes independentes possuem distribuição de probabilidade são não-gaussiana.

   No modelo ICA, não é assumido que as distribuições de probabilidade das componentes independentes são conhecidas, porém, é preciso assumir que elas sejam não-gaussianas. As distribuição gaussiana são simétricas. Não há uma direção de maior concentração de valores que possa ser privilegiada na estimativa do modelo ICA. Embora seja assumido que as distribuições das componentes independentes sejam não-gaussianas, certamente as distribuições das misturas observadas serão.

   Considere um sistema de mistura (2) ortogonal e duas fontes gaussianas ($s_1 e s_1$). Logo, como os sinais misturados resultantes ($x_1 e x_2$) são gaussianos, não correlacionados e possuem variância unitária. Como a f.d.p. conjunta de (x_1, x_2) é simétrica, ela não contém nenhuma informação sobre a matriz de mistura A. Desta forma, pode-se dizer que a matriz A não pode ser estimada, se mais de uma das fontes originais for gaussiana [58]. De modo bem mais formal, pode-se comprovar que as distribuição das variáveis gaussianas não é afetada por qualquer transformação ortogonal e, que as variáveis são independentes. De modo que, para variáveis gaussianas, pode-se apenas estimar o modelo para ICA a menos de uma transformação ortogonal. Isso mostra e enfatiza a afirmação feita de que a separação em componentes independentes só é possível se no máximo uma fonte original for gaussiana.

   A função densidade de probabilidade para a distribuição gaussiana ou normal é definida por 14:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{14}$$

Onde μ e σ são, respectivamente, a média e o desvio padrão.

3. Para fins de simplicidade, será assumido que o número de componentes independentes equivale ao número de misturas observadas, isto é, a matriz de mistura é quadrada e, portanto, pode possuir inversa. Se esse não for o caso, haverão misturas redundantes que poderão ser omitidas.

2.7.3 Estimação de Componentes Independentes

1. Estimação Através de Maximização de Não-Gaussianidade

   A não-gaussianidade é um elemento chave para se estimar as componentes independentes no modelo ICA, pois a matriz A não é conhecida quando as componentes independentes possui distribuição gaussiana [56]. Supondo que x seja uma das amostras do sinal, distribuído de acordo com o modelo ICA em 5 e que todas as componentes independentes s possuem distribuições iguais. Para estimar as componentes independentes, basta encontrar as combinações lineares corretas de x_i, de modo que

$$S = A^{-1}X \tag{15}$$

   Supondo ainda uma combinação linear qualquer dos vetores x_i dada por

$$y = b^T X \tag{16}$$

   como $X = AS$, pode-se escrever:

$$\begin{aligned} y &= \textstyle\sum_i b_i x_i \\ &= b^T AS \end{aligned} \tag{17}$$

   onde b deve ser determinado. A partir de 17, pode-se observar que y é uma combinação linear de s_i, com coeficientes dados por $q = b^T A$. Logo obtém-se:

$$\begin{aligned} y &= q^T s \\ &= \textstyle\sum_i q_i s_i \end{aligned} \tag{18}$$

   se b corresponder a uma das linhas da inversa de A, então y será uma das componentes independentes e, nesse caso, apenas um dos elementos de q será igual a 1, enquanto todos os outros serão iguais a zero.

Como a penas o vetor de mistura X é conhecido e por isso b não pode ser determinado exatamente é preciso encontrar um estimador que forneça uma aproximação de b.

Uma maneira de se determinar b é variar os coeficientes em q e então verificar como a distribuição de $y = q^T s$ muda. De acordo com o teorema do limite central [57], a soma de duas variáveis aleatórias independentes é mais gaussiana que as variáveis originais [57], $y = q^T s$ normalmente é mais gaussiana que qualquer uma das s_i e menos gaussianas quando se iguala a umas das s_i. Nesta situação, apenas um dos elementos q_i de q é diferente de zero [49].

Sabe-se que na prática os valores de q são desconhecidos e sabe-se ainda que através das equações 17 e 18, obtém-se:

$$b^T x = q^T s \tag{19}$$

pode-se variar o valor de b, um vetor que maximiza a não-gaussianidade de $b^T x$, sendo que esse vetor necessariamente corresponde a $q = A^T S$ vetor esse que possui apenas uma de suas componentes diferente de zero. Isto que dizer que o y em 17 é igual a uma das componentes independentes. De modo que, a maximização da não-gaussianidade de $b^T x$ permite encontrar uma das componentes.

2. Negentropia como medida de Não-gaussianidade

A entropia é um conceito da Teoria de Informação e mede o grau de informação que pode ser obtida através da observação de uma variável e está relacionada com a quantidade de informação que essa variável possui. Sendo y um vetor aleatório com função densidade de probabilidade $f(y)$, a sua entropia diferencial é dada por:

$$H(y) = -\int f(y)\, log f(y)\, dy \tag{20}$$

Como um dos principais resultados da Teoria da Informação, sabe-se que uma variável gaussiana tem a maior entropia entre todas as variáveis aleatórias de igual variância [49][57]. Isto que dizer uma versão modificada da entropia diferencial pode ser usada como medida de não-gaussianidade. Essa medida é denominada de negentropia, definida por

$$J(y) = H(y_{gauss} - H(y)) \tag{21}$$

sendo y_{gauss} uma variável aleatória de mesma matriz de covariância que y. A negentropia sempre é não-negativa, tem valor igual a zero se, somente se, y possui distribuição gaussiana é invariante para transformações lineares inversíveis.

Apesar de sua qualidade como medida de não-gaussianidade, a negentropia é de difícil estimação. Logo, é necessário a utilização de aproximações utilizando momentos de alta ordem. Assim,

$$j(y) \approx \frac{1}{12} E\{y^3\}^2 + \frac{1}{48} kurt(y)^2 \tag{22}$$

sendo $kurt(y)$, ou seja, a *kurtosis* de y, definida como o momento de quarta ordem da variável aleatória y, expresso por

$$kurt(y) = E\{y^4\} - 3\left(E\{y^2\}\right)^2 \tag{23}$$

2.8 Redução da Dimensionalidade

O aumento da dimensionalidade do espaço de características, ou seja, do número de características, torna o sistema mais complexo, custa um maior tempo de treinamento e pode reduzir a capacidade de generalização do sistema[33]. As vantagens em reduzir a dimensionalidade da representação do padrão refletem-se na medida de custo e precisão do classificador. Além disso, uma pequena quantidade de características pode avaliar o problema da dimensionalidade, quando o número de exemplos de treinamento é pequeno. Porém, um reduzido número de características pode levar a uma fraca discriminação e consequentemente a uma precisão inferior no sistema de reconhecimento resultante. Mas a redução de dimensionalidade é necessária quando, por exemplo, é possível construir dois padrões arbitrários similares, codificando-os a partir de um grande números de características redundantes [34]. 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 [48]. A escolha das características ideais para a classificação é realizada através da observação dos objetos conhecidos e de suas respectivas classes, isto é, o conjunto de treinamento, porém toda redução de dimensionalidade resulta numa perda de informação e por isso, o objetivo principal da redução de dimensionalidade é preservar o máximo possível da informação relevante dos dados. 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 traz nenhuma informação nova para o vetor de características [10].

A redução de características consiste na escolha de um subconjunto das características mais informativas produzidas a partir dos sinais originais sem que se perca sua capacidade discriminante [59], isto é, aqui é realizada a escolha de um subconjunto ótimo de características, que representa a informação importante, contida nos dados, segundo al-

gum critério. Esta seleção é importante em casos que a medição da mesma é custosa, pois pode permitir que um subconjunto representativo, e menor que o original seja selecionado, melhorando a qualidade dos dados e os modelos construídos, tornando-se mais compreensíveis, deste modo o processamento se torna mais rápido.

2.8.1 Máxima Relevância e Mínima Redundância

Com o objetivo de reduzir o número de característica do problema aplicou-se, isto é, determinar um subconjunto ótimo utilizou-se algoritmo de máxima relevância e Mínima Redundância.

Considere v um vetor de características dos dados, com n amostras e m características, representado por $v = v_i, i = 1, ..., m$ e seja c um vetor de classe (rótulo).

O objetivo da seleção de características é determinar um conjunto de observação de dimensão v, um subconjunto de m características que represente c, de maneira satisfatória e efetiva.

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 c, é dada por [10]:

$$maxD(v, c), D = \frac{1}{|v|} \sum v_i \epsilon dI(v_i; c) \tag{24}$$

É provável que as características selecionadas de acordo com o critério acima descrito 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 [60], sendo assim, tem-se:

$$minR(v), R = \frac{1}{|v|^2} \sum_{v_i, v_j \epsilon v} I(v_i, v_j) \tag{25}$$

Os critérios descritos em 24 e 25 são chamados conjuntamente de Máxima-Relevância-Mínima-Redundância (mRMR) [48].

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

$$max\phi(D, R), \phi = D - R \tag{26}$$

2.9 Classificação

A escolha das características ideais para a classificação é realizada através da observação dos objetos conhecidos e de suas respectivas classes, isto é, o conjunto de treinamento.

Para construir um classificador, realizam-se três atividades: aquisição do corpus de documentos, que consiste na coleta de exemplos que identificam cada uma das classes a serem aprendidas; criação da representação dos documentos, extraindo características que conseguem identificar unicamente cada um dos exemplos; e indução do classificador, que usa um algoritmo de aprendizado para definir a partir de um conjunto de representações a qual categoria esse conjunto pertence, tornando possível a predição de classes para valores não utilizados durante o treinamento [40].

O desempenho desejado de um classificador f é que o mesmo obtenha o menor erro durante o treinamento, sendo o erro mensurado pelo número de predições incorretas de f. Sendo assim definimos como risco empírico $R_{emp}(f)$, como sendo a medida de perda entre a resposta desejada e a resposta real. A equação 27 mostra a definição do risco empírico.

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} c(f(x_i), y_i) \tag{27}$$

onde $c(.)$ é a função de custo relacionada a previsão $f(x_i)$ com a saída desejada y_i [38], onde uma função de custo é a "perda 0/1" definida pela equação 28. O procedimento pela busca de uma função f' que represente um menor valor de $R_{emp}(f)$ é denominado de Minimização do Risco Empírico.

$$c(f(x_i), y_i) = \begin{cases} 1, se\, y_i\, f(x_i) < 0 \\ 0, se\, y_i\, f(x_i) > 0 \end{cases} \tag{28}$$

Quanto a hipótese de que os padrões de treinamento (x_i, y_i) são gerados por uma distribuição de probabilidade $P(x, y)$ em $R^N \times \{-1, +1\}$ sendo P desconhecida. A probabilidade de classificação incorreta do classificador f é denominada de risco funcional, que quantifica a capacidade de generalização, conforme é mostrado pela equação 29 [69].

$$R(f) = \int c(f(x_i), y_i) dP(x_i, y_i) \tag{29}$$

Durante a fase de treinamento, $R_{emp}(f)$, pode ser facilmente obtido, ao contrário de $R(f)$, pois em geralmente a distribuição de probabilidade P é desconhecida [70].

2.9.1 Máquina de Vetores de Suporte

As Máquinas de Vetores de Suporte (SVMs, do Inglês Support Vector Machines) constituem uma técnica de aprendizado que vem recebendo crescente atenção da comunidade de Aprendizado de Máquina (AM) [39]. A SVM é uma técnica de aprendizado estatístico,

baseada no princípio da Minimização do Risco Estrutural (SRM), e pode ser usada para resolver problemas de classificação [67]. As SVMs constituem uma das técnicas de aprendizado de máquinas de maior sucesso e aplicação pela comunidade de Inteligência Computacional. Têm apresentado resultados equivalentes e, muitas vezes, superiores aos alcançados por outros algoritmos de aprendizado, inclusive outros tipos de RNAs [68].

A máquina de Vetores de Suporte (SVM) é utilizados em vários problemas de classificação regressão como em [63] [64] [65]. Se baseia na teoria de aprendizagem estatística de [61], que constrói um modelo que representa os exemplos como pontos em um hiper-espaço, por meio de funções *Kernel* [66]. No hiper espaço são obtidos vetores de suporte que representam os limiares de decisão do algoritmo. Para se realizar a classificação, um novo modelo é mapeado para um ponto no hiper-espaço e recebe a classe de acordo com a posição do modelo em relação ao limiar de decisão.

Algumas das principais características das SVMs [62], são:

- Boa capacidade de generalização - os classificadores de uma SVM em geral alcançam bons resultados em termo de generalização. Essa capacidade é medida por sua eficiência na classificação de dados que não pertençam ao conjunto utilizado em seu treinamento, portanto, é evitado o overfitting (memoriza os padrões de treinamento, gravando suas peculiaridades e ruídos, ao invés de extrair as características gerais que permitirão a generalização ou reconhecimento de padrões não vistos durante o treinamento).

- Robustez em grandes dimensões - as SVMs são robustas diante de objetos de grandes dimensões, como por exemplo, imagens. Comumente há a ocorrência de overfitting nos classificadores gerados por outros métodos inteligentes sobre esses tipos de dados.

- Teoria bem definida - as SVMs possuem uma base teórica bem estabelecida dentro da Matemática e Estatística.

As SVMs tem como fundamento a determinação de um hiperplano que maximize a margem de separação de um conjunto de dados em classes distintas [54][38]. 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 [78].

Um conjunto de treinamento S é linearmente separável, se for possível separar os padrões das classes diferentes contidos no mesmo por pelo menos um hiperplano. Aos classificadores que realizam a separação dos dados por meio de um hiperplano são denominados lineares. Para simplificar, aqui será considerado o caso de classificação binária, isto é, em duas classes, sendo este modelo linear definido ela equação 30.

$$f(\vec{x}) = \vec{w}^T\vec{x} + b = 0 \tag{30}$$

onde w é o vetor normal ao hiperplano descrito em 30 e $\frac{b}{\|w\|}$ é a distância do hiperplano à origem,conforme a figura 2.9.1.

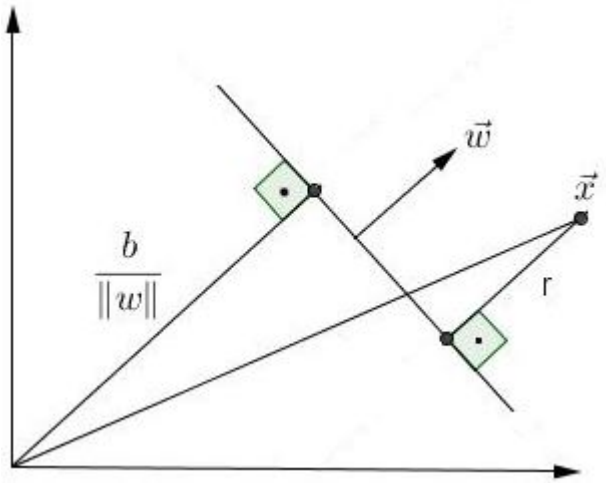


Figura 8: Representação geométrica de um hiperplano de duas dimensões.

O resultado de b resulta da constatação de que a distância r de um ponto à superfície expressa pela equação 30 é determinada por [79].

$$r = \frac{\vec{w}^T\vec{x} + b}{\|\vec{w}\|} \tag{31}$$

Quando utilizado em problemas de classificação de padrões, por exemplo, observa-se que o vetor divide o espaço em dois subespaços, uma para cada classe. De modo que a classificação de cada um será dada de acordo com sua posição com relação as margens de separação, conforme expresso na equação 32.

$$\begin{cases} \vec{w}^T\vec{x} + b \geq 0 \, se \, y_i = +1 \\ \vec{w}^T\vec{x} + b < 0 \, se \, y_i = -1 \end{cases} \tag{32}$$

Os pontos que satisfazem a primeira ou segunda expressão da equação 32 são satisfeito com uma igualdade, dá-se o nome de *vetores de suporte*, ou *support vectors* [54]. Logo, a equação 31 tem-se que as distâncias dos vetores de suporte ao hiperplano de separação podem ser determinadas por [54] 33.

$$\begin{cases} \frac{1}{\|w\|}, se \, y_i = +1 \\ -\frac{1}{\|w\|} \, se \, y_i = -1 \end{cases} \tag{33}$$

De modo que as margens de separação entre as classes, conforme observado na figura abaixo, pode ser expressa pela equação 34, definida como a margem geométrica do classificador [80].

$$\rho = \frac{2}{\|\vec{w}\|} \tag{34}$$

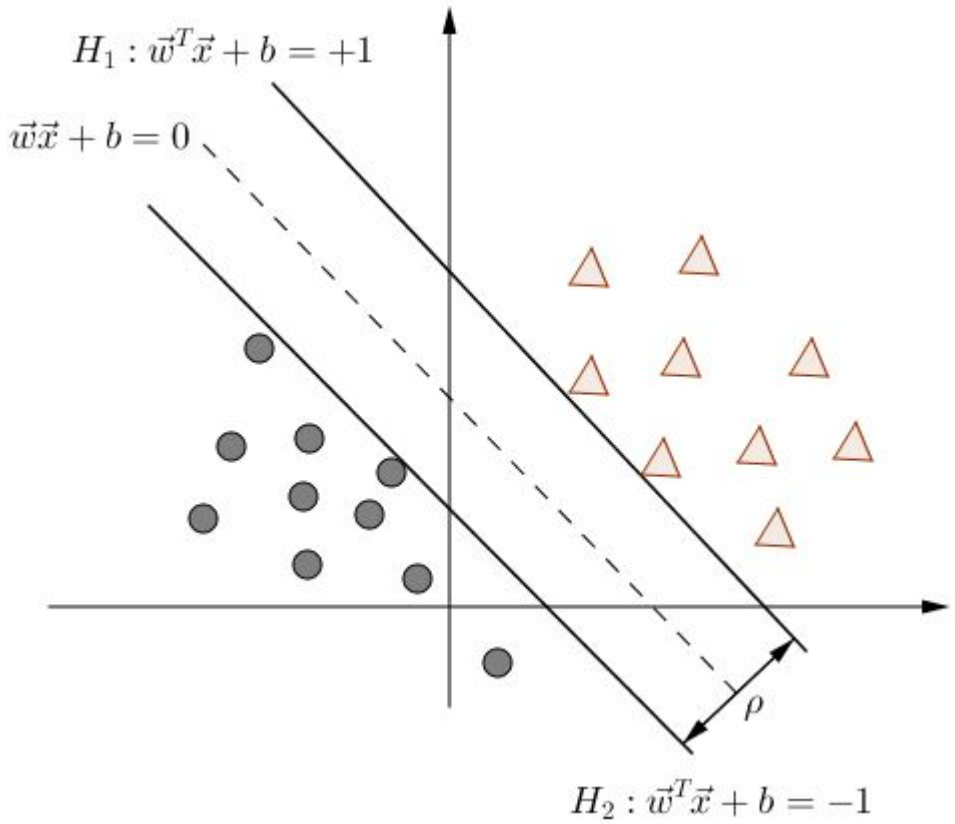


Figura 9: Hiperplano ótimo, com dois vetores de suporte H_1 e H_2.
Fonte: Adaptada de [10]

As SVMs obtêm sua solução através da maximização da margem de separação entre as classes. Com base na equação 34, conclui-se dai que tal resultado pode ser obtido por meio da minimização da margem do vetor de pesos $\|\vec{w}\|$ [81], obtendo o seguinte problema de otimização primal [62]:

$$\begin{gathered} \text{Minimizar através de } (\vec{w}, b) : \tfrac{1}{2}\|\vec{w}\|^2 \\ \text{sujeito a} : \forall_{i=1}^{n} : y_i \left[\vec{w}\,\vec{x_i} + b\right] \geq 1 \end{gathered} \tag{35}$$

A solução direta da equação 35 não é nada fácil, em problemas quadráticos como esse, cuja a função objetiva convexa e os pontos que satisfazem as formam um conjunto convexo de restrições lineares em $\vec{w}$, podem ser solucionadas utilizando o método dos multiplicadores de Lagrange [82], que transforma o problema inicial em um outro problema de mais simples solução. Este método agrega as restrições à função custo, associadas a variáveis auxiliares denominada multiplicadores de Lagrange, que resulta numa função Lagrangeana expressa em 36 [62].

$$J(\vec{w}, b, \vec{x_i}, \vec{\alpha}) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i(\vec{w}^T\vec{x_i} + b) - 1\right] \tag{36}$$

onde os multiplicadores de Lagrange são expressos por α_i.

A equação 36 pode ser minimizada em relação a $\vec{w}$ e b, e maximizada em relação a $\vec{\alpha}$ [83]. Aqui busca-se o ponto de sela onde:

$$\frac{\partial J}{\partial \vec{w}} = 0 \; e \; \frac{\partial J}{\partial b} = 0 \tag{37}$$

Obtendo como solução da equação:

$$\vec{w} = \sum_{i=1}^{N} \alpha_i y_i \vec{x_i} \tag{38}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{39}$$

Assim, substituindo as equações 38 e 39 na equação 36, obtém-se um problema de otimização dual que será representado por:

$$\begin{gathered} \text{Maximizar através de } \vec{\alpha} : \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i \vec{x_i}^T \vec{x_j} \\ \text{sujeito a: } \sum_{i=1}^{N} \alpha_i y_i = 0, \\ \forall i = 1, 2, ..., N \end{gathered} \tag{40}$$

De modo que, por se tratar de um ponto de sela, tem-se que o produto de cada multiplicador de Lagrange e sua correspondente restrição se anula de modo a obter a equação 41:

$$\alpha_i \left[y_i(\vec{w}^T \vec{x_i} + b) - 1 \right] = 0, \forall i = 1, 2, ...N \tag{41}$$

Logo, αi é diferente de zero para os dados que se encontram sobre as margens, de modo a anular o termo entre colchetes na equação 41. Estes pontos são os que estão mais próximos à superfície de separação e dos vetores de suporte, os dados mais importantes do conjunto de treinamento para as SVMs [81]. Os demais pontos podem ser excluídos de acordo com a definição da solução, pois os mesmos não influenciam na formulação da função de decisão.

Os parâmetros do problema primal será obtido através do problema dual. Para se obter $\vec{w^*}$, aplicar a equação 37. A partir dai observa-se que os pontos onde $\alpha \neq 0$, os vetores de suporte, participam do cálculo $\vec{w^*}$, de modo a enfatizar o fato de que os demais pontos podem ser descartados. O valor de b^* é obtido através da equação 41, computando-se a média para todos os vetores de suporte(SVs) [62], de acordo com a equação 42:

$$b^* = \frac{1}{n_{SV}} \sum_{\vec{x} \epsilon SV} \frac{1}{y_j} - \vec{w^*}.\vec{x_j} = \frac{1}{n_{SV}} \sum_{\vec{x_j} \epsilon SV} \left(\frac{1}{y_j} - \sum_{\vec{x_i} \epsilon SV} \alpha_i^* y_i \vec{x_i}.\vec{x_j} \right) \tag{42}$$

Daí, se obtém a função de decisão da SVM 43:

$$f(\vec{x}) = sign(f(x)) = sign\left(\sum_{\vec{x_i}\epsilon SV} \alpha_i^* y_i \vec{x_i}.\vec{x} + b^*\right) \quad (43)$$

2.9.2 Análise Discriminante Linear

A análise discriminante é uma técnica utilizada para discriminar e classificar objetos quando as variáveis independentes são quantitativas (métricas) é um dos métodos estatísticos supervisionados mais conhecidos para classificação de dados [86]. Uma técnica da estatística multivariada que estuda a separação de objetos de uma população em duas ou mais classes [84]. O problema da discriminação entre dois ou mais grupos, objetivando posterior classificação foi inicialmente abordado por Fisher (1936). Consiste em encontrar funções matemáticas capazes de classificar um indivíduo X em uma de varias populações $\pi_i, (i = 1, 2, ..., n)$, com base em medidas de um número p de características, sempre buscando minimizar a probabilidade má classificação, ou seja, minimizar a probabilidade de classificar erroneamente um indivíduo em população π_i, quando realmente pertence a população π_j, $(i = j)$ $i, j = 1, 2, ..., n$ [85].

Tem por objetivo, buscar uma combinação linear de características observadas que apresente o maior poder de discriminação entre as populações, com base em conjunto de treinamento de dados que melhor explique aqueles dados [88, 87]. Esta combinação linear é denominada função discriminante e está função tem uma propriedade muito interessante que é a de minimizar as probabilidades de má classificação, quando as populações são normalmente distribuídas com média μ e variância $\sum$ conhecidas. No entanto, a média e variância das populações não são conhecidas, sendo assim há a necessidade de se estimar estes parâmetros. De acordo com a seleção das funções discriminantes podem ser definidas como lineares ou quadráticas. No caso particular da função de Fisher assume-se que as matrizes de covariância são iguais e é denominada função discriminante linear de Fisher.

A função discriminante é uma combinação de variáveis com os pesos relativos que aperfeiçoa a habilidade dos preditores de diferenciarem entre grupos [71], sendo descrita da seguinte forma:

$$Z_{jk} = a + W_1 x_{1k} + W_2 x2k + ... + W_n x_{nk} \quad (44)$$

Sendo:

Z_{jk} – Escore Z discriminante da função discriminante f para o objeto k;

a – Intercepto;

W_i – Peso discriminante para a variável independente i;

x_{ik} – Variável independente i para o objeto k.

O escore Z é obtido do somatório da multiplicação de cada variável independente por seu peso correspondente. Calculando a média dos escore Z obtém-se o centroide. Existe um centroide para cada grupo identificado e cada centroide indica o local mais

típico de uma unidade pertencente a um grupo identificado. O teste de significância estatística é a medição da distância entre os centroides dos grupos identificados. Quanto mais afastada é a distribuição dos escores discriminantes para os grupos, ou seja, quanto menor a sobreposição dos escores discriminantes, melhor separação dos grupos [72].

A função discriminante é diferente da função discriminante linear de Fisher, uma vez que, enquanto a primeira é utilizada como um meio de facilitar a interpretação dos parâmetros das variáveis explicativas, a função discriminante linear de Fisher é utilizada para classificar as observações nos grupos, assim os valores das variáveis explicativas de uma observação são inseridos nas funções de classificação e, consequentemente, um escore de classificação é calculado para cada grupo, para aquela observação. Dadas as p variáveis e g grupos, é possível estabelecer $m = min(g-1; p)$ funções discriminantes que são combinações lineares das p variáveis, de modo que a função linear de Fisher seja dada por:

$$Z_n = W_1X_1 + W_2X_2 + ... + W_nX_n \tag{45}$$

em que W_i representa o vetor de pesos das variáveis para as funções discriminantes e são estimados de modo que a variabilidade dos escore da função discriminante seja máxima entre os grupos e mínima dentro dos grupos [73].

Os coeficientes $W_1, W_2, W_3, ..., W_n$ podem ser determinados maximizando a razão entre a variabilidade entre as populações e a variabilidade comum dentro das populações, que consiste num problema de maximização de uma razão de formas quadráticas, Quando esta abordagem é usada é possível determinar varias combinações lineares para separar grupos. O número de soluções s disponíveis é o mínimo entre p e $m-1$. Estas combinações lineares são denominadas funções discriminantes canônicas ou eixos discriminante de Fisher [74].

As características das populações são estimadas a partir das correspondentes amostras aleatórias, também denominadas de amostras de treinamento, isto quando as densidades não são conhecidas.

Assim, considere $X_{i1}, X_{i2}, ..., X_{in}$ como uma amostra aleatória da i-ésima população seus estimadores da média amostral da i-ésima população e da média global de todas as populações, respectivamente, em X_{ij} representa a observação da j-ésima unidade amostral aleatória da i-ésima população, são dados pelas equações 46 47 .

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} \tag{46}$$

$$\bar{X}_i = \frac{\sum_{i=1}^{m} n_i \bar{X}_i}{\sum_{i=1}^{m} n_i} \tag{47}$$

Com o objetivo da determinação da regra discriminante a principio são calculadas a matriz de somas de quadrados cruzados dentro da amostra W, dadas por

$$S_b = \sum_{i=1}^{m} n_i(\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T \tag{48}$$

$$S_W = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T \tag{49}$$

Neste momento, necessita-se calcular os autovalores e auto vetores da matriz $S_W^{-1} S_b$. O objetivo é encontrar a razão entre o determinante da matriz S_B e a matriz S_W, conhecido como critério de Fisher [89].

$$Fisher\_criterion = max \frac{|S_b|}{|S_W|}, \tag{50}$$

Logo, a determinação de uma base vetorial W_otm que maximiza o critério de Fisher pode ser resolvido como um problema de autovalores e pode ser determinado conforme equação 51:

$$\left(S_W^{-1} S_B\right) \phi = \Lambda \phi, \tag{51}$$

onde ϕ e a matriz de autovetores e Λ a matriz de autovalores de $S_W^{-1} S_b$.

2.9.3 Overfitting

O *overfitting* ou super ajuste é um fenômeno que surge como resultado de *overtrainig* (super treino), mas não exclusivamente nesses casos, pois, o mesmo pode ocorrer quando utiliza-se muitos parâmetros para determinar um conjunto de características (modelo). E como principal consequência tem-se a memorização pela rede dos padrões, isto é, ela perde a capacidade de generalização.

Para detectar e evitar o overfitting o conjunto de dados deve ser dividido em dois subconjuntos um para treinamento e o outro para os testes, permitindo assim uma avaliação final e a obtenção de uma taxa real de acertos na classificação [75]. Neste trabalho utilizou-se a técnica de validação cruzada para realizar a avaliação do modelo proposto.

2.9.4 Validação Cruzada

É uma técnica que avalia a capacidade de generalização de um dado modelo, a partir de um conjunto de dados. Essa é uma das principais técnicas e é amplamente empregada em problemas onde o objetivo da modelagem é a classificação. Permite determinar a precisão de um modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados.

A técnica consiste na divisão do conjunto de dados em subconjuntos mutuamente exclusivos, e posteriormente, na utilização de alguns teste subconjuntos para o treinamento

e o restante dos subconjuntos para validação do teste. O modelo é avaliado a partir dos resultados obtidos desta combinação.

Existem varias maneiras de realizar a divisão dos dados, as três mais utilizadas são: *holdout*, o *k-fold* e o *leave-one-out* [76].

- *Holdout* este método é simples e consiste em dividir o conjunto de dados em dois subconjuntos mutuamente exclusivos, um para treinamento e outro para teste. O conjunto de dados fornece dados para o treinamento da técnica utilizada e o conjunto de teste fornece dados para o treinamento da técnica utilizada e o conjunto de teste fornece dados novos, para testar a generalização do modelo. Geralmente utiliza-se $\frac{2}{3}$ dos dados para o treinamento e $\frac{1}{3}$ restante para teste [76]. Depois da divisão dos conjuntos, a estimação do modelo é realizada (treinamento) e, posteriormente, os dados de teste são aplicados (validação) e o erro de predição calculado [77].

- *K-fold* este método consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos e do mesmo tamanho e, a partir dai, um subconjunto é utilizado para teste e os $k-1$ restantes são utilizados para estimação dos parâmetros. O processo é realizado k vezes permutando de forma circular o subconjunto de teste. Assim, como exemplo se $k = 5$, será realizado quatro treinos, na primeira vez o primeiro grupo será usado para teste e os outros para treinamento. Enquanto que na segunda vez, o segundo grupo será para teste e os outros quatros serão utilizados para treinamento, e assim ocorrerá de forma análoga até que todas as permutações sejam realizadas [76]. No final das k interações calcula-se a acurácia sobre os erros encontrados.

- *Leave-one-out* o método é uma simplicaçãodo *k-fold*, com k, com k igual ao número total de dados N [76]. Os N padrões são divididos em dois conjuntos, em que o primeiro possui somente um padrão e o segundo com todos os outros restantes $(N-1)$. Testa-se a rede com o primeiro grupo que contem um elemento e treina-se a rede com outros $(N-1)$ padrões, segundo conjunto, e este processo é refeito para todos os padrões do modelo. Neste modelo realiza-se realiza-se N cálculos de erro, uma para cada dado.

3 OBJETIVOS

3.1 Objetivo geral

Propor um método de classificação de câncer de próstata baseado em reconhecimento de padrões proteômicos para auxiliar o profissional da saúde no diagnóstico precoce e eficaz em conjunto com outros métodos já existente, tais como, toque retal e PSA por exemplo.

3.2 Objetivos específicos

- Melhorar a qualidade de vida dos pacientes e diminuir o número de óbitos causados por esta patologia;
- Comparar os resultados obtidos pelos classificadores SVM e LDA quanto a poder de generalização dos resultados e decidir pelo melhor conjunto de técnicas para o método proposto;
- Avaliar a eficiência do método proposto por meio das medidas de acurácia, sensibilidade e especificidade;

4 Materiais e Método

Neste capítulo será realizado a descrição dos materiais utilizados, onde podem ser encontrados e como eles foram obtidos juntamente com o conjunto de técnicas empregadas para a implementação do método proposto que tem por objetivo principal auxiliar os profissionais da saúde no diagnóstico precoce e eficaz do câncer de próstata. O método consiste basicamente em extrair as características dos sinais proteômicos que estão presentes na base de dados [91], pela técnica Análise de Componentes Independentes (ICA), reduzir o conjunto de características, selecionando e extraindo as características irrelevantes para o problema em questão, fazendo uso da técnica Máxima Relevância e Mínima Redundância(mRMR), e classificar os pacientes nos grupos controle (não portador do câncer de próstata) e ativo (portador do câncer de próstata) fazendo uso dos classificadores Máquina de Vetores de Suporte (SVM) e Análise Discriminante Linear (LDA) de modo a comparar os resultados e optar pelo melhor conjunto de técnicas a serem utilizadas com o objetivo de otimizar os resultados.

Os resultados obtidos por todas as técnicas abordadas neste trabalho serão descristo nas subsecções a seguir. O diagrama de blocos do método proposto está descrito na figura [10].

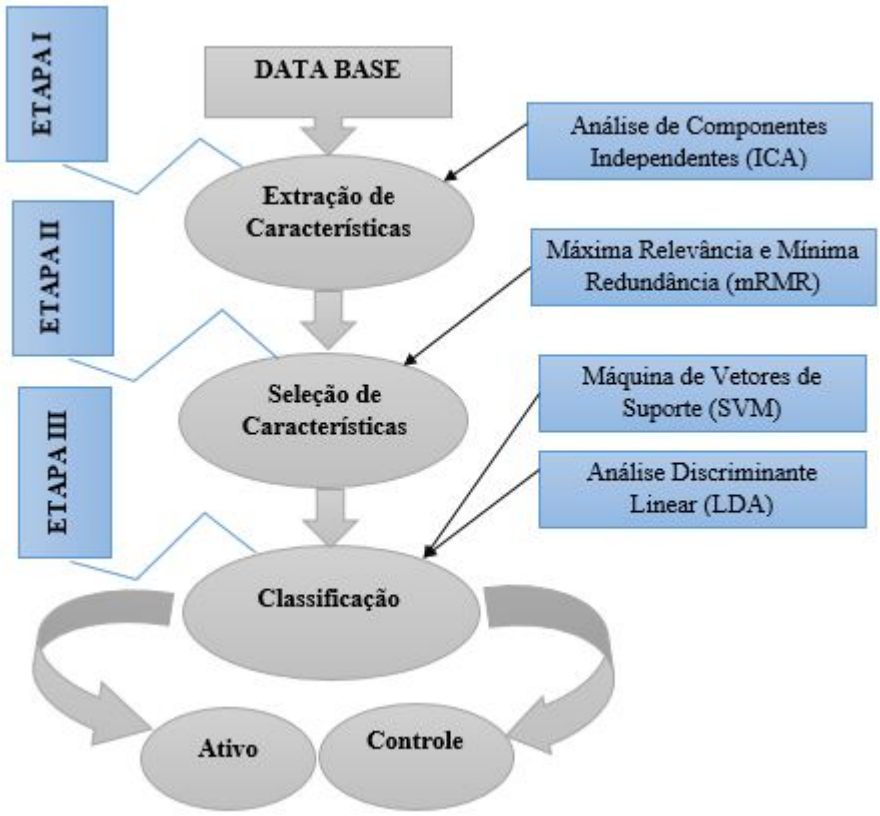


Figura 10: Diagrama de blocos do método proposto

4.1 Base de Dados

Os dados utilizados para aplicação do método proposto é baseado em padrões proteômicos e os mesmo foram obtidos utilizando um espectrômetro de massas fazendo uso da técnica *SELD-TOF*. As bases de dados são públicas, gratuitas, e podem ser obtidas em [Seldi-MS 2002] [91].

Estar disponível para o câncer de próstata quatro bases de dados, e as mesmas foram separadas de acordo com o nível de PSA. A primeira base "No evidence"contém 63 amostras com PSA menor que $1ng/ml$, a segunda "Benign"e contém 190 amostras com PSA menor que $4ng/ml$, a terceira "Prostate Cancer 4-10"e contém 26 amostras com PSA variando entre (4-10) ng/ml, e a quarta base "Prostate Cancer 10"e contém 24 amostras com PSA maior que 10 ng/ml, totalizando 306 casos. Cada um dos casos possui 15.154 níveis de intensidade ou características diferentes. Para a realização deste trabalho utilizou-se a base de dados "Benign"que contém a maior quantidade de amostras sem o câncer de próstata, e juntou-se as amostras das bases "Prostate Cancer 4-10"e "Prostate Cancer 10", amostras com câncer de próstata.

Na figura11 pode-se observar quatro sinais multinível, que foram escolhidos de forma aleatória da base de dados [SELD-MS] [91] e através do software "MatLab"foi plotado seus gráficos, onde dois dos sinais são do grupo controle e os outros dois do grupo ativo. Tem-se ainda que o eixo horizontal corresponde a razão massa carga (m/z) e o eixo vertical à intensidade do sinal.

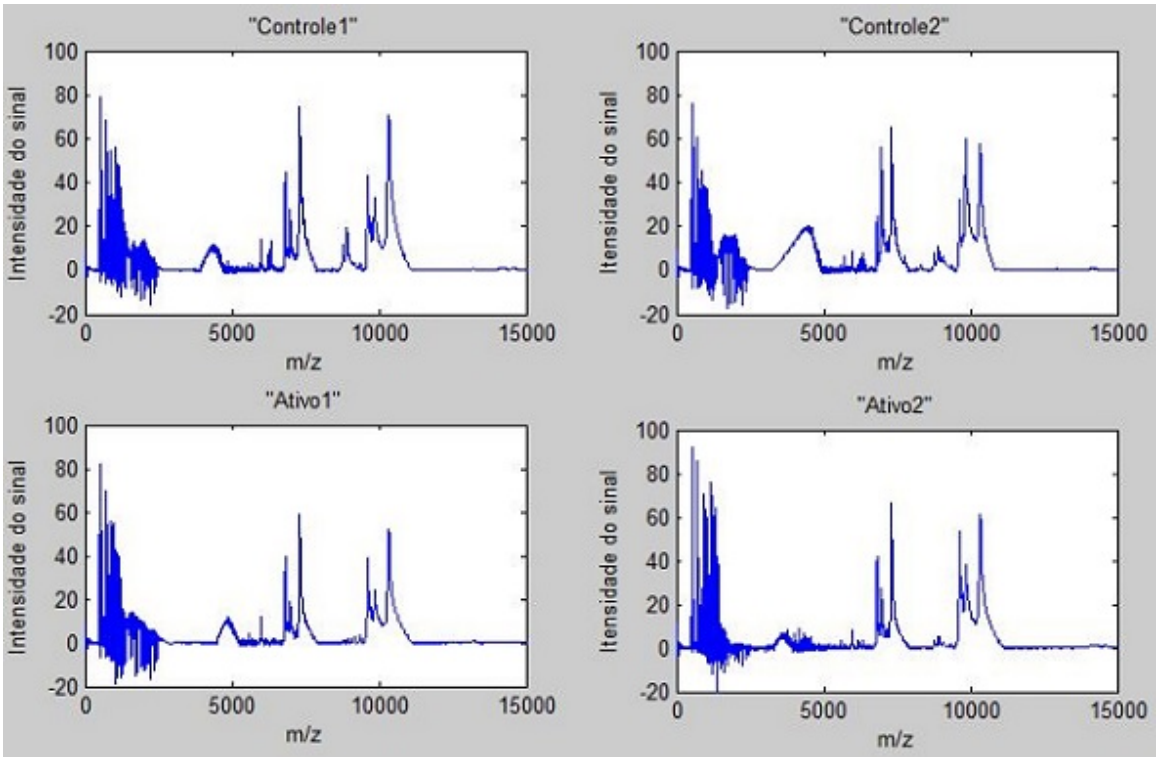


Figura 11: Sinais multinível extraídos da base de dados[91] dos casos ativos e controle.

Observa-se ainda na figura11 que apenas com a representação dos sinais extraídos do espectro de massa, através da comparação visual entre eles, não é possível predizer se um indivíduo é ou não portador do câncer de próstata, sendo necessário a realização de todo o método descrito neste trabalho.

4.2 Extração de Características

Na fase de extração de características aplicou-se a técnica de Análise de Componentes Independentes (ICA), descrita na seção 2.7. Considerando que cada amostra, isto é, cada sinal proteômico é estatisticamente independente e que cada uma seja definida pela função 52:

$$x_i = a_{i1}s_1 + a_{i2}s_2 + ... + a_{in}s_n \tag{52}$$

ou seja, uma combinação linear das outras amostras. A figura 12 ilustra a apresentação de um sinal proteômico como combinação linear de suas características.

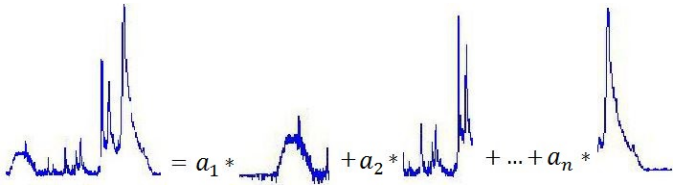


Figura 12: Sinal proteômico como mistura de suas características

Objetivando obter a matriz X do modelo ICA, descrito na fase de extração de características, uniu-se a matriz dos casos controle, de dimensão 190×15154 com a matriz dos casos ativo de 69×15154, gerando assim, a matriz de dimensão 259×15154. Em seguida, utilizou-se a matriz X como entrada para o algoritmo FastICA, objetivando as funções base da matriz quadrada A, de ordem 259×259 que contem todas as características dos sinais proteômicos, onde cada linha representa uma amostra e cada coluna um nível de intensidade do sinal proteômico, isto é, representa as características dos vetores de cada sinal proteômico, parâmetro para o classificador, que permitirá a futura classificação do paciente, com a presença ou ausência do câncer de próstata.

4.3 Seleção das características mais significativa e redução da dimensionalidade

Nesta fase tem-se por objetivo escolher um subconjunto das características mais informativas produzidas a partir dos sinais originais sem que se perca sua capacidade discriminante. Muitas das vezes as características extraídas possuem muita redundância, isto é, informações mútuas já obtidas por outras, então tem-se a necessidade de excluir

essas características redundantes. Esta seleção é importante em casos que a medição da mesma é custosa, pois pode permitir que um subconjunto representativo, e menor que o original seja selecionado, melhorando a qualidade dos dados e o modelo construído, tornando-se mais compreensível, deste modo o processamento se torna mais rápido. Para realizar esta seleção de características utilizou-se o a técnica Máxima Relevância e Minima Redundância, descrita na seção 2.8.1.

4.4 Classificação

Na última fase serão utilizadas duas técnicas de classificação. Máquina de Vetores de Suporte (SVM) e Análise Discriminante Linear (LDA) descritas respectivamente em 2.9.1, 2.9.2 que analisará cada linha da matriz de características, já reduzida através da técnica mRMR, e vai atribuir uma classificação, ou seja, vai rotular nos grupos ativo ou controle. Para fins de comparação, os classificadores receberam o mesmo conjunto de treinamento e teste, escolhidos entre as amostras de forma aleatória. E ainda objetivando aumentar a confiabilidade do método proposto e do classificador em relação á sua capacidade generalização, será aplicado a técnica estatística de validação cruzada 10 - *fold-cross validation* [76].

Em processo de aprendizado de máquinas, mineração de dados, os conjunto de características é dividido em dois subconjuntos, denominados grupo de treinamento e grupo de teste. Na fase inicial um algoritmo de indução de conhecimento é aplicado à base de treinamento. Assim, obtém-se um modelo "treinado", que de certa forma representa o conhecimento extraído. Na fase seguinte o modelo é aplicado no outro subconjunto denominado grupo de testes. Como a base de teste é previamente rotulada, se pode medir a taxa de acerto do modelo, comparando-se o resultado obtido com a rotulação disponível na base de teste [92].

A técnica de Validação Cruzada *10-fold-cross validation* consiste na divisão da base de dados em x partes *(folds)*. Onde, $x - 1$ partes são usadas para treinamento e a outra utilizada para teste. O processo é repetido x vezes, de modo que cada parte seja utilizada uma vez como conjunto de teste. No final, é realizada a correção total calculando a média dos resultados obtidos em cada etapa, obtendo-se assim estimativa de qualidade do modelo permitindo a análise estatística. Assim, dividiu-se as amostras igualmente em 10 subconjuntos, e foram utilizados 9 grupos para treino e um para teste, e este procedimento foi repetido diversas vezes permutando os grupos de teste e treino até que todos fossem testados pelos classificadores, com o objetivo de validar os resultados obtidos.

4.5 Validação do método de classificação

Afim de avaliar o classificador em relação à sua capacidade de generalização utilizou-se algumas medidas estatísticas. 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 [93]. A validade de um teste dar-se em termos quantitativos ou qualitativos, e possui por finalidade diagnosticar um evento ou predizê-lo. Para avaliação de um teste tem-se 4 possíveis interpretações para o resultado do mesmo: duas em que o teste é dito correto e duas que está incorreto, quando o resultado é dado como positivo na presença da doença, resultado verdadeiro positivo (V_P), ou negativo na ausência da doença, resultado verdadeiro negativo (V_N). Assim, também o teste será considerado incorreto quando ele é positivo na ausência da doença, resultado falso positivo (F_P) ou ainda negativo na presença da doença, resultado falso negativo (F_N), este último pode ser considerado um resultado péssimo, pois o mesmo pode fazer com que o paciente doente não inicie o tratamento na fase inicial da patologia, o que prejudicaria seriamente o tratamento, levando a doença a um estágio mais avançado e algumas vezes ao óbito.

Sensibilidade, Especificidade e Acurácia são as medidas mais utilizadas para descrever um sistema de diagnóstico. Sendo a Sensibilidade (S) a capacidade de um teste detectar os indivíduos verdadeiramente positivo,isto é, diagnosticar corretamente os doentes. Especificidade (E) capacidade que o teste tem em detectar os verdadeiros negativos, ou seja, diagnosticar corretamente os indivíduos sadios. Acurácia (A) proporção de acertos, ou seja, o total de verdadeiros positivos e verdadeiros negativos em relação à amostra estudada.

As equações para calcular a sensibilidade, a especificidade e a acurácia são respectivamente [56]:

$$Sen = \frac{V_p}{V_p + F_N} \tag{53}$$

$$Esp = \frac{V_N}{V_N + F_p} \tag{54}$$

$$Acu = \frac{V_p + V_N}{V_p + V_N + F_p + F_N} \tag{55}$$

5 Resultados e Discussões

Neste capítulo será exposto os resultados obtidos em cada etapa do conjunto de técnicas empregadas. O método foi implementado fazendo uso da linguagem MatLab, R2015a.

5.1 Descrição da utilização da base de dados

A base de dados utilizada como já comentado na seção 4.1 , possui quatro diretórios, onde foram separados de acordo com o nível de PSA de cada uma das amostras, dois desses diretórios são de casos de pacientes sem o câncer de próstata e os outros dois com o câncer. Assim, para aplicação do método proposto necessita-se de um conjunto de dados com ausência desta patologia que será denominado grupo controle, e outro com a presença desta patologia que será denominado grupo ativo. De modo que para o grupo controle utilizou-se o diretório com PSA variando entre um e quatro, e o mesmo possui 190 amostras e para o grupo controle uniu-se os dois diretórios com câncer de próstata para se obter um número significativo de amostras para aplicação do presente método proposto.

As amostras de ambos os grupos controle e ativo possuem duas informações cada uma: relação entre a massa de um determinado íon e o número de cargas elementares que ele carrega (m/z) e também a intensidade do sinal do íon, as informações citadas estão separadas por vírgula, conforme Tabela 6.

Tabela 6: Duas amostras parciais obtidas de forma aleatória dos casos controle e ativo respectivamente.

M/Z,intensidade do íon	M/Z,intensidade do íon
-7.8602611e-005,-0.53176471	-7.8602611e-005,-0.55027451
2.1773576e-007,-0.32	2.1773576e-007,-0.40909804
9.6021472e-005,-0.35137255	9.6021472e-005,-0.079686275
0.00036601382,-0.069019608	0.00036601382,-0.63654902
0.00081019477,-0.36705882	0.00081019477,-0.15811765
0.0014285643,0.84078431	0.0014285643,0.50854902
0.0022211225,-2.2729412	0.0022211225,-2.3228235
0.0031878693,-2.665098	0.0031878693,-2.1816471
0.0043288047,-2.5866667	0.0043288047,-2.9973333
0.0056439287,-3.8964706	0.0056439287,-3.7581176

De um modo geral, o grupo que foi denominado ativo ficou com 69 amostras, pacientes com câncer de próstata, e o grupo controle ficou com 190 amostras, pacientes sem o câncer de próstata.

5.1.1 Extração de características

A etapa de extração de características foi realizada fazendo uso da técnica de Análise de Componentes Independentes (ICA), vista na seção 2.7.

Com o objetivo de se obter a matriz X do modelo ICA, primeiro gerou-se duas matrizes, matriz dos casos controle, com os dados dos pacientes não portador do câncer de próstata e a matriz do caso ativo, com os dados dos pacientes portador do câncer.

De modo a se obter as matrizes, ativo de dimensão 69×15154 e controle de dimensão 190×15154. Estas matrizes podem ser vista nas figuras 13(a) e 13(b) respectivamente.

Em seguida com a união das matrizes ativo e controle obteve-se uma matriz X de dimensão 259×15154 denominada matriz de mistura, descrita no modelo ICA. Para a extração de parâmetros através do ICA usou-se o algorítimo fastICA, muito utilizado para fazer separação de fontes cegas (BSS), de modo que as funções bases são estimadas a partir da matriz de mistura X. A figura 14 apresenta de forma parcial a matriz de mistura X que é a união da matrizes do caso ativo com o caso controle.

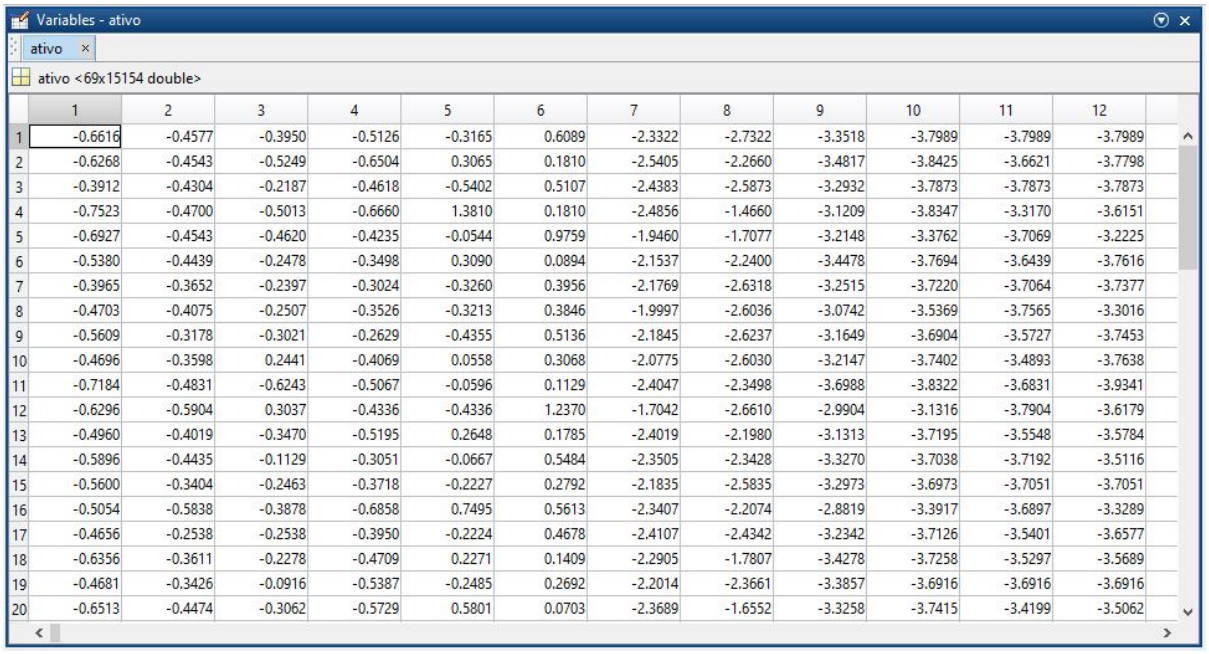
Fazendo uso da matriz X como parâmetro para o algorítimo fastICA , obteve-se as funções de base da matriz A de dimensão 259×259, que contém todas as características de cada amostra. Onde cada linha desta matriz representa uma amostra da base de cados e cada uma das colunas um parâmetro. A figura 15 apresenta uma parte da matriz A, com 20 amostras e 12 características de cada amostra.

5.2 Seleção das características

Como já descrito, o objetivo desta etapa é reduzir o custo computacional e o tempo de processamento, selecionando as características que melhor represente o conjunto de dados após a etapa de extração. Para reduzir a dimensionalidade da matriz de características A, utilizou-se o algorítimo de Máxima Relevância e de Mínima Redundância descrito na seção 2.8.1.

Os testes realizados com o objetivo de reduzir o vetor de características ocorreram incrementando, de um em um, o número de características até o limite de 100, sendo que cada vetor gerado foi testado com o classificador da Máquina de Vetores de Suporte (SVM), com o intuito de encontrar o vetor de melhor desempenho.

Percebeu-se que a partir da centésima característica não ocorria mais variações positivas para os resultados do classificador, e com isso, gerou-se uma nova matriz A_r, de dimensão 259×100, com as características reorganizadas da mais relevante e menos redundante para a menos relevante e mais redundante conforme a figura 16.



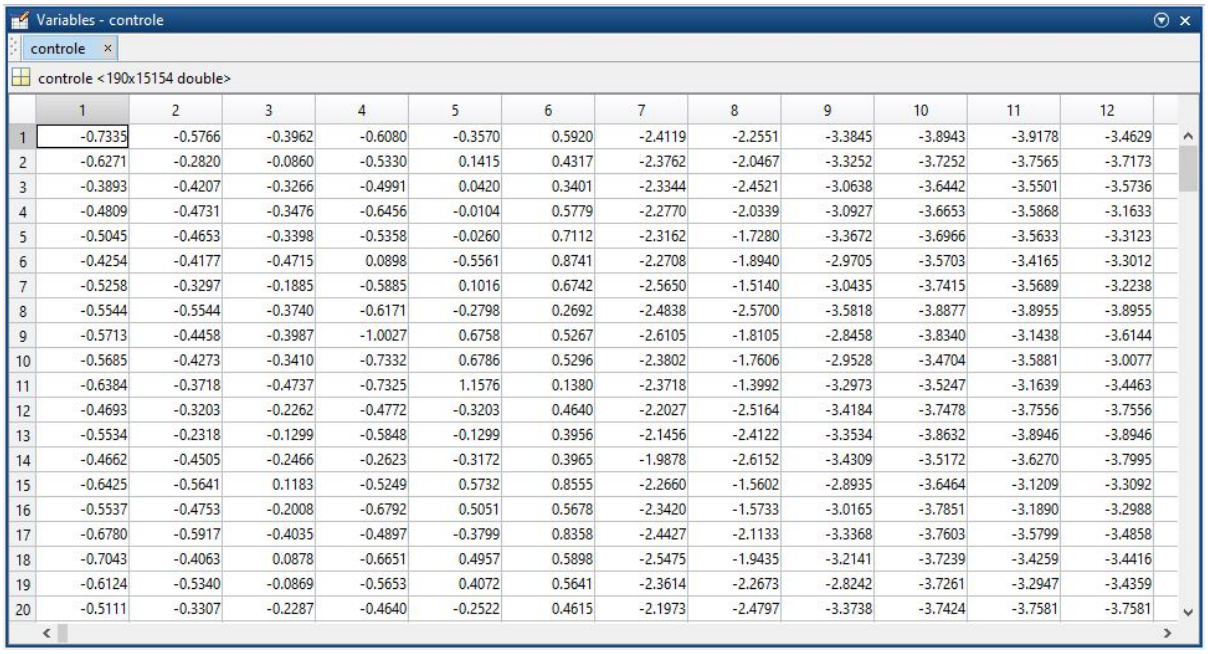
Variables - ativo

ativo

ativo <69x15154 double>

	1	2	3	4	5	6	7	8	9	10	11	12
1	-0.6616	-0.4577	-0.3950	-0.5126	-0.3165	0.6089	-2.3322	-2.7322	-3.3518	-3.7989	-3.7989	-3.7989
2	-0.6268	-0.4543	-0.5249	-0.6504	0.3065	0.1810	-2.5405	-2.2660	-3.4817	-3.8425	-3.6621	-3.7798
3	-0.3912	-0.4304	-0.2187	-0.4618	-0.5402	0.5107	-2.4383	-2.5873	-3.2932	-3.7873	-3.7873	-3.7873
4	-0.7523	-0.4700	-0.5013	-0.6660	1.3810	0.1810	-2.4856	-1.4660	-3.1209	-3.8347	-3.3170	-3.6151
5	-0.6927	-0.4543	-0.4620	-0.4235	-0.0544	0.9759	-1.9460	-1.7077	-3.2148	-3.3762	-3.7069	-3.2225
6	-0.5380	-0.4439	-0.2478	-0.3498	0.3090	0.0894	-2.1537	-2.2400	-3.4478	-3.7694	-3.6439	-3.7616
7	-0.3965	-0.3652	-0.2397	-0.3024	-0.3260	0.3956	-2.1769	-2.6318	-3.2515	-3.7220	-3.7064	-3.7377
8	-0.4703	-0.4075	-0.2507	-0.3526	-0.3213	0.3846	-1.9997	-2.6036	-3.0742	-3.5369	-3.7565	-3.3016
9	-0.5609	-0.3178	-0.3021	-0.2629	-0.4355	0.5136	-2.1845	-2.6237	-3.1649	-3.6904	-3.5727	-3.7453
10	-0.4696	-0.3598	0.2441	-0.4069	0.0558	0.3068	-2.0775	-2.6030	-3.2147	-3.7402	-3.4893	-3.7638
11	-0.7184	-0.4831	-0.6243	-0.5067	-0.0596	0.1129	-2.4047	-2.3498	-3.6988	-3.8322	-3.6831	-3.9341
12	-0.6296	-0.5904	0.3037	-0.4336	-0.4336	1.2370	-1.7042	-2.6610	-2.9904	-3.1316	-3.7904	-3.6179
13	-0.4960	-0.4019	-0.3470	-0.5195	0.2648	0.1785	-2.4019	-2.1980	-3.1313	-3.7195	-3.5548	-3.5784
14	-0.5896	-0.4435	-0.1129	-0.3051	-0.0667	0.5484	-2.3505	-2.3428	-3.3270	-3.7038	-3.7192	-3.5116
15	-0.5600	-0.3404	-0.2463	-0.3718	-0.2227	0.2792	-2.1835	-2.5835	-3.2973	-3.6973	-3.7051	-3.7051
16	-0.5054	-0.5838	-0.3878	-0.6858	0.7495	0.5613	-2.3407	-2.2074	-2.8819	-3.3917	-3.6897	-3.3289
17	-0.4656	-0.2538	-0.2538	-0.3950	-0.2224	0.4678	-2.4107	-2.4342	-3.2342	-3.7126	-3.5401	-3.6577
18	-0.6356	-0.3611	-0.2278	-0.4709	0.2271	0.1409	-2.2905	-1.7807	-3.4278	-3.7258	-3.5297	-3.5689
19	-0.4681	-0.3426	-0.0916	-0.5387	-0.2485	0.2692	-2.2014	-2.3661	-3.3857	-3.6916	-3.6916	-3.6916
20	-0.6513	-0.4474	-0.3062	-0.5729	0.5801	0.0703	-2.3689	-1.6552	-3.3258	-3.7415	-3.4199	-3.5062

(a) Ativo



Variables - controle

controle

controle <190x15154 double>

	1	2	3	4	5	6	7	8	9	10	11	12
1	-0.7335	-0.5766	-0.3962	-0.6080	-0.3570	0.5920	-2.4119	-2.2551	-3.3845	-3.8943	-3.9178	-3.4629
2	-0.6271	-0.2820	-0.0860	-0.5330	0.1415	0.4317	-2.3762	-2.0467	-3.3252	-3.7252	-3.7565	-3.7173
3	-0.3893	-0.4207	-0.3266	-0.4991	0.0420	0.3401	-2.3344	-2.4521	-3.0638	-3.6442	-3.5501	-3.5736
4	-0.4809	-0.4731	-0.3476	-0.6456	-0.0104	0.5779	-2.2770	-2.0339	-3.0927	-3.6653	-3.5868	-3.1633
5	-0.5045	-0.4653	-0.3398	-0.5358	-0.0260	0.7112	-2.3162	-1.7280	-3.3672	-3.6966	-3.5633	-3.3123
6	-0.4254	-0.4177	-0.4715	0.0898	-0.5561	0.8741	-2.2708	-1.8940	-2.9705	-3.5703	-3.4165	-3.3012
7	-0.5258	-0.3297	-0.1885	-0.5885	0.1016	0.6742	-2.5650	-1.5140	-3.0435	-3.7415	-3.5689	-3.2238
8	-0.5544	-0.5544	-0.3740	-0.6171	-0.2798	0.2692	-2.4838	-2.5700	-3.5818	-3.8877	-3.8955	-3.8955
9	-0.5713	-0.4458	-0.3987	-1.0027	0.6758	0.5267	-2.6105	-1.8105	-2.8458	-3.8340	-3.1438	-3.6144
10	-0.5685	-0.4273	-0.3410	-0.7332	0.6786	0.5296	-2.3802	-1.7606	-2.9528	-3.4704	-3.5881	-3.0077
11	-0.6384	-0.3718	-0.4737	-0.7325	1.1576	0.1380	-2.3718	-1.3992	-3.2973	-3.5247	-3.1639	-3.4463
12	-0.4693	-0.3203	-0.2262	-0.4772	-0.3203	0.4640	-2.2027	-2.5164	-3.4184	-3.7478	-3.7556	-3.7556
13	-0.5534	-0.2318	-0.1299	-0.5848	-0.1299	0.3956	-2.1456	-2.4122	-3.3534	-3.8632	-3.8946	-3.8946
14	-0.4662	-0.4505	-0.2466	-0.2623	-0.3172	0.3965	-1.9878	-2.6152	-3.4309	-3.5172	-3.6270	-3.7995
15	-0.6425	-0.5641	0.1183	-0.5249	0.5732	0.8555	-2.2660	-1.5602	-2.8935	-3.6464	-3.1209	-3.3092
16	-0.5537	-0.4753	-0.2008	-0.6792	0.5051	0.5678	-2.3420	-1.5733	-3.0165	-3.7851	-3.1890	-3.2988
17	-0.6780	-0.5917	-0.4035	-0.4897	-0.3799	0.8358	-2.4427	-2.1133	-3.3368	-3.7603	-3.5799	-3.4858
18	-0.7043	-0.4063	0.0878	-0.6651	0.4957	0.5898	-2.5475	-1.9435	-3.2141	-3.7239	-3.4259	-3.4416
19	-0.6124	-0.5340	-0.0869	-0.5653	0.4072	0.5641	-2.3614	-2.2673	-2.8242	-3.7261	-3.2947	-3.4359
20	-0.5111	-0.3307	-0.2287	-0.4640	-0.2522	0.4615	-2.1973	-2.4797	-3.3738	-3.7424	-3.7581	-3.7581

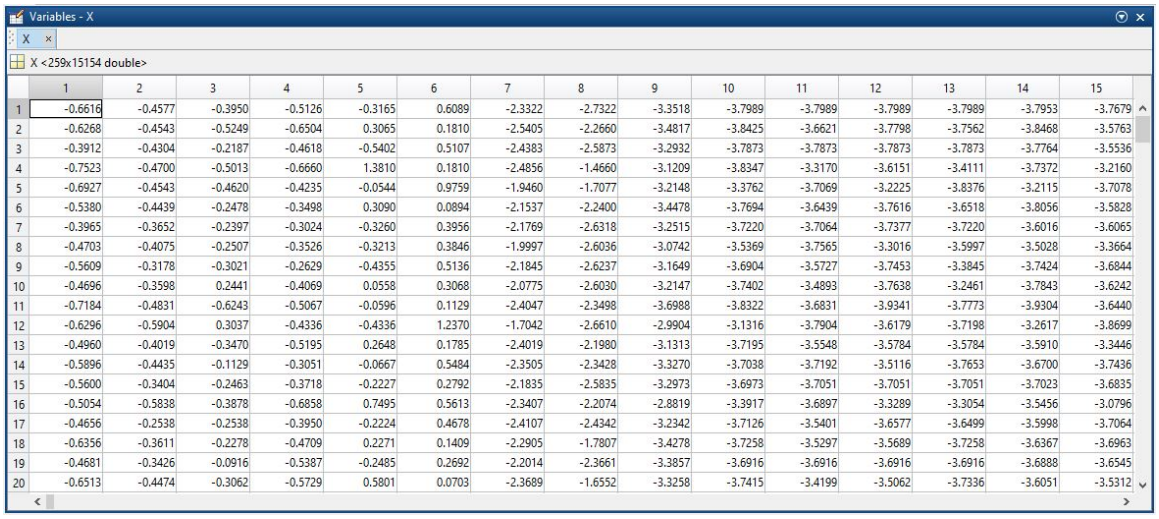
(b) Controle

Figura 13: Matrizes geradas da base de dados dos pacientes do grupo ativo e grupo controle respectivamente.

5.3 Classificação pela SVM e validação do método

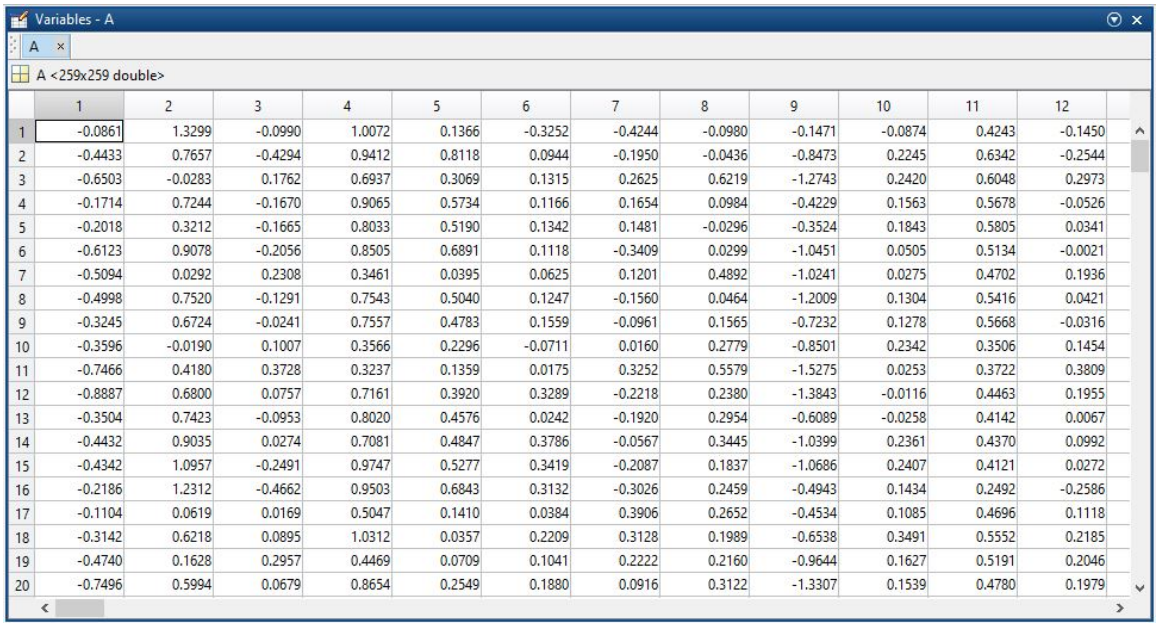
A etapa de classificação analisa cada vetor de característica, isto é, as linhas da matriz A_r, e após essa análise pode realizar-se a sua devida classificação: em controle ou ativo dos indivíduos.

Com o objetivo de aumentar a confiabilidade nos resultados do classificador, usou-se uma técnica estatística de validação cruzada *10-fold-cross validation*, na qual divide-se



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-0.6616	-0.4577	-0.3950	-0.5126	-0.3165	0.6089	-2.3322	-2.7322	-3.3518	-3.7989	-3.7989	-3.7989	-3.7989	-3.7953	-3.7679
2	-0.6268	-0.4543	-0.5249	-0.6504	0.3065	0.1810	-2.5405	-2.2660	-3.4817	-3.8425	-3.6621	-3.7798	-3.7562	-3.8468	-3.5763
3	-0.3912	-0.4304	-0.2187	-0.4618	-0.5402	0.5107	-2.4383	-2.5873	-3.2932	-3.7873	-3.7873	-3.7873	-3.7873	-3.7764	-3.5536
4	-0.7523	-0.4700	-0.5013	-0.6660	1.3810	0.1810	-2.4856	-1.4660	-3.1209	-3.8347	-3.3170	-3.6151	-3.4111	-3.7372	-3.2160
5	-0.6927	-0.4543	-0.4620	-0.4235	-0.0544	0.9759	-1.9460	-1.7077	-3.2148	-3.3762	-3.7069	-3.2225	-3.8376	-3.2115	-3.7078
6	-0.5380	-0.4439	-0.2478	-0.3498	0.3090	0.0894	-2.1537	-2.2400	-3.4478	-3.7694	-3.6439	-3.7616	-3.6518	-3.8056	-3.5828
7	-0.3965	-0.3652	-0.2397	-0.3024	-0.3260	0.3956	-2.1769	-2.6318	-3.2515	-3.7220	-3.7064	-3.7377	-3.7220	-3.6016	-3.6065
8	-0.4703	-0.4075	-0.2507	-0.3526	-0.3213	0.3846	-1.9997	-2.6036	-3.0742	-3.5369	-3.7565	-3.3016	-3.5997	-3.5028	-3.3664
9	-0.5609	-0.3178	-0.3021	-0.2629	-0.4355	0.5136	-2.1845	-2.6237	-3.1649	-3.6904	-3.5727	-3.7453	-3.3845	-3.7424	-3.6844
10	-0.4696	-0.3598	0.2441	-0.4069	0.0558	0.3068	-2.0775	-2.6030	-3.2147	-3.7402	-3.4893	-3.7638	-3.2461	-3.7843	-3.6242
11	-0.7184	-0.4831	-0.6243	-0.5067	-0.0596	0.1129	-2.4047	-2.3498	-3.6988	-3.8322	-3.6831	-3.9341	-3.7773	-3.9304	-3.6440
12	-0.6296	-0.5904	0.3037	-0.4336	-0.4336	1.2370	-1.7042	-2.6610	-2.9904	-3.1316	-3.7904	-3.6179	-3.7198	-3.2617	-3.8699
13	-0.4960	-0.4019	-0.3470	-0.5195	0.2648	0.1785	-2.4019	-2.1980	-3.1313	-3.7195	-3.5548	-3.5784	-3.5784	-3.5910	-3.3446
14	-0.5896	-0.4435	-0.1129	-0.3051	-0.0667	0.5484	-2.3505	-2.3428	-3.3270	-3.7038	-3.7192	-3.5116	-3.7653	-3.6700	-3.7436
15	-0.5600	-0.3404	-0.2463	-0.3718	-0.2227	0.2792	-2.1835	-2.5835	-3.2973	-3.6973	-3.7051	-3.7051	-3.7051	-3.7023	-3.6835
16	-0.5054	-0.5838	-0.3878	-0.6858	0.7495	0.5613	-2.3407	-2.2074	-2.8819	-3.3917	-3.6897	-3.3289	-3.3054	-3.5456	-3.0796
17	-0.4656	-0.2538	-0.2538	-0.3950	-0.2224	0.4678	-2.4107	-2.4342	-3.2342	-3.7126	-3.5401	-3.6577	-3.6499	-3.5998	-3.7064
18	-0.6356	-0.3611	-0.2278	-0.4709	0.2271	0.1409	-2.2905	-1.7807	-3.4278	-3.7258	-3.5297	-3.5689	-3.7258	-3.6367	-3.6963
19	-0.4681	-0.3426	-0.0916	-0.5387	-0.2485	0.2692	-2.2014	-2.3661	-3.3857	-3.6916	-3.6916	-3.6916	-3.6916	-3.6888	-3.6545
20	-0.6513	-0.4474	-0.3062	-0.5729	0.5801	0.0703	-2.3689	-1.6552	-3.3258	-3.7415	-3.4199	-3.5062	-3.7336	-3.6051	-3.5312

Figura 14: Matriz de mistura X (parcial) da união dos casos controle com o casos ativo.



	1	2	3	4	5	6	7	8	9	10	11	12
1	-0.0861	1.3299	-0.0990	1.0072	0.1366	-0.3252	-0.4244	-0.0980	-0.1471	-0.0874	0.4243	-0.1450
2	-0.4433	0.7657	-0.4294	0.9412	0.8118	0.0944	-0.1950	-0.0436	-0.8473	0.2245	0.6342	-0.2544
3	-0.6503	-0.0283	0.1762	0.6937	0.3069	0.1315	0.2625	0.6219	-1.2743	0.2420	0.6048	0.2973
4	-0.1714	0.7244	-0.1670	0.9065	0.5734	0.1166	0.1654	0.0984	-0.4229	0.1563	0.5678	-0.0526
5	-0.2018	0.3212	-0.1665	0.8033	0.5190	0.1342	0.1481	-0.0296	-0.3524	0.1843	0.5805	0.0341
6	-0.6123	0.9078	-0.2056	0.8505	0.6891	0.1118	-0.3409	0.0299	-1.0451	0.0505	0.5134	-0.0021
7	-0.5094	0.0292	0.2308	0.3461	0.0395	0.0625	0.1201	0.4892	-1.0241	0.0275	0.4702	0.1936
8	-0.4998	0.7520	-0.1291	0.7543	0.5040	0.1247	-0.1560	0.0464	-1.2009	0.1304	0.5416	0.0421
9	-0.3245	0.6724	-0.0241	0.7557	0.4783	0.1559	-0.0961	0.1565	-0.7232	0.1278	0.5668	-0.0316
10	-0.3596	-0.0190	0.1007	0.3566	0.2296	-0.0711	0.0160	0.2779	-0.8501	0.2342	0.3506	0.1454
11	-0.7466	0.4180	0.3728	0.3237	0.1359	0.0175	0.3252	0.5579	-1.5275	0.0253	0.3722	0.3809
12	-0.8887	0.6800	0.0757	0.7161	0.3920	0.3289	-0.2218	0.2380	-1.3843	-0.0116	0.4463	0.1955
13	-0.3504	0.7423	-0.0953	0.8020	0.4576	0.0242	-0.1920	0.2954	-0.6089	-0.0258	0.4142	0.0067
14	-0.4432	0.9035	0.0274	0.7081	0.4847	0.3786	-0.0567	0.3445	-1.0399	0.2361	0.4370	0.0992
15	-0.4342	1.0957	-0.2491	0.9747	0.5277	0.3419	-0.2087	0.1837	-1.0686	0.2407	0.4121	0.0272
16	-0.2186	1.2312	-0.4662	0.9503	0.6843	0.3132	-0.3026	0.2459	-0.4943	0.1434	0.2492	-0.2586
17	-0.1104	0.0619	0.0169	0.5047	0.1410	0.0384	0.3906	0.2652	-0.4534	0.1085	0.4696	0.1118
18	-0.3142	0.6218	0.0895	1.0312	0.0357	0.2209	0.3128	0.1989	-0.6538	0.3491	0.5552	0.2185
19	-0.4740	0.1628	0.2957	0.4469	0.0709	0.1041	0.2222	0.2160	-0.9644	0.1627	0.5191	0.2046
20	-0.7496	0.5994	0.0679	0.8654	0.2549	0.1880	0.0916	0.3122	-1.3307	0.1539	0.4780	0.1979

Figura 15: Matriz de características A (Parcial) gerada pelo algorítimo fastICA.

as amostras igualmente em 10 grupos, utilizando 9 grupos para treino e 1 para teste, esse procedimento foi repetido várias vezes, permutando os grupos até que todos fossem utilizados.

A média dos melhores resultados encontrados pela técnica de validação cruzada em conjunto com a SVM pode ser vista na tabela 7. Observa-se que os vetores com os melhores desempenhos foram os de 26, 27 e 30 características, durante a fase de classificação pela SVM.

Assim, considerando o vetor de 27 características, representado na linha 3 da tabela 1, conclui-se que das 190 amostras sem o câncer de próstata, 183 foram classificadas corretamente (Verdadeiro Negativo), errando apenas dezessete casos, dos quais os indivíduos

Variables - Ar

Ar

Ar <259x100 double>

	1	2	3	4	5	6	7	8	9
1	0.0104	-0.1544	-0.0730	0.2623	-0.4607	-0.4598	0.4243	0.4805	-1.0687
2	0.2583	0.1193	-0.6429	-0.0526	-0.4705	-1.1512	0.6342	-0.0841	0.0167
3	0.4838	0.0764	-0.4699	-1.0270	0.2526	-0.6882	0.6048	-0.4087	0.0353
4	0.1003	0.1178	-0.2902	-0.2315	-0.3317	-0.7840	0.5678	0.3285	-0.3061
5	0.1338	0.1664	-0.5225	0.1104	-0.2883	-0.7486	0.5805	0.1165	-0.0520
6	0.4569	0.1827	-0.6276	-0.5810	-0.6714	-0.7075	0.5134	0.2261	-0.1967
7	0.5328	0.1688	-0.4619	-0.9011	0.0609	-0.7622	0.4702	-0.0456	0.0679
8	0.6009	0.0581	-0.6970	-0.5415	-0.6458	-0.7296	0.5416	0.0869	-0.2414
9	0.2636	0.0142	-0.6094	-0.5638	-0.5856	-0.6189	0.5668	0.0615	-0.3541
10	0.4379	0.3084	-0.3751	-0.6709	0.2012	-0.5553	0.3506	-0.3098	0.2222
11	0.7976	-0.0434	-0.0784	-1.3480	0.4297	-0.5089	0.3722	0.2518	0.2819
12	0.7181	0.1545	-0.3458	-1.0719	-0.5797	-0.7192	0.4463	0.5655	0.2300
13	0.3407	0.1051	-0.7638	-0.5299	-0.6203	-0.6254	0.4142	0.0831	-0.5976
14	0.4246	0.0515	-0.5829	-0.6183	-0.5517	-0.7126	0.4370	0.4612	-0.0949
15	0.3544	-0.1854	-0.4210	-0.3328	-0.7900	-0.7010	0.4121	0.5124	-0.5665
16	0.1255	0.0058	-0.1886	-0.4772	-0.6317	-0.9522	0.2492	0.4521	-0.6646
17	0.2216	0.0764	-0.3997	-0.2134	0.3155	-0.6315	0.4696	0.0099	0.2222
18	0.2196	0.1673	0.0584	-0.3479	-0.0211	-0.7126	0.5552	0.5538	0.3455
19	0.4891	0.1535	-0.2989	-0.6251	0.3025	-0.5173	0.5191	0.3495	0.2306
20	0.5328	0.0700	-0.2745	-0.7740	-0.4208	-0.5600	0.4780	0.3067	-0.1003

Figura 16: Matriz de características A_r (parcial),com as características reorganizadas da mais relevante e menos redundante para a menos relevante e mais redundante.

eram do grupo controle, e foram classificados no grupo ativo; já nos 69 casos com o câncer de próstata, 52 foram classificadas corretamente quanto a presença da doença (Verdadeiros positivos), obtendo métricas para acurácia, sensibilidade e especificidade, respectivamente de 89,21%, 83,68% e 95,08%.

5.3.1 Classificação pela LDA e validação do método

A classificação pela LDA também analisa cada uma das amostras, isto é, cada vetor de características e logo em seguida realiza a classificação das amostras nos seus respectivos grupos controle ou ativo. Assim como na classificação com SVM, também usou-se a técnica de Validação Cruzada para validar o bom desempenho do método aplicado. Deste modo os resultados obtidos foram melhorando com o aumento do número de características conforme descrito na Tabela 8, onde os mesmos também são avaliados pelas métricas de validação sensibilidade, especificidade e acurácia obtendo respectivamente 100%, 100% e 100%.

Considerando o vetor de 77 características, observa-se que as 69 amostras com câncer de próstata, todas foram classificadas corretamente no grupo ativo (VP), nenhuma no grupo controle (FP) e das 190 amostras sem o câncer de próstata todas foram classificadas corretamente no grupo controle (VN), nenhuma no grupo ativo (FP) tendo uma margem

Tabela 7: Desempenho do classificador para os melhores resultados do método proposto

Características	VP	FP	VN	FN	Acurácia%	Sensibilidade%	Especificidade%
25	48	8	182	21	88,81	80,24	94,56
26	50	6	184	19	88,80	84,28	93,76
27	**52**	**7**	**183**	**17**	**89,21**	**83,68**	**95,08**
28	49	9	181	20	88,81	83,63	95,00
29	48	9	181	21	88,45	80,43	95,04
30	49	7	183	20	89,62	81,99	95,30

Tabela 8: Resultados obtidos pela LDA em ordem crescente do número de características.

Características	VP	FP	VN	FN	Acurácia%	Sensibilidade%	Especificidade%
10	48	11	179	21	87,64	89,50	81,35
20	58	9	181	11	92,27	94,27	86,56
30	60	6	184	9	94,21	95,34	90,91
40	61	8	182	8	93,82	95,79	88,41
50	64	4	186	5	96,52	97,38	94,12
60	65	3	187	4	97,30	97,90	95,59
70	68	2	188	1	98,84	99,47	97,14
77	**69**	**0**	**190**	**0**	**100**	**100**	**100**
78	69	0	190	0	100	100	100

de acerto de 100% em ambos os grupos o que mostra um excelente resultado atingido pela aplicação da LDA. Observa-se que com o vetor de 77 características a LDA acertou todos os resultados tanto para o grupo controle quanto para o grupo ativo.

6 Conclusão

Neste trabalho foi realizada uma aplicação de reconhecimento de padrões proteômicos auxiliado por técnicas computacionais, objetivando o diagnóstico precoce do câncer de próstata.

As técnicas computacionais aqui empregadas mostraram-se bastante eficazes como a Análise de Componentes Independentes (ICA) na fase de extração das características dos sinais obtidos através de um espectro de massas, pela técnica *SELDI-TOF*, assim como a técnica de seleção de atributos Máxima Relevância e Mínima Redundância (mRMR) que selecionou as melhores características e mostrou que a redução do conjunto de característica não afetou de forma negativa os resultados, e, ainda diminuiu o custo computacional e agilizou o processo, permitindo um bom desempenho dos classificadores Máquina de Vetores de Suporte (SVM) e Análise Discriminante Linear (LDA), que foram capazes de realizar a classificação de forma eficiente dos pacientes nos grupos controle e ativo.

De um modo geral realizou-se ainda um estudo comparativo na etapa de classificação com base na revisão bibliográfica e nos resultados obtidos, com o objetivo de avaliar

o desempenho de diferentes abordagens do método proposto. Com base neste estudo, levantou-se a hipótese de que o uso em conjunto das técnicas ICA, mRMR e LDA podem compor de modo satisfatório um método que auxilia no diagnóstico precoce do câncer de próstata, baseado nos resultados.

A capacidade de generalização da LDA mostrou-se neste problema muito mais eficaz que a SVM, os resultados apresentados mostram que a LDA teve 100% de acerto em ambos os casos, na classificação das amostras do grupo controle, assim como no grupo ativo. Enquanto que a SVM estimasse que a mesma teve uma capacidade de generalização de apenas 75,36% de acertos com relação a classificação das amostras do grupo ativo e 96,31% com relação a classificação do grupo controle. Em compensação, durante o uso da LDA necessitou-se utilizar uma maior quantidade de características (no mínimo 72 para se obter o melhor desempenho do classificador) tornando o processo mais robusto enquanto que com a SVM o melhor resultado foi obtido com um vetor com apenas 27 características (vetor de características ótimo).

Futuros estudos e novas aplicações destes classificadores em outras bases de dados mais complexas são necessárias para que se possa estimar ou até mesmo validar a hipótese de que a LDA tem melhor capacidade de generalização quando se trabalhar com sinais proteômicos e o conjunto de técnicas ICA e mRMR. Supõe-se que o excelente resultado obtido pela LDA na etapa de classificação ocorreu devido ao fato de que o conjunto das técnicas empregadas se correlacionaram muito bem, isto é, a ICA extraiu de forma satisfatória as características e o mRMR selecionou o conjunto ótimo de características para a análise discriminante Linear permitindo um bom desempenho deste classificador.

O toque retal é o teste mais utilizado no diagnóstico do câncer de próstata, apesar de suas limitações, uma vez que somente as porções posterior e lateral da próstata podem ser palpadas, deixando de 40% a 50% dos tumores fora do seu alcance. As estimativas de sensibilidade variam entre 55% e 68%. O valor preditivo positivo é estimado entre 25% e 28%. Quando utilizado em associação à dosagem do PSA com valores entre 1,5 ng/ml e 2,0 ng/ml, sua sensibilidade pode chegar a 95% [6].

Os resultados aqui apresentados demonstram que o método proposto alcançou um bom desempenho, se comparado com o principal teste utilizado nos dias atuais: o toque retal, mesmo quando realizado em conjunto com outros teste como o PSA, mas deve-se levar em consideração o tabu ainda existente entre os homens e o teste do toque retal, cercado, ainda no século XXI, de muitos preconceitos, levando a comunidade masculina ao adiamento deste teste e até mesmo a sua não realização, quando ocorre a suspeita da presença desta doença, o que pode gerar complicações futuras para o tratamento, prejudicando na qualidade de vida do paciente e familiares.

Os resultados apresentados neste trabalho nos encorajam a continuar as pesquisas com uma perspectiva de novas aplicações em outras bases de dados para obter um melhor modelo do problema e posteriormente o desenvolvimento de um *software* que possa

auxiliar os profissionais da saúde num diagnóstico precoce e eficaz, diminuindo, com isso, os impactos negativos causados pelo câncer de próstata e até mesmo outras doenças.

REFERÊNCIAS

[1] BERTOLDO, Sandra Alves; PASQUINI, Valdiléia Zorub. Câncer de Próstata: um desafio para saúde do homem. Rev. Enfermagem Unisa, v. 11, n. 2, p. 138-142, 2010. Disponível em: <http://www.unisa.br/graduacao/biologicas/enfer/revista/arquivos/2010-2-15.pdf>.

[2] ESTIMATIVA, 2016: Incidência de câncer no Brasil /Instituto Nacional de Câncer. José Alencar Gomes Silva - Rio de Janeiro: INCA, 2015. Disponível em: <http://www.inca.gov.br/estimativa/2016/> Acesso: Mai. 2016.

[3] INCA, Informativo: DETECÇÃO PRECOCE, Instituto Nacional de Câncer José de Alencar Gomes da Silva. Boletim ano 5, n.2 maio/agosto 2014. Disponível em: <http://controlecancer.bvs.br/>. Acesso: Jan. 2016.

[4] SROUGI, Miguel. Próstata: isso é com você. Publifolha, 2003.

[5] SOCIEDADE BRASILEIRA DE UROLOGIA. Câncer da próstata: consenso – Rio de Janeiro: INCA, 2002. 20p.

[6] INCA 2016, Câncer de próstata. Instituto Nacional do Câncer [Online] . Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/prostata>.

[7] GOMES, Romeu; DO NASCIMENTO, Elaine Ferreira; DE ARAÚJO, Fábio Carvalho. Por que os homens buscam menos os serviços. Cad. saúde pública, v. 23, n. 3, p. 565-574, 2007.

[8] FILHO RTF e, DAMIÃO R. Câncer de próstata. Revista Hospital Universitário Pedro Ernesto. 2010;9 (Supl. 1):20-27. Disponível em: <http://revista.hupe.uerj.br/detalhe_artigo.asp?id=249>.

[9] INCA 2016, Câncer. Instituto Nacional de Câncer José de Alencar Gomes da Silva. Boletim ano 5, n.2 maio/agosto 2014. Disponível em: <http://www1.inca.gov.br/conteudo_view.asp?id=322>.

[10] CAMPOS, L. F. A. Método de Detecção de Câncer em Mamas Densas Utilizando Diagnóstico Auxiliado por Computador. Tese (Doutorado em Biotecnologia) Universidade Federal do Maranhão, São Luís, 2013.

[11] ONCOGUIA 2015. ONG Instituto de Oncoguia [Online]. Disponível em: <http://www.oncoguia.org.br/conteudo/cancer/12/1/>.

[12] SOARES, D. A. S. Câncer de próstata para a realização do toque retal. Monografia (Especialização em Atenção Básica em Saúde da Família) Universidade Federal de Minas Gerais, Governador Valadares, 2014.

[13] DA SILVA FRANCA, Carlos Antônio; VIEIRA, Sérgio Lannes; PENNA, Antônio Belmiro Rodrigues Campbell. Definição de recidiva bioquímica após tratamento radioterápico do câncer de próstata localizado: revisão de literatura. Revista Brasileira de Cancerologia, v. 54, n. 1, p. 57-61, 2008.

[14] MOYER V. A. Screening for Prostate Cancer: U.S. Preventive Services Task Force Recommendation Statement. Annals of Internal Medicine; vol.157, 2: 120-134, 2012.

[15] DA CRUZ, Fernanda SalvatoI Mayra Costa; DE CARVALHOII, Gallo. Métodos e estratégias em proteômica e suas aplicações na área vegetal. Ciência Rural, v. 40, n. 3, p. 727-734, 2010.

[16] EMIDIO, Nayara Braga et al. PROTEÔMICA: UMA INTRODUÇÃO AOS MÉTODOS E APLICAÇÕES. HU Revista, v. 41, n. 3 e 4, 2015.

[17] HEIN, M. Y. et al. Chapter 1 - Proteomic Analysis of Cellular Systems. In: Walhout, A. J. M., Vidal, M., et al.Handbook of Systems Biology. San Diego: Academic Press, 2013, p.3-25.

[18] CARVALHO, Paulo Costa et al. Máquina de agrupamento por elipsóide: uma linha de frente para auxiliar no diagnóstico de doenças. Revista Eletrônica de Comunicação, Informação & Inovação em Saúde, v. 1, n. 2, 2007.

[19] BARBOSA, Eduardo Buzolin et al. Proteômica: metodologias e aplicações no estudo de doenças humanas. Revista da Associação Médica Brasileira, v. 58, n. 3, p. 366-375, 2012.

[20] LIU, H.; SADYGOV, R.G.; YATES, J.R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Analytical Chemistry, v.76, p.4193-4201, 2004.

[21] ARN, P. H. Phenylketonuria (PKU). In: Aminoff, M. J. e Daroff, R. B. Encyclopedia of the Neurological Sciences (Second Edition). Oxford: Academic Press, 2014, p.887-889.

[22] WEATHERALL, D. J. Sickle Cell Anemia. In: Hughes, S. M. Brenner?s Encyclopedia of Genetics (Second Edition). San Diego: Academic Press, 2013, p.429-431.

[23] 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.

[24] MAY C, BROSSERON F, CHARTOWSKI P, Schumbrutzki C, Schoenebeck B, Marcus K. Instruments and methods in proteomics. Methods Mol Biol. 2011;696:3-26.

[25] MCLAFFERTY, F. ?Tandem mass spectrometry.?.Science, New York,v. 214, no. 4518, p. 280-287, 1981.

[26] GALDOS-RIVEROS, Alvaro Carlos et al. Proteômica: novas fronteiras na pesquisa clínica. Enciclopédia Biosfera, v. 6, n. 11, p. 1-24, 2010.

[27] PETRICOIN, Emanuel F. et al. Serum proteomic patterns for detection of prostate cancer. Journal of the National Cancer Institute, v. 94, n. 20, p. 1576-1578, 2002.

[28] CARVALHO, PAULO COSTA. Ambiente Computacional para Proteômica. Tese (Doutorado). Programa de Engenharia de Sistema e Computação. Universidade Federal do Rio de Janeiro, 2010.

[29] TOU, J. N.; GONZALEZ. R. C. (Ed.) Pattern Recognition Principles. Massachusetts, USA: Addison-Wesley Publishing Company, 1981.

[30] LIMA, Diogo Borges. MÉTODO COMPUTACIONAL PARA IDENTIFICAÇÃO DE PEPTÍDEOS MARCADOS COM FENIL-ISOTIOCIANATO E ANALISADOS POR CROMATOGRAFIA LÍQUIDA ACOPLADA A ESPECTROMETRIA DE MASSA EM TANDEM. 2013. Tese de Doutorado. Universidade Federal do Rio de Janeiro.

[31] CASTRO, ARMANDO ANTONIO MONTEIRO; DO PRADO, PEDRO PAULO LEITE. Algoritmos para reconhecimento de padrões. Revista Ciências Exatas, v. 8, n. 2002.

[32] QUEIROZ, Suellem Stephanne Fernandes; PINTO, Kayo Luann Nogueira. Extração de características e reconhecimento de padrões e objetos. VETOR-Revista de Ciências Exatas e Engenharias, v. 24, n. 2, p. 2-13, 2016.

[33] THEODORIDIS, S.; KOUTROUMBAS, K. BPattern Recognition, Academic. New York, 2003.

[34] WATANABE, Satosi. Pattern recognition: human and mechanical. John Wiley & Sons, Inc., 1985.

[35] HAYKIN, S. Neural Networks and Learnig Machines. [S.1.]: Pearson, 2008.

[36] FAYYAD, U., PIATETSKY-SHAPIRO, G. and SMYTH, P. (1997) "From Data Mining to Knowledge Discovery in Databases", In: Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI Press/MIT Press books, USA.

[37] SYMEONIDIS, A. L., Mitkas, A. P. (2005). In: Agente Intelligence Through Data Mining, Springer, New York.

[38] LORENA, A. C.; CARVALHO, A. C. P. L. F. Uma Introdução às Support Vector Machines. Revista de Informática Teórica e Aplicada, vol.14, no2, pp 43-67, 2007. Disponível em: ¡http://seer.ufrgs.br/index.php/rita/article/viewPDFInter stitial/rita_v14_n2_p43-67/3543¿. Acesso em: 16 out.2016.

[39] MITCHELL, Tom M. Machine Learning: McGraw-Hill, 1997.Disponível em: <http://www.cs.cmu.edu/ tom/mlbook/NBayesLogReg.pdf>. Acesso em: jan. 2016.

[40] DE SOUZA, Lima L. Class-Test: Classificação automática de teste para auxílio à criação de suítes de teste. Master's thesis. Universidade Federal de Pernambuco (2009).

[41] SANTOS, Cícero Nogueira dos. Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro. Rio de Janeiro, 2005. Disponível em: <http://www2.comp.ime.eb.br/dissertacoes/2005-Cicero_Santos.pdf>. Acesso em: Fev, 2016.

[42] SOUTO, M. C. P.; LORENA, A. C.; DELMBEM, A. C. B.; CARVALHO, A. C. P. L. F. Técnicas de Aprendizado de Máquina para problemas de Biologia Molecular. Minicursos de Inteligência Artificial, Jornada de Atualização Científica em Inteligência Artificial, XXIII Congresso da Sociedade Brasileira de Computação, 2003. 103?152 p. Disponível em: <http://www.dimap.ufrn.br/ marcilio/DAAD/BIB/GENE EXPRESSION/jaia2003-14-03-08.pdf>. Acesso em: fev. 2016.

[43] BATISTA, Gustavo Enrique de Almeida Padro Alves. Pré-Processamento de Dados em Aprendizado de Máquina Supervisionado. Tese de Doutorado, 2003. ICMC-USP.

[44] FÜRNKRANZ, Johannes. Separate-and-conquer rule learning. Artificial Intelligence Review, v. 13, n. 1, p. 3-54, 1999.

[45] MONARD, M. C. & Baranauskas, J. A. Indução de Regras e Árvores de Decisão (1ed.), Chapter 5, pp. 115-140. volume 1 of Rezende (2003).

[46] CHEESEMAN, Peter; STUTZ, John; HANSON, Robin. Bayesian classification theory. 2011-05-23]. http://citeseerx. ist. psu. edu/viewdoc/download, 1990.

[47] SANTOS, Daiane Sampaio. Predição de mínimos e máximos locais para investimento em bolsa de valores utilizando aprendizado de máquina. 2014.

[48] PENG H, LONG F, DING C. *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, aug. 2005

[49] HYVARINEN, A., KARHUNEN, J., OJA, E. Independent Component Analysis. John Wiley & Sons, 2001.

[50] G.DARMOIS. "Analyse generale des liaisons stochastiques", Rev. Inst. Internationale Statist., v. 21, n. 2-8, 1953.

[51] KAGAN, A. M., LINNIK, Y. V., RAO, C. R. Characterization Problems in Mathematical Statistics. New York, Wiley, 1973.

[52] HERAULT, J., JUTTEN, C. "Space or time adaptive signal processing by neural network models". In: Neural networks for computing: Proceedings of the AIP Conference, pp. 206-211, New York: American Institute of Physics, 1986.

[53] COMON, P. "Independent component analysis - a new concepte"Signal Processing,v. 36, pp. 287-314, 1994.

[54] HAYKIN, S. Neural Networks: A comprehensive Foundation (2nd Edition), Prentice Hall.

[55] LEE, T.W.., Independent Component Analysis Theory and Applications, Kluwer, 1998.

[56] GUILHON, D. Rodrigues. Compressão e Sinais de Eletrocardiograma Utilizando Análise de Componentes Independentes. Tese Mestrado, Universidade Federal do Maranhão.

[57] PAPOULIS, Athanasios; PILLAI, S. Unnikrishna. Probability, Randon Variable and Stochastic Processes. 4e.d. Nova York: McGraw-Hill. 2002. 852p.

[58] HYVARINEM, A. One Unit Contast Function for Independent Component Analysis: A Statistical Analysis. Proc. IEE Workshop on Neural Network for Signal Processing, p.388-397, Florida, 1997(b).

[59] CATARINO, Francisco Manuel Inácio Ferreira. *Segmentação da íris em imagens com ruído.* 2009. Tese de Doutorado.

[60] DONALD, David et al. *Bagged super wavelets reduction for boosted prostate cancer classification of seldi-tof mass spectral serum profiles.* Chemometrics and intelligent laboratory systems, v. 82, n. 1, p. 2-7, 2006.

[61] VAPNIK, V.N. The Nature of Statistical Learning Theory. Berlin: Springer-Verlag, 1995.

[62] SMOLA, A. J. e SCHOOLKOPF, B. Learning with Kernels. MIT Press, 2002.

[63] CHOUDHRY, R. e GARG, K. A hybrid machine learning sytem for stock market forecasting. In Proceedings of World Academy of Science, Engineerting and Technology, 2008.

[64] HUANG, C. L. e Tsai, C. Y. A hybrid sofm-svr with a filter-based feature selection for stock market forecasting. Expert Systems With Applications, (2009).

[65] XIE., W. Yu, L., XU, S., WANG, S. A new method for crude oil price forecasting based on support vector machines. In Computacional Science-ICCS, Springer, 2006.

[66] SCHOLKOPF, B. Statistical learnig and kernel methods.

[67] CRISTIANINI, N., SHAWE-TAYLOR J., An introduction to support Vector Machines: and other kernel-based learning methods. New York: Cambridge University Press, 1999. 189p.

[68] BRAGA, A. P., T. B. LUDEMIR, & A. C. P. L. F. de Carvalho. Redes Neurais Artificiais: Teoria e aplicações. LTC - Livros Técnicos e Científicos Editora S. A, 2000.

[69] SMOLA, A. J. et al. Introduction to large margin classifiers. In: Morgan-Kauffman, 1999. cap. 1, p. 1-28.

[70] LORENA, A. C.; CARVALHO, A. C. P. L. Introdução aos Classificadores de MArgens Largas. São Carlos - SP, Maio 2003.

[71] CRAMER, D. & HOWIT, D.L. *The SAGE dictionary of Statistics.* London: Sage Publications.

[72] HAIR, J.F., BLACK. W.C. , BABIN, B.J. & ANDERSON, R.E. *Multivariate Data Analysis* (7th ed.) Upper Saddle River: Prentice Hall.

[73] MAROCO, J. *Análise estatística coom utilização do SPSS*. 3.ed. Lisboa: Edições Sílabo, 2007.

[74] LEITE, Isabel Cristina Costa. *Análise de componentes independentes aplicada a avaliação de imagem radiográfica de sementes.* Universidade Federal de Lavras. Tese (doutorado), 2013.

[75] REZENDE, S. O., Sistemas Inteligentes:Fundamentos e aplicações, Ed. Manole, p. 535, 2005.

[76] KOHAVI, R. A study a cross validation a bootstrap for accuracy estimation and a model selection. In: International Joint Conference on Artificial Intelligence (IJCAI). [S.l.: s.n.], 1995.

[77] THEODORIDIS, S. e KOUTROUMBAS, K. Pattern Recognition. Elsevier, second edition, 2003.

[78] DING, C. and PENG, H. "Minimum Redundancy Feature Selection from Microarray Gene Expression Data". Proc. Second IEEE Computational Systems Bioinformatics Conf. 2003. Pág. 523-528, August.

[79] DUDA, R. O. & P. E. HART. Pattern Classification and Scene Analysis. John Wiley & Sons Inc.1973.

[80] CAMPBELL, C. (2001). An introduction to kernel methods. pp. 155-192.

[81] BURGESS, C. J. C. A tutorial on support vector machines for pattern recognition, 1998. Data Mining And Knowledge Discovery 2, 121-167.

[82] BERSTSEKAS, D. P. Constrained Optimization and Lagrange Multiplier Methods, 1982. Academic Press.

[83] MULLER, K. R., S. MIKA, G. Ratsch, K. Tsuda, & B. Schölkopf (2001). An introduction to kernel-based learning algorithms. Neural Networks, IEEE Transactions on 12(2), 181-201.

[84] KHATTREE, R.A. & NAIK, D.N. *Multivariate data reduction and discrimination with SAS software.* Cary, NC, USA: SAS Institute Inc., 2000. 558p.

[85] REGAZZI, A.J. *Análise multivariada, notas de aula INF 766.* Departamento de Informática da Universidade Federal de Viçosa, v.2, 2000.

[86] XU, P.; BROCK, N.; PARRISH, R.S. *Modified linear discriminant analysis approaches for classification of high-dimensional microarray data.* Computacional Statistics & Data Analysis, v.53, n.5, p.1674-1687, 2008. <http://dx.doi.org/10.1016/j.csda.2008.02.005>

[87] MARTINEZ, A.M.; KAK, A.C. PCA versus LDA. IEE Transactions on Pattern Analysis and Machine Intelligence, v.23, n.2, p.228-233, 2001. Disponível em<http://dx.doi.org/10.1109/34.908974>.

[88] FISHER, R.A. *The use of multiple measurements in taxonomic problems.* Annals of Eugenics, v.7,n.2p.179-188, 1936. <http://dx.doi.org/101111/j.1469-1809.1936.tb02137.x>

[89] KITANI, E.C.; THOMAZ, C.E. Análise de discriminante lineares para modelagem e reconstrução de imagens de face. In: Congresso da SBC, 27., 2007, Rio de Janeiro. Anais.

[90] BELHUMEUR, P.; HESPANHA, J. P. N.; KRIEGMAN, D. J. *Eigenfaces vs. Fisherfaces: recognition using class specific linear projection.* IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 19, n. 7, pp. 711-720, July 1997.

[91] *Data Base*. Disponível em: <home.ccr.cancer.gov/ncifdaproteomics/ppatterns>.

[92] CABENA, P., HADJINIAN, P., STADLER, R,. VERHEES,. J e ZANASI, A. *Discovering Data Mining: From Concept to Implementation*, Prentice Hall, 1997.

[93] BUSHBERG, J.T. and BOONE, J.M. *The essential physics of medical imaging.* Lippincott Williams & Wilkins(2011).