

2

Conceitos Básicos

Neste capítulo será dada uma visão geral de alguns conceitos que serão de fundamental importância na compreensão dos capítulos seguintes, como Imagem Médica, Nódulo Pulmonar Solitário, os dois algoritmos de classificação utilizados: Análise Discriminante Linear de Fisher e Rede Neural Perceptron de Múltiplas Camadas, o procedimento de seleção de variáveis (medidas) *passo a passo*, a técnica para validação do modelo *deixa um de fora* e, por último, a técnica de avaliação de diagnóstico Curva ROC.

2.1

Imagem Médica

Esta seção dá uma visão geral de alguns conceitos importantes sobre imagem médica, como aquisição da imagem, formas de tratamento de imagens em Computação Gráfica, o padrão DICOM e a técnica de interpolação linear.

2.1.1

Aquisição da Imagem

As técnicas de aquisição de imagens médicas podem ser divididas em invasivas e não invasivas, de acordo com a forma como são obtidas. Os métodos invasivos caracterizam-se pela introdução de um instrumento no interior do corpo humano, de forma a obter as imagens pretendidas. Nesta categoria incluem-se as angiografias e as imagens de medicina nuclear. Nos métodos não invasivos incluem-se os raios X, ultra-sonografia, tomografia computadorizada e ressonância magnética.

Os dados volumétricos extraídos desses métodos são geralmente adquiridos na forma de imagens de fatias paralelas uniformemente espaçadas, representando cortes transversais ao eixo longitudinal do paciente. Comumente nas regiões de maior interesse são feitos cortes mais

próximos, permitindo uma maior visualização dos dados. Cada imagem gerada está associada a uma localização k , $k = 1, 2, \dots, l$, no eixo z e uma espessura $\Delta z = e$ em torno desta localização, formando um cubóide. O cubóide é subdividido em outros cubóides pequenos chamados *voxels*. O *voxel* é equivalente a *pixel* em 3D e representa uma abreviação para *volume element*. Cada *pixel* da imagem está associado a um *voxel*. O valor associado a cada *pixel* representa a média das atenuações do raio X no volume interno do corpo correspondente ao *voxel*. Os valores destas atenuações são expressos em Unidades de Hounsfield (UH) [38]. Tais valores são obtidos pela exposição do corpo ao bombardeamento de raios X em várias direções.

O valor associado a cada *voxel* é um número inteiro, proporcional ao tom de cinza do *pixel* na imagem correspondente, e representa a integração de alguma propriedade física que está sendo mensurada no interior do volume associado ao *voxel*. No caso da tomografia computadorizada, por exemplo, a grandeza física medida é a densidade do tecido. Quanto maior for a densidade do tecido, maior serão as atenuações e, portanto, maior serão os valores dos *pixels* nas imagens dos cortes referentes a este tecido.

Nas próximas seções serão abordadas as características gerais de quatro métodos não invasivos de aquisição de imagens médicas.

Raio X

Em 1895, o físico alemão Wilhelm Rontgen descobriu os raios X, descoberta que viria a revolucionar o meio científico, e em especial a Medicina [79].

Na formação de uma imagem de raio X é emitida uma determinada fonte de radiação, que atravessa o corpo humano e é projetada num filme sensível. Os diferentes tecidos do corpo humano absorvem a radiação emitida em quantidades distintas, de forma que os raios atingem o filme com diferentes intensidades, dependendo da radiação absorvida.

Ressonância Magnética

A ressonância magnética é principalmente aplicada a “tecidos moles”. No interior do corpo humano, todos os núcleos atômicos possuem um determinado campo magnético, o que significa que eles se comportam como pequenos ímãs. Quando o paciente é colocado no interior de um tubo capaz de gerar um elevado campo magnético, os núcleos alinham-se na direção deste campo, vibrando em torno do seu eixo com uma frequência que

depende fundamentalmente do tipo de núcleo, o que permite distinguir os diversos tipos de tecidos.

Ultra-sonografia

Nas imagens produzidas por ultra-som são usados impulsos sonoros de alta frequência, em vez de energia de radiação.

Um emissor é manipulado por um operador sobre o corpo do paciente, permitindo obter imagens em tempo real. Assim que uma onda sonora encontra um tecido, uma parte dela é refletida, sendo o tempo que leva a regressar ao ponto de origem (eco) proporcional à distância a que se encontra o tecido. A amplitude do sinal de eco depende das propriedades acústicas dos tecidos e manifesta-se na imagem gerada sob a forma de diferentes intensidades no brilho produzido.

Tomografia Computadorizada

A Tomografia, derivada da palavra grega “Tomos”, que significa corte ou fatia, e “Grafos”, que significa desenhar uma imagem ou gráfico, emprega os mesmos princípios da radiografia convencional com o objetivo de criar uma representação anatômica baseada na quantidade de atenuação sofrida pela radiação incidente. O nome Tomografia Computadorizada (TC) deve-se ao fato dessa técnica ser altamente dependente de computadores para realizar os cálculos matemáticos relativamente complexos referentes às informações coletadas durante a emissão e rotação dos raios X.

Na TC, o feixe de raios X que atravessa o corpo é muito colimado e fino, reduzindo sobremaneira a produção de raios secundários que degradariam a imagem. Diferentemente do estudo radiológico convencional, os raios X não impressionam filmes após atravessarem o corpo, mas são captados por detectores de fótons e as medidas de atenuação tissular são calculadas e armazenadas no computador. Tais mensurações são feitas em Unidades de Hounsfield (UH). A Figura 2.1 ilustra o funcionamento de uma TC.

Quanto mais densas as regiões do corpo, maiores serão seus valores de atenuação em UH. Assim, o ar contido nas vias respiratórias e no tubo digestivo tem valores mais negativos, como -800 UH ou -1000 UH, e os ossos, os mais positivos, tais como 400 UH ou 500 UH. A água é usada para a calibração do equipamento e seus valores de atenuação estão entre 0 e ± 10 UH [14], [80].

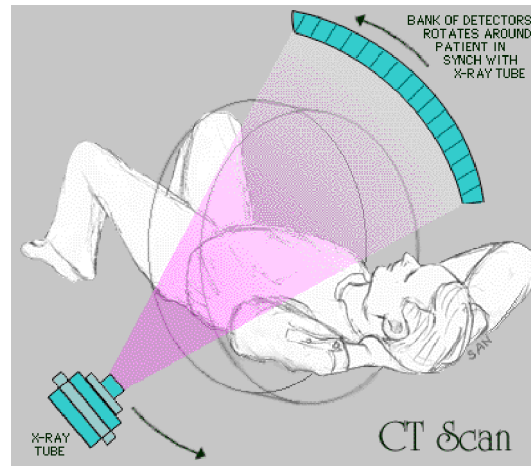


Figura 2.1: Funcionamento de uma TC (Fonte: <http://www.geocities.com/siumingrd/CT>).

Na realidade, a imagem obtida com equipamentos de TC é o resultado da disposição na tela do monitor de uma enorme quantidade de números lado a lado e em linhas, que representam coeficientes de atenuação tissular, produtos de cálculos efetuados pelo computador enquanto o feixe de raios X atravessa a área estudada. Cada valor numérico corresponde a uma tonalidade em escala de cinza, que vai do preto ao branco. As áreas mais escuras indicam menor densidade e as mais claras indicam maior densidade. A Figura 2.2 exemplifica uma TC do tórax e mostra algumas estruturas encontradas no exame.

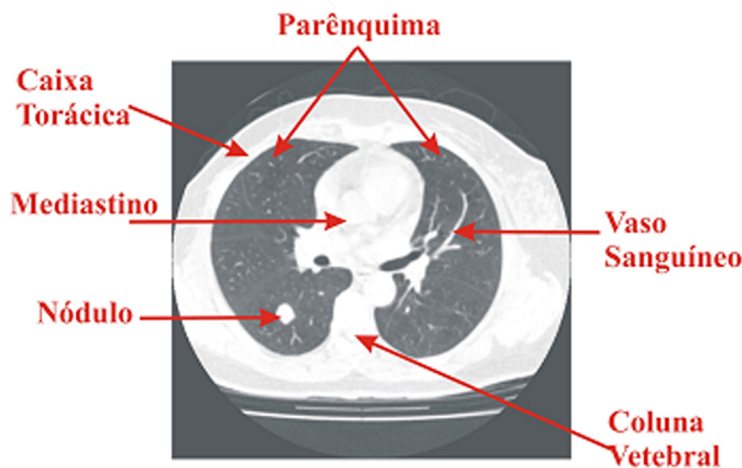


Figura 2.2: Tomografia computadorizada do tórax.

A TC é de fundamental importância no diagnóstico precoce do Nódulo Pulmonar Solitário, pois é muito sensível a diferenças em densidades, podendo identificar lesões menores que 1 mm^3 . Devido a essas

características, a TC identifica calcificações com mais precisão do que outros métodos radiográficos, e possibilita um diagnóstico mais confiável, trazendo como consequência maior sobrevida para o paciente.

2.1.2

Computação Gráfica e Medicina

Podem-se identificar quatro técnicas básicas de Computação Gráfica que são amplamente utilizadas na área médica: representação de dados, processamento de imagens, reconstrução e visualização [61].

Representação de dados

A representação da imagem trata da caracterização da quantidade de *pixels* que representa a imagem e como ela é representada de forma compacta para armazenamento e transmissão. O modo de representar e armazenar uma imagem em computador influi decisivamente no desempenho dos algoritmos que implementam as operações de manipulação e análise. Determina, também, o espaço de memória (estática ou dinâmica), o que, em algumas aplicações, é importante devido ao grande volume de dados que constituem a imagem.

Processamento e Análise de Imagens

Uma vez equacionado o problema de aquisição e representação de dados, a fase seguinte consiste em efetuar o processamento dos dados de forma a obter os resultados desejados. Dentre os métodos de processamento, podemos destacar a segmentação e o registro.

O problema de segmentação consiste em classificar regiões de uma imagem com diferentes atributos (cor, opacidade, profundidade, textura, etc.). Isto é conseguido através de um particionamento do domínio da imagem baseado em propriedades da função de atributos. Um particionamento muito usado consiste em determinar regiões do domínio da imagem nas quais alguns dos atributos têm valores diferenciados dos demais.

O problema de registro de imagens consiste em alinhar objetos em duas ou mais imagens. Essas imagens podem ter sido obtidas, por exemplo, em instantes diferentes, por sensores diferentes ou de ângulos diferentes. Para registrar duas imagens, faz-se necessário determinar uma transformação tal

que cada ponto na primeira imagem possa ser mapeado em um ponto na segunda. Esse mapeamento deve alinhar as duas imagens da melhor maneira possível, sendo que o significado de “melhor maneira” depende dos objetos a serem alinhados nas duas imagens.

Reconstrução

A reconstrução consiste em obter a geometria e a topologia de um objeto gráfico a partir de suas amostras. Os equipamentos médicos de aquisição de dados, por exemplo, capturam “amostras” dos diversos órgãos, e é preciso desenvolver técnicas que possibilitem uma reconstrução tridimensional do órgão a partir dessas amostras.

Portanto, o problema de reconstrução consiste em recuperar um objeto representado por um conjunto de dados amostrados. Para reconstruir um objeto, é necessário que a representação inclua ainda, um modelo de como a geometria varia entre as amostras. Em geral, esta a variação é obtida com o uso de algum método de interpolação aplicado aos dados amostrados.

Visualização

A visualização volumétrica consiste em obter informações visuais sobre dados médicos de naturezas diversas. A visualização volumétrica de objetos anatômicos elucida a sua estrutura tridimensional.

Dados volumétricos são valores estruturados geometricamente em um volume e, em geral, são obtidos a partir de três tipos de processos: a) *scanners* tridimensionais (ressonância magnética, tomografia computadorizada, etc.), b) simulações baseadas em modelos computacionais, e c) da conversão de um modelo geométrico.

Existem duas classes de técnicas de visualização de volumes, que se traduzem nas que trabalham com a extração de uma isosuperfície representada através de primitivas gráficas e nas que trabalham gerando a imagem diretamente a partir do volume.

Técnicas de visualização através de superfícies envolvem a extração e a representação de uma isosuperfície que é posteriormente visualizada através da utilização de técnicas convencionais da Computação Gráfica. Entre os algoritmos de visualização através de superfícies destacam-se o de conexão de contornos [21] e o *marching cubes* [13]. Este último foi o algoritmo de visualização adotado neste trabalho.

A segunda classe, visualização direta de volume, consiste em representar o volume através de *voxels* 3D que são projetados diretamente em *pixels* 2D e armazenados como uma imagem, dispensando o uso de primitivas geométricas. Os algoritmos que fazem parte deste grupo são [90], [21]: *ray casting*, *splatting*, *shear-warp*, *shell rendering*, *cell-projection* e *V-Buffer*.

2.1.3 Padrão de Imagens DICOM

O padrão DICOM (*Digital Imaging and Communications in Medicine*) é uma especificação detalhada que descreve um meio de formatar e trocar imagens juntamente com informações associadas. É dirigido aos mecanismos de operação da interface usados para transferir dados de e para um determinado dispositivo de imagem.

Essa especificação relaciona ligações de redes normatizadas e dispositivos de armazenamento (*Media Storage Devices*), responsáveis pela comunicação e arquivo de imagens digitais, provenientes de tomografia computadorizada, ressonância magnética, medicina nuclear, ultra-sonografia, raios X, etc.

A comissão ACR-NEMA (*American College of Radiology - National Electrical Manufacturers Association*) foi criada em 1983 com a missão de desenvolver uma interface entre os equipamentos de imagens médicas (tais como tomografia computadorizada, ressonância magnética, medicina nuclear e ultra-sonografia) e qualquer outro dispositivo com que se quisesse comunicar. Além das especificações para a ligação do hardware, o padrão a desenvolver deveria incluir um dicionário de elementos de dados, que possibilitasse a interpretação e a visualização correta da imagem [71].

A especificação do padrão DICOM 3.0 encontra-se dividida em 13 partes. Tal divisão permite que cada parte possa expandir-se individualmente sem haver necessidade de reeditar todo o padrão. Dentro das partes, as seções sujeitas a adições ou modificações encontram-se em suplementos, reduzindo assim o esforço de edição necessário quando da sua atualização [39].

A adoção do padrão DICOM pelas indústrias de imagem médica abre novas oportunidades para organizações de cuidados à saúde para aumentar a qualidade e a eficiência nos cuidados aos pacientes. O sistema DICOM permite que informações sobre um paciente viajem entre lugares diferentes do mundo via modem, o que é mais barato e mais rápido do que outros

meios de transporte. Além disso, as imagens não perdem a definição e, conseqüentemente, a interpretação das imagens pelas entidades médicas é mantida, já que a qualidade gráfica não se altera. A Figura 2.3 ilustra a estrutura da imagem no formato DICOM. Este formato de imagem é o utilizado neste trabalho.

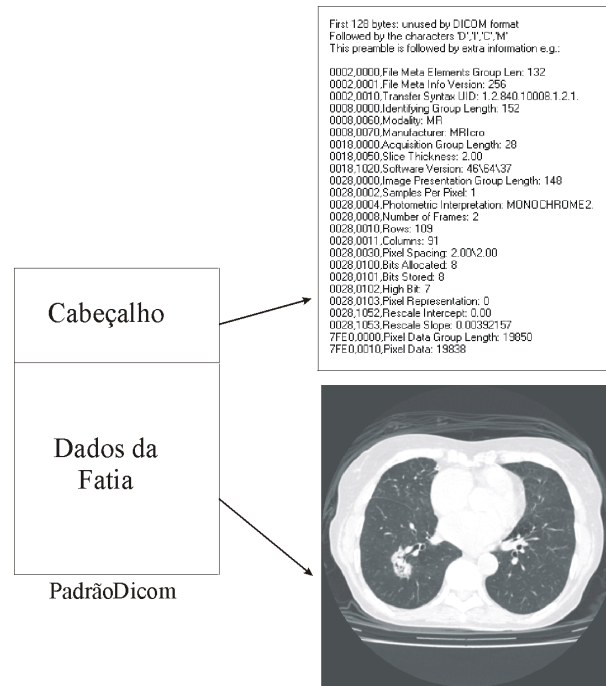


Figura 2.3: Estrutura da imagem no formato DICOM.

2.1.4 Interpolação

A Figura 2.4 mostra um dado volumétrico. Pode-se notar nesta figura que d define a qualidade da amostragem na direção z e que a dimensão p dos *pixels* define a qualidade nas direções x e y . A relação entre d e p dita o grau de anisotropia da amostragem. A interpolação tem por objetivo melhorar a qualidade da amostragem, estimando valores amostrados em uma nova escala e gerando uma amostragem isotrópica. Esta correção de escala é importante neste trabalho para calcular as medidas propostas, em imagens de TC com espaçamentos diferentes entre fatias. Desta forma, a interpolação uniformiza o máximo possível as imagens que contêm os nódulos.

A Figura 2.5 ilustra a transformação ocorrida no espaço do *voxel* na operação de interpolação. O espaço tem resolução de $2 \times 2 \times 2$ *voxels* ($m = n = l = 2$) e as dimensões dos *voxels* são $\Delta x = \Delta y = p$ e $\Delta z = 2p$.

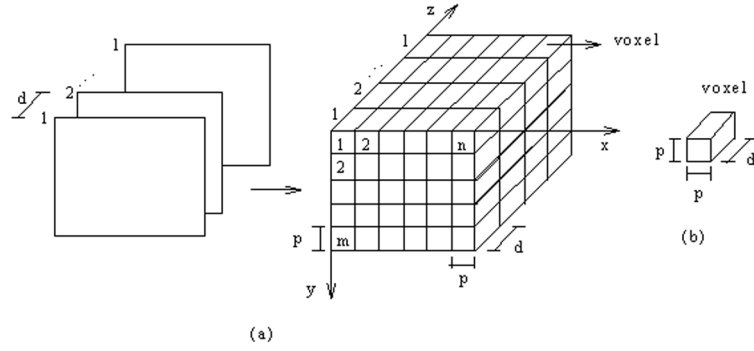


Figura 2.4: Espaço amostral do *voxel* [56].

Para obter *voxels* cúbicos com dimensões $\Delta x = \Delta y = \Delta z = p/2$, novas amostras podem ser interpoladas nas fatias 1 e 2, aumentando a resolução das fatias para 4×4 *pixels*, e novas fatias com resolução 4×4 *pixels* podem ser interpoladas entre as fatias 1 e 2. A base para a interpolação são as densidades dos 8 *voxels* do espaço original.

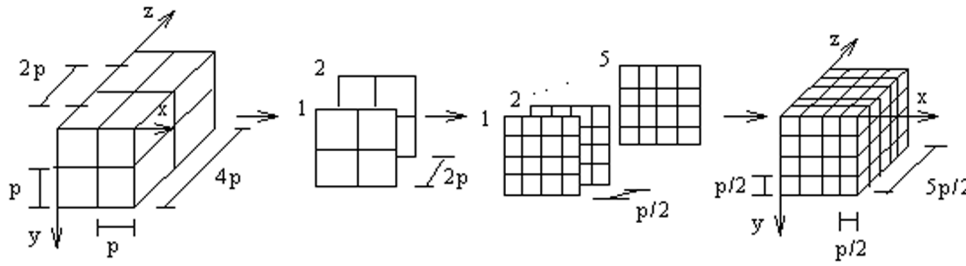


Figura 2.5: Exemplo de interpolação [56].

O exemplo da Figura 2.5 mostra que, para conseguir *voxels* cúbicos, com dimensões $\Delta x = \Delta y = \Delta z = p$, basta apenas interpolar amostras na direção z . Esta é a forma mais comum de interpolação. Entretanto, a interpolação nas direções x , y e z é a mais genérica.

Neste trabalho será usada apenas uma interpolação linear em relação a z . A Figura 2.6 ilustra o processo de interpolar uma fatia m entre as fatias n e $n+1$. A interpolação linear assume que a variação de densidade é linear na direção z entre os *voxels* v_n e $v_n + 1$. A densidade $d_i(v_m)$ é obtida por:

$$d_i(v_m) = d_o(v_n) + \frac{(d_o(v_{n+1}) - d_o(v_n)) l_i}{l_s + l_i} \quad (2-1)$$

onde d_i é a densidade interpolada, d_o é a densidade original, $l_s + l_i$ (espaçamento entre as fatias n e $n+1$). O procedimento é repetido para os outros *voxels* da fatia m a serem interpolados.

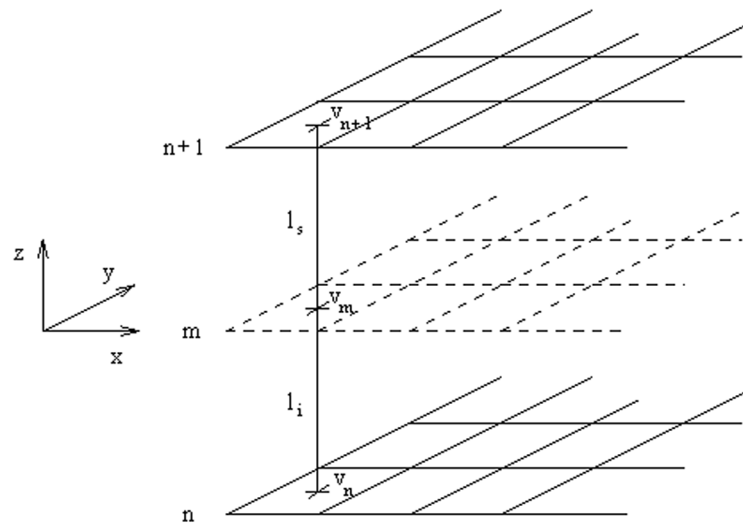


Figura 2.6: Interpolação linear [56].

2.2 Nódulo Pulmonar Solitário

A identificação de um Nódulo Pulmonar Solitário (NPS) é um problema freqüente na prática radiológica. O radiologista tem um papel determinante na avaliação adequada das características morfológicas deste tipo de lesão e na orientação da conduta mais apropriada para o seu tratamento. É importante salientar que se o câncer de pulmão for identificado e diagnosticado com tamanho inferior a 3 cm, há uma chance de sobrevivência do paciente de 80% [81].

O NPS é caracterizado como uma imagem discreta (isolada), aproximadamente esférica, com densidade maior que a do ar, com contornos definidos e tamanho de até 3 cm. Se o nódulo tiver mais de 3 cm é denominado “massa”. A massa tem as demais características semelhantes às do nódulo e deve parecer ter atingido essas dimensões por crescimento.

Várias enfermidades podem se manifestar nos NPS, mas as principais causas são o carcinoma broncopulmonar ¹ (44%), seguido de tuberculose pulmonar (23%), tumores benignos (13%), a metástase (9%) e os abscessos ² (5%) [81].

¹Um tipo de câncer de pulmão formado por células primárias do pulmão.

²São como um furúnculo no pulmão.

2.2.1

Natureza do Nódulo

Para se fazer uma hipótese diagnóstica, devem ser levados em consideração três fatores principais: características radiográficas, dados clínicos e frequência estatística de determinados processos [38].

As características radiológicas da lesão são de grande auxílio, até mesmo podendo definir a natureza benigna ou maligna de um NPS. Nessa avaliação utilizam-se os princípios gerais de descrição de qualquer imagem médica, que englobam seis aspectos fundamentais: 1) forma, 2) densidade (estrutura), 3) limite, 4) localização, 5) número, e 6) evolução (mudança) [38], [62].

Os cinco primeiros aspectos descrevem as características geoespaciais do nódulo, enquanto o último se refere à sua condição evolutiva-temporal.

Com muita frequência, os médicos especialistas não contam com as características listadas acima para classificar o nódulo como benigno. A lesão será, então, considerada de natureza indeterminada, o que é insuficiente para definir uma conduta a ser adotada. Geralmente, há a expectativa de que seja avaliada a probabilidade de determinada lesão ser ou não de natureza maligna. Essa avaliação deve ser compreendida como uma opção reservada, repleta de limitações, que será importante basicamente naqueles casos em que o risco cirúrgico for muito grande, quando comprovado a um eventual erro de diagnóstico.

Os principais dados a serem considerados para inferir sobre a probabilidade de benignidade e malignidade de uma lesão, sob a ótica do diagnóstico por imagem, são [38]:

a) modificação temporal-evolutiva;

- o tempo de duplicação de uma lesão situa-se entre 30 e 450 dias, e seu achado será sugestivo de malignidade. Para que um nódulo atinja 1 cm de diâmetro, a partir de uma única célula neoplásica, o tempo exigido é de dois anos e meio a 25 anos;
- uma lesão que se duplique em menos de sete dias sugere fortemente que sua natureza seja benigna.

b) presença e tipos de calcificação(ões);

- a calcificação difusa, da subtotalidade da lesão, é um achado muito sugestivo de benignidade;
- mais de 30% das lesões não calcificadas podem ser consideradas de natureza benigna;

- a calcificação em nódulos malignos é um achado pouco freqüente, mas não raro, e pode ocorrer por alteração distrófica, ossificação do tumor ou inclusão de granuloma calcificado previamente existente, que em geral é excêntrico e discreto.

c) tamanho absoluto da lesão;

- menos de 5% das lesões benignas têm mais de 3 cm;
- lesões menores de 1 cm, identificadas na tomografia computadorizada do tórax, tanto podem ser de natureza maligna como benigna.

d) interface nódulo-parênquima;

- configuração lobulada com limites espiculados são indícios fortes de malignidade;
- contornos lisos e regulares (não lobulados), sem infiltração do parênquima circunjacente, são sugestivos, porém não conclusivos, de benignidade;
- configuração regular com limites precisos (circunscritos, sem espículas) em uma lesão que sofre de modificação da forma com a mudança de decúbito ³, é muito sugestiva de lesão cística de conteúdo líquido, e em sua grande maioria, benigna.

e) variação de densidade após impregnação de contraste endovenoso.

Será considerado benigno do ponto de vista radiológico um nódulo em que se identifique [62], [72], [40], [38]:

- i) calcificação difusa, central ou em camadas;
- ii) limites precisos (liso, circunscrito) em uma lesão que sofre modificação em sua forma com a mudança de decúbito;
- iii) tempo de duplicação de uma lesão menor que sete dias;
- iv) ausência de crescimento por mais de dois anos.

Os dados clínicos e a freqüência estatística de determinados processos mórbidos são fatores essenciais para o diagnóstico do nódulo. Por exemplo, a incidência de determinadas doenças em relação à faixa etária, sexo ou ao habitat do paciente. Assim, seria pouco provável que um NPS em uma

³Posição adotada pelo paciente no leito: ele pode estar deitado com a barriga para cima (decúbito dorsal), de barriga para baixo (decúbito ventral), ou de lado (decúbito lateral).

criança correspondesse a carcinoma brônquico. Por outro lado, um nódulo teria grande probabilidade de ser carcinoma brônquico se identificado em paciente tabagista ativo ou passivo, com história familiar de neoplasia, com mais de 40 anos, com emagrecimento ou ainda com pneumonias de repetição, num mesmo local.

As Figuras 2.7 e 2.8 resumem a provável natureza do nódulo para diagnóstico do NPS em relação ao coeficiente de atenuação e à forma, respectivamente [38]. Observa-se que cada uma das características pode sugerir um ou mais tipos de lesão.






HOMOGÊNEO	HETEROGÊNEO		
MALIGNO / BENIGNO	Não Escavado (com calcificação)	Difusa: raramente MALIGNO	-
		Central: BENIGNO	
		Laminar concêntrica: raramente MALIGNO	
		Pipoca: BENIGNO	-
		Puntiforme: MALIGNO / BENIGNO	-
		Excêntrica: MALIGNO	-
	Cavitário	Com necrose: MALIGNO / BENIGNO	
		Com gordura: BENIGNO	-
		Com ar: MALIGNO / BENIGNO	
		Com nível líquido: MALIGNO / BENIGNO	

Figura 2.7: Provável diagnóstico do NPS em relação ao coeficiente de atenuação.

Na maior parte das ocasiões os NPS são indeterminados, isto é, não existem dados suficientes para que os médicos o diagnostiquem como malignos ou benignos. Assim, é preciso utilizar um algoritmo que permita combinar a precaução de não deixar de estudar processos malignos e evitar técnicas desnecessárias em processos benignos. A Figura 2.9 resume este algoritmo [82].






CONFIGURAÇÃO	Regular	Esférico: MALIGNO / BENIGNO		-
		Ovóide: MALIGNO / BENIGNO		-
		Lobulado: MALIGNO / BENIGNO		-
	Irregular: MALIGNO			-
LIMITES	Preciso (liso, circunscrito): BENIGNO			
	Impreciso	Espiculado: MALIGNO		-
		Nebuloso: MALIGNO		-

Figura 2.8: Provável diagnóstico do NPS em relação à forma.

2.3

Técnicas para Analisar, Discriminar e Classificar

Em termos gerais, o reconhecimento de padrões é a ciência que compreende a identificação ou classificação de medidas de informações em categorias. Categorias têm por característica representar entidades ou padrões de informação que apresentam similaridades. O reconhecimento de padrões é composto de um conjunto de técnicas e abordagens que são usadas de forma integrada na solução de diversos problemas práticos, como por exemplo a identificação de um nódulo pulmonar como maligno ou benigno. Entre as abordagens que podem ser empregadas na classificação de problemas pode-se destacar a Análise Discriminante Linear de Fisher e Rede Neural Perceptron de Múltiplas Camadas (*Multilayer Perceptrons*).

Neste trabalho foram utilizadas duas técnicas para discriminar e classificar os NPS. A Análise Discriminante Linear de Fisher foi escolhida devido ao seu grande potencial em classificação, e é muito utilizada nos trabalhos analisados (Seção 1.4). A Rede Neural Perceptron de Múltiplas Camadas foi escolhida pelo fato de ser simples e nos últimos anos estar sendo amplamente utilizada como ferramenta de diagnóstico. Com essas duas técnicas, será realizada a comparação entre elas, com o objetivo de verificar a eficiência na classificação dos NPS.

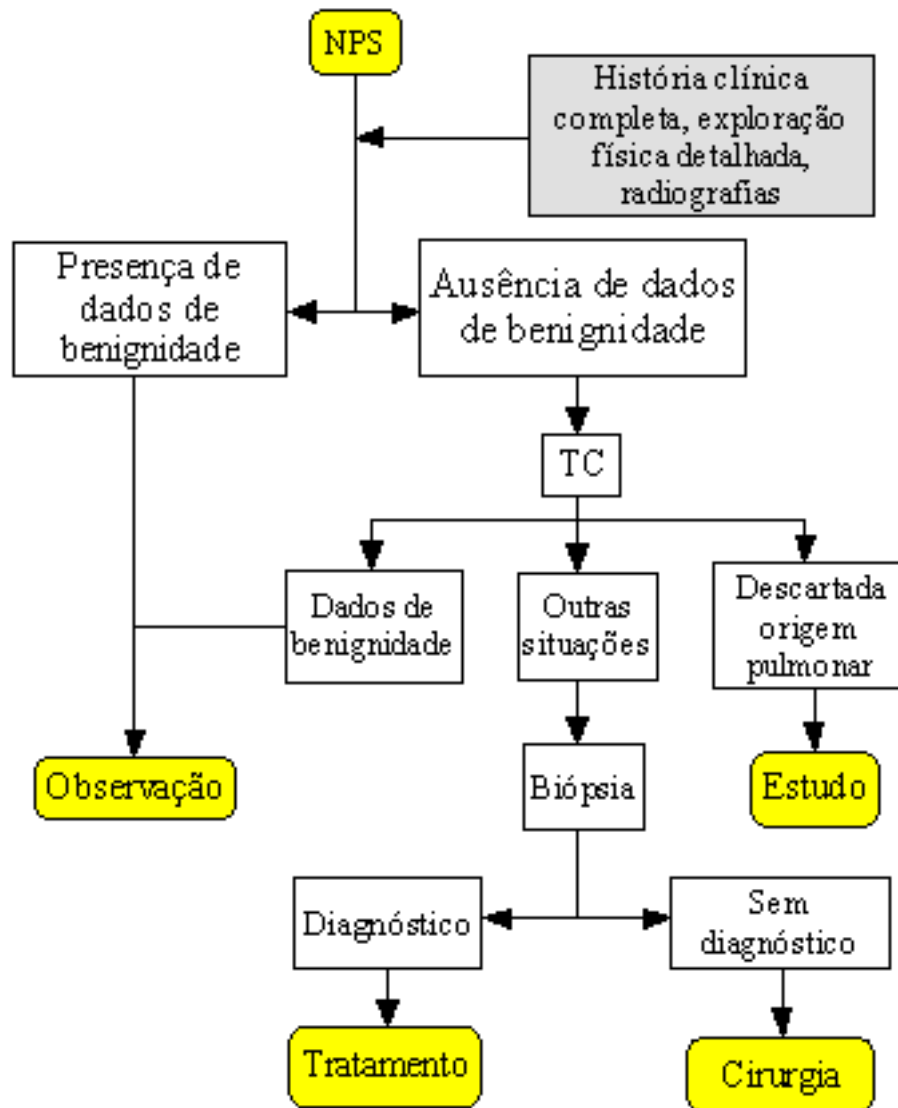


Figura 2.9: Algoritmo para diagnóstico dos NPS.

A Análise Discriminante Linear de Fisher (ALDF) é uma técnica estatística que permite discriminar e classificar indivíduos pertencentes a dois ou mais grupos mutuamente exclusivos definidos *a priori*, com base em um número de variáveis independentes observáveis. Essas variáveis observáveis são chamadas de “variáveis discriminantes”. Para isso é calculada uma “função discriminante”, que é uma função composta por índices, onde cada índice tem um peso específico. Esses pesos são calculados por uma metodologia estatística não subjetiva.

Redes Neurais Artificiais são técnicas computacionais que têm se mostrado extremamente eficientes na solução de problemas para os quais os métodos tradicionais da computação convencional não têm apresentado soluções satisfatórias, sendo uma de suas áreas de maior potencial de

aplicação justamente problemas ligados ao reconhecimento de padrões. Uma rede neural pode ser vista como um conjunto de elementos processadores simples, baseados em neurônios, que são ligados uns aos outros através de conexões análogas às sinapses. Estas conexões guardam o “conhecimento” da rede e os diversos padrões de conectividade expressam os vários objetos representados pela rede. O conhecimento da rede é adquirido por meio de um processo de treinamento no qual as conexões entre as unidades são variadas através das mudanças de pesos. Dentre os diversos algoritmos de redes neurais, o algoritmo Perceptron de Múltiplas Camadas (MLP) é um dos mais utilizados devido à sua simplicidade e eficiência.

2.3.1

Análise Discriminante Linear de Fisher

A técnica multivariada da análise discriminante trata dos problemas relacionados com a separação de conjuntos distintos de objetos (ou observações) e a alocação de novos objetos (observações) em conjuntos previamente definidos. Essa técnica está inserida em um contexto mais amplo, que é o do reconhecimento de padrões. Seu objetivo é construir uma regra de reconhecimento de padrões e classificação.

A análise discriminante e a de classificação são técnicas multivariadas interessadas, respectivamente, na separação de uma coleção de objetos distintos e na alocação de novos objetos em grupos previamente definidos [28]. Apesar de estarem claramente interligadas, não devem ser confundidas. A análise discriminante se refere aos métodos de atribuição de classes a determinados conjunto de dados. Por exemplo, pode-se considerar NPS benignos e malignos; cada um seria um grupo, diferenciado pela função discriminante. Já a classificação se refere à alocação de novos NPS nos seus devidos grupos correspondentes.

As discriminações podem ser feitas através dos processos supervisionados que são utilizados quando se conhece o padrão (dados para treinamento) ou através dos processos não supervisionados, sendo estes recomendados quando não se tem um padrão reconhecido. A análise discriminante é um método supervisionado de concepção estatística.

Ela deve ser empregada quando as seguintes condições puderem ser atendidas [28], [2]:

- a) os grupos sob investigação são mutuamente exclusivos;
- b) cada grupo é obtido de uma população normal multivariada;

- c) as matrizes de covariância relativas a cada grupo são iguais;
- d) devem existir no mínimo dois grupos: $g \geq 2$, onde g é número de grupos;
- e) devem existir pelo menos dois indivíduos por grupo: $N_i \geq 2$, onde N_i é o número de indivíduos do grupo i ;
- f) duas medidas não podem ser perfeitamente correlacionadas ($r_{ij} \neq 1$);
- g) o número máximo de variáveis é igual ao número de observações menos dois: $0 < n < (N - 2)$.

O objetivo da análise discriminante é determinar um conjunto de coeficientes discriminantes para um conjunto de variáveis independentes que forneçam uma ponderação linear capaz de extrair a maior quantidade possível de informação quanto à classificação dos indivíduos nos grupos. Ela visa maximizar a variância entre grupos (intergrupar) em relação à variância dentro dos grupos (intragrupar), considerando-se amostras previamente classificadas dos diversos grupos.

Como resultado, a análise discriminante é um sistema de escores. O escore é determinado multiplicando-se o peso discriminante pelo valor de cada variável independente do indivíduo e somando-se os resultados. Uma vez que esse escore é determinado, o indivíduo é classificado como pertencente a um dos grupos analisados.

A análise discriminante envolve derivar combinações lineares de variáveis independentes que irão discriminar entre grupos definidos *a priori* tal que as taxas de má classificação sejam minimizadas. É importante salientar que a eficiência de uma técnica é proporcional à qualidade das informações disponíveis, enfatizando-se a importância da fase de coleta de dados.

Análise Discriminante de Fisher para Discriminação entre Dois Grupos

Este trabalho tem por objetivo classificar os NPS como benignos ou malignos. Estes grupos serão designados por π_1 e π_2 , respectivamente. Os nódulos são separados e classificados com base em suas medidas, associadas a p variáveis aleatórias $X^T = [X_1, X_2, \dots, X_p]$.

O objetivo é achar a combinação linear de $Y = b^T X$ para a qual a razão entre a variância da diferença entre as médias dos dois grupos π_1 e π_2 e a variância total seja maximizada. Isto é, deseja-se obter um vetor de

pesos b que maximize [3]:

$$\Delta = \frac{|b^T (\mu_1 - \mu_2)|^2}{b^T \Sigma b} \quad (2-2)$$

onde μ_1 e μ_2 são as médias de π_1 e π_2 , respectivamente, e Σ é a matriz de covariância de X_1, X_2, \dots, X_p .

Como geralmente os parâmetros da população não são conhecidos, usa-se \bar{X} em vez de μ , e S em vez de Σ .

Pode-se mostrar que b é dado por [3]:

$$b = S^{-1} (\bar{X}_1 - \bar{X}_2) \quad (2-3)$$

onde b é o vetor de pesos, S^{-1} é inversa da matriz de covariância amostral da população, \bar{X}_1 é o vetor da média amostral de π_1 , e \bar{X}_2 é a média amostral de π_2 .

O cálculo de S pode ser obtido de duas formas [3]. A primeira forma é através da equação:

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \quad (2-4)$$

onde S é matriz de covariância conjunta, S_1 e S_2 são as matrizes de covariância de π_1 e π_2 , respectivamente, e n_1 e n_2 são os números de indivíduos de π_1 e π_2 , respectivamente.

A segunda forma é através da equação:

$$S = W + B \quad (2-5)$$

onde W é matriz de covariância intragrupo e B é a matriz de covariância intergrupo.

A matriz de covariância intragrupo (W) é definida por:

$$W = \sum_{i=1}^p \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \quad (2-6)$$

$i = 1, \dots, p; j = 1, \dots, n_i$, e

$$\bar{X}_i = \left(\frac{1}{n_i} \right) \sum_{j=1}^{n_i} X_{ij} \quad (2-7)$$

onde p é o número de amostras, n_i é o tamanho da i -ésima amostra, X_{ij} observações (j -ésima observação da i -ésima amostra), e \bar{X}_i é a média amostral para a i -ésima amostra.

A definição da matriz B de variância intergrupo das n variáveis calculada sobre a nuvem dos centros de gravidade ponderados é dada por:

$$B = \sum_{i=1}^p \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 \quad (2-8)$$

$$\bar{X} = \left(\frac{1}{n}\right) \sum_{i=1}^p \sum_{j=1}^{n_i} X_{ij} \quad (2-9)$$

$$n = \sum_{i=1}^p n_i \quad (2-10)$$

onde n é o tamanho da amostra e \bar{X} é a média amostral global.

Classificação

A regra de classificação, a partir da função discriminante (Y), que aloca cada indivíduo das amostras em um dos grupos é [15], [2]:

- Aloca o indivíduo (X_0) no grupo π_1 se

$$\hat{Y}_0 = b^T X_0 \Rightarrow \hat{Y}_0 = (\bar{X}_1 - \bar{X}_2)^T S^{-1} X_0 \geq \hat{m} \quad (2-11)$$

onde

$$\hat{m} = \frac{1}{2} (\bar{Y}_1 + \bar{Y}_2) = \frac{1}{2} \left[(\bar{X}_1 - \bar{X}_2)^T S^{-1} (\bar{X}_1 + \bar{X}_2) \right]$$

- Caso contrário, aloca o indivíduo (X_0) no grupo π_2 .

O método de Fisher pode ser estendido para mais de duas populações, mas como isso não faz parte do escopo deste trabalho são sugeridos outros trabalhos [28], [4], [15] e [3] para se obter um melhor aprofundamento do assunto.

Seleção de Medidas

No início de uma análise, dispõe-se de uma grande quantidade de medidas preditoras. Dessa forma, é necessário fazer uma seleção para identificar quais as principais medidas, e que, conseqüentemente, farão parte da função discriminante e da rede neural MLP.

Embora se possa utilizar tantas medidas quanto quisermos, na prática nem todas acrescentam informação no sistema estudado. Existem várias técnicas para selecionar variáveis para o modelo [4], [28], [15], mas neste trabalho será utilizado o procedimento de seleção de variáveis *passo a passo* para a análise discriminante. No caso de dois grupos (que é o relevante para este trabalho), este procedimento é equivalente ao de regressão linear *passo a passo* [4].

A decisão sobre as variáveis que entram e saem do modelo é baseada na denominada estatística F, que é empregada para verificar a adequação do modelo de discriminação. Ela tem este nome porque, sob a hipótese de que as médias de todos os grupos sejam iguais, ela tem uma distribuição F [47], [6]. A estatística F avalia a relação existente entre a variância da função de discriminação $Y = b^T X$ entre grupos (intergrupar) e a variância dentro dos grupos (intragrupar). Quanto maior a diferença entre os grupos, maior será o valor de F. A Tabela 2.1 mostra as equações utilizadas para calcular o valor de F.

Fonte de Variação	Somatório dos Quadrados	Graus de Liberdade	Erro quadrático Médio	Valor de F
Intergrupo	$B = \sum_{j=1}^p n_i (\bar{Y}_i - \bar{Y})^2$	$p - 1$	$M_1 = \frac{B}{p-1}$	$\frac{M_1}{M_2}$
Intragrupo	$W = \sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	$n - p$	$M_2 = \frac{W}{n-p}$	
Total	$S = B + W$ $S = \sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$	$n - 1$		

onde p é o número de grupos, n_i é o número de observações no i -ésimo grupo, n é o número total de observações $\left(\sum_{i=1}^p n_i\right)$, \bar{Y}_i é média da função discriminante para o i -ésimo grupo $\left(\sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i}\right)$, e \bar{Y} é a média global $\left(\sum_{i=1}^p \sum_{j=1}^{n_i} \frac{Y_{ij}}{n}\right)$.

Tabela 2.1: Cálculo da variância e do valor de F.

O procedimento *passo a passo* utiliza, na verdade, a chamada estatística F-parcial. Suponhamos que o processo de discriminação seja feito com base nas variáveis X_1, \dots, X_r e que desejemos examinar se vale a pena introduzir a nova variável X_{r+1} . A estatística F-parcial é definida como $\frac{B_2 - B_1}{W}$, onde W é calculado como na Tabela 2.1, enquanto B_1 e B_2 representam a variância intergrupo para as funções discriminantes calculadas com base nas variáveis X_1, \dots, X_r e X_1, \dots, X_r, X_{r+1} , respectivamente. Deste modo, a diferença $B_2 - B_1$ descreve a redução no erro de classificação ocasionada pela introdução de X_{r+1} . Quanto maior é esta redução, mais atraente é a introdução de X_{r+1} no modelo.

Em cada passo do método, é calculado um valor “F para entrar” para cada variável ainda não incluída no modelo, que corresponde à estatística F-parcial descrita acima. Por outro lado, é calculado um valor “F para sair” para cada variável já incluída no modelo e que corresponde à estatística F-parcial relativa a esta variável, considerando o modelo obtido com sua exclusão. Novas variáveis cujo “F para entrar” seja superior a um valor α_1 previamente especificado são incluídas no modelo, enquanto variáveis cujo “F para sair” seja inferior a um outro valor α_2 são excluídas. O processo termina quando não há novas variáveis a incluir ou excluir.

A Figura 2.10 resume o procedimento de seleção de variáveis *passo a passo* descrito anteriormente.

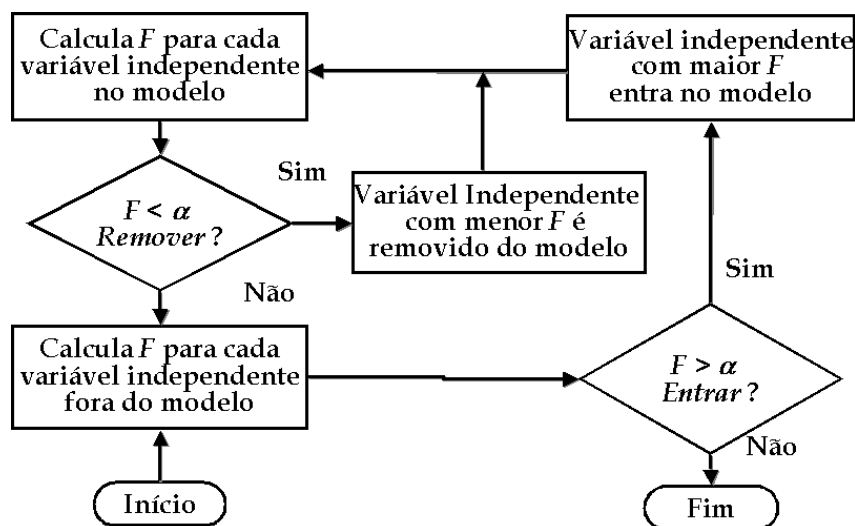


Figura 2.10: Procedimento de seleção de variáveis *passo a passo*.

As medidas selecionadas com o procedimento de seleção de variáveis *passo a passo* para a análise discriminante, serão as mesmas utilizadas como entrada para a Rede Neural Perceptron de Múltiplas Camadas [63].

2.3.2

Redes Neurais Perceptrons de Múltiplas Camadas

Redes neurais artificiais têm sido aplicadas com sucesso nos mais diversos problemas [73], [32], [64], [19]. Embora existam inúmeras arquiteturas de redes neurais, a arquitetura Perceptron de Múltiplas Camadas (*Multilayer Perceptron*) é, sem dúvida, a mais freqüentemente encontrada na literatura. Entre as razões para sua popularidade podemos destacar sua flexibilidade para formar soluções de qualidade para uma ampla classe de problemas, a partir de um mesmo algoritmo de aprendizado.

As Redes Neurais Perceptrons de Múltiplas Camadas (MLP) são arquiteturas nas quais os neurônios são organizados em duas ou mais camadas de processamento, já que sempre vai existir uma camada de entrada e uma de saída. As redes com apenas duas camadas, uma de entrada e outra de saída, apresentam limitações importantes e podem ser aplicadas com sucesso a uma classe restrita de problemas [33]. No entanto, com a utilização da MLP com mais de duas camadas (pelo menos uma escondida), muitas das limitações apresentadas pelo perceptrons foram solucionadas [83]. A Figura 2.11 exemplifica uma rede neural com uma camada escondida. Esta arquitetura é geralmente referida como 3-4-1, ou seja, 3 neurônios de entrada, 4 neurônios escondidos e 1 neurônio de saída. Para generalizar, podemos dizer que uma rede com p entradas, h_1 neurônios na primeira camada escondida, h_2 na segunda camada escondida e q neurônios na camada de saída é descrita por $p-h_1-h_2-q$.

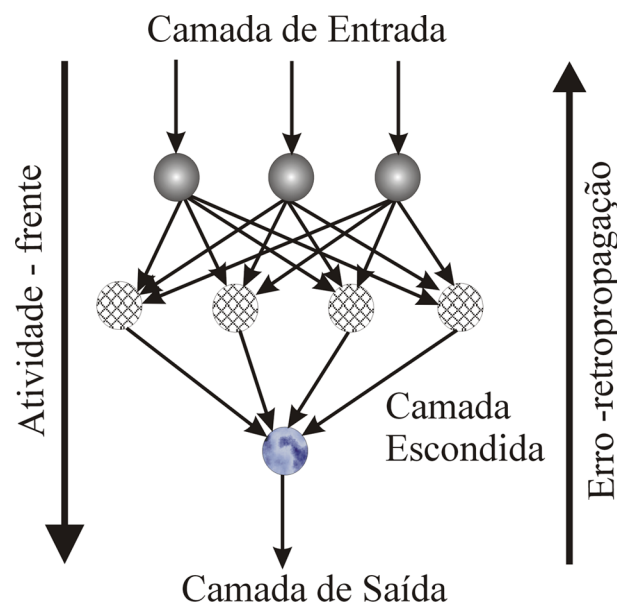


Figura 2.11: Modelo de uma rede MLP (3-4-1).

Algoritmo de Treinamento – Retropropagação (Backpropagation)

A mais importante propriedade de uma rede neural é sua capacidade de aprendizado. Uma rede aprende através de um processo iterativo de ajustes aplicados aos seus pesos sinápticos e limiares.

O processo de aprendizagem de uma rede neural implica na seguinte seqüência de eventos [31]:

1. A rede é estimulada pelo ambiente de informação;
2. A estrutura da rede é alterada como resultado do estímulo;
3. Em virtude das alterações que ocorreram em sua estrutura interna, a rede tem modificada sua resposta aos estímulos do ambiente.

Um tipo particular de aprendizagem que será utilizado neste trabalho é o supervisionado. Esse tipo de aprendizado é caracterizado pela presença de um “professor” externo. A função do “professor” durante o processo é suprir a rede neural com uma resposta desejada a um determinado estímulo.

O algoritmo de aprendizagem por retropropagação (*Backpropagation*) é baseado na regra de aprendizagem por correção de erros. O algoritmo utiliza pares de entradas e saídas desejadas e, por meio de um mecanismo para correção dos erros, ajusta os pesos da rede. Para a minimização do erro obtido pela rede e o ajuste dos pesos, o algoritmo utiliza a regra de delta generalizada, com aplicação do gradiente [83], [63], [34].

Durante o treinamento com o algoritmo de retropropagação, a rede opera em uma seqüência de dois passos. Primeiro, um padrão é apresentado à camada de entrada da rede. A atividade resultante flui através da rede, camada por camada (*feed-forward*), até que a resposta seja produzida pela camada de saída. No segundo passo, a saída obtida é comparada à saída desejada para esse padrão particular. Se esta não estiver correta, o erro é calculado. O erro é propagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo modificados conforme o erro é retropropagado (*feed-backward*).

Os passos abaixo resumem o algoritmo de aprendizagem por retropropagação:

1. Ajustar os pesos dos elementos de processamento com pequenos valores aleatórios.
2. Apresentar as entradas, um vetor x_0, x_1, \dots, x_N de medidas, e especificar um vetor d_1, d_2, \dots, d_N de saída desejado.

3. Calcular as saídas reais da rede, y_1, y_2, \dots, y_N , definida pela equação:

$$y_k = f \left[\sum_{j=1}^m x_{jk}(p) w_{jk}(p) - \theta_k \right], \text{ onde } f \text{ é a função de ativação, } x \text{ é o vetor de entrada, } w \text{ é o vetor peso e } \theta \text{ é o bias.}$$

4. Reajustar os pesos. Usar um algoritmo recursivo começando pelos elementos de processamento de saída, trabalhando para trás no sentido da primeira camada. Os pesos são ajustados através da equação $w_{ij}(t+1) = w_{ij}(t) + \eta \delta_j x'_i$, onde w_{ij} é o peso do elemento de processamento oculto j no tempo t ; x'_i pode ser tanto um elemento de processamento de saída quanto um de entrada; η denota um termo de ganho (velocidade da aprendizagem); e δ_j é um termo de erro para o elemento de processamento j . Se j for um elemento de saída, então $\delta_j = y_j(1 - y_j)(d_j - y_j)$, onde d_j denota a saída desejada e y_j é a saída real da rede; se o elemento j for um elemento oculto, então $\delta_j = x'_j(1 - x'_j) \sum_k \delta_k w_{jk}$, onde k denota todos os elementos acima dos elementos j . Os limiares delta dos elementos internos são ajustados de forma semelhante. A convergência algumas vezes pode ser mais rápida se um termo de momento for adicionado e os pesos alterados de forma mais suave, pela equação: $w_{ij}(t+1) = w_{ij}(t) + \eta \delta_j x'_i + \alpha(w_{ij}(t) - w_{ij}(t-1))$, onde $0 < \alpha < 1$.

5. Repetir retornando para o passo 2.

Uma demonstração mais detalhada do algoritmo de retropropagação pode ser vista em [83], [63], [34].

A regra delta generalizada funciona quando são utilizadas na rede unidades com uma função de ativação semilinear, que é uma função diferenciável e não decrescente. Uma função de ativação amplamente utilizada, nestes casos, é a função sigmóide. Duas funções sigmóide muito utilizadas são a função logística, definida por $(y = \frac{1}{1+e^{-x}})$, e a tangente hiperbólica, definida por $(y = \frac{1-e^{-x}}{1+e^{-x}})$ [63].

A taxa de aprendizagem essencialmente, influencia a magnitude das mudanças dos pesos, desempenhando papel fundamental no desempenho do aprendizado. Uma taxa de aprendizado pequena implica em pequenas variações, tornando o treinamento lento e aumentando as chances de paradas em mínimo locais; altas taxas de aprendizado, no entanto, podem levar a MLP a saturação ou mesmo à oscilação, comprometendo todo o processo de aprendizado. Esta taxa de aprendizagem é introduzida na rede com o objetivo de permitir maior rapidez na convergência ao erro desejado, enquanto o erro estiver diminuindo, e ao mesmo tempo evita que a rede

venha a oscilar, diminuindo a taxa de aprendizagem quando o erro tende a aumentar.

O treinamento das redes MLP com retropropagação pode demandar muitos passos no conjunto de treinamento, resultando num tempo de treinamento consideravelmente longo. Se for encontrado um mínimo local, o erro para o conjunto de treinamento pára de diminuir e estaciona em um valor maior que o aceitável. Uma maneira de aumentar a taxa de aprendizado sem levar à oscilação é modificar a regra delta generalizada para incluir o termo momento, uma constante que determina o efeito das mudanças passadas dos pesos na direção atual do movimento no espaço de pesos [83], [35].

Desta forma, o termo momento leva em consideração o efeito de mudanças anteriores de pesos na direção do movimento atual no espaço de pesos. O termo momento torna-se útil em espaços de erro que contenham longas gargantas, com curvas acentuadas ou vales com descidas suaves [83].

Utilização da rede MLP

Depois que a rede estiver treinada e o erro estiver em um nível satisfatório, a rede poderá ser utilizada como uma ferramenta para classificação de novos dados. Para isto, a rede deverá ser utilizada apenas no modo progressivo (*feed-forward*). Nesta fase, novas entradas são apresentadas à camada de entrada e são processadas nas camadas intermediárias, e os resultados são apresentados na camada de saída, como no treinamento, mas sem a retropropagação do erro. A saída apresentada é o modelo dos dados na interpretação da rede. A Figura 2.11 ilustra este processo.

Limitações da rede MLP

As redes neurais que utilizam retropropagação, assim como muitos outros tipos de redes neurais artificiais, podem ser vistas como “caixas pretas”, nas quais quase não se sabe porque a rede chega a um determinado resultado, uma vez que os modelos não apresentam justificativas para suas respostas. Neste sentido, muitas pesquisas vêm sendo realizadas visando a obtenção de conhecimentos sobre as redes neurais artificiais e a criação de procedimentos explicativos, nos quais se tenta justificar o comportamento das redes em determinadas situações [83], [63], [34].

Outra limitação refere-se ao tempo de treinamento de redes neurais utilizando retropropagação, que tende a ser muito longo. Algumas vezes são necessários milhares de ciclos (épocas) para se chegar a níveis de erros aceitáveis, o que pode demandar um longo período de tempo [31].

Uma terceira limitação é a dificuldade de definir a arquitetura ideal da rede de forma que ela seja tão grande quanto o necessário para conseguir obter as representações internas necessárias e, ao mesmo tempo, pequena o suficiente para apresentar um treinamento rápido. Não existem regras claras para a definição de quantas unidades devem existir nas camadas intermediárias, quantas camadas, ou como devem ser as conexões entre essas unidades [83], [31], [35].

2.3.3

Comparação entre ALDF e MLP

Existem muitas similaridades conceituais entre ALDF e MLP [29]:

- O treinamento de uma MLP é semelhante, no método estatístico, a aprender no modelo da ALDF. Os dois modelos buscam um ajuste dos pesos (parâmetros) baseados no conjunto de dados que são apresentados a eles.
- Em uma rede neural, cada nodo de entrada da rede pode ser visto, na estatística, como uma variável independente, explanatória ou preditiva.
- Existem similaridades entre os pesos da MLP, utilizados nas camadas adjacentes, para o cálculo da saída com os chamados coeficientes de regressão em estatística.
- O *bias*, que nas MLP tem o efeito de aumentar ou diminuir a entrada líquida da função de ativação, dependendo de se ele é positivo ou negativo, em estatística é conhecido como “intercepto”.
- O erro em uma MLP é calculado através da diferença entre a saída real e a saída desejada da rede. Essa diferença (erro) é semelhante ao conceito de minimização de resíduos na regressão estatística.
- No modelo de ALDF, o processo converge quando a função de probabilidade é maximizada, enquanto em MLP a função de erro dos mínimos quadrados é minimizada.

A Tabela 2.2 resume os principais termos semelhantes nos dois modelos estudados.

MLP	ALDF
Treinamento, aprendizagem	Estimação de parâmetros
Unidades de entrada	Variáveis independentes, explanatórias, preditivas
Camada de saída	Variável dependente, valores previstos
Pesos nas conexões	Coefficientes de regressão
Bias	Intercepto
Erro	Resíduo
Casos de treinamento, padrões	Observação
Características	Variáveis

Tabela 2.2: Termos similares entre MLP e ALDF.

2.4

Validação do Modelo

A validação do modelo com o próprio conjunto de dados que serviu para fazer o treinamento do modelo classificador induz uma estimativa de qualidade pouco realista. Para evitar esta validação tendenciosa, é necessário dividir (reamostrar) o conjunto de dados original em um para treinamento e outro para teste.

Deixa um de fora é um caso especial de reamostragem que é uma técnica elegante para estimar taxas de erros de classificador [10]. Como é computacionalmente cara, é frequentemente reservada para problemas em que o tamanho da amostra é relativamente pequeno. Para uma amostra de tamanho n , um classificador é projetado usando $(n - 1)$ casos e testado no único caso restante. Isto é repetido n vezes, cada vez gerando um classificador e deixando um de fora. Assim, cada caso na amostra é usado como um caso de teste, e os demais são usados para projetar o classificador. A taxa de erro é o número de erros dividido por N . A Figura 2.12 ilustra esta técnica.

2.5

Curva ROC (Receiver Operating Characteristic)

A avaliação dos métodos propostos neste trabalho pode ser feita por comparação com técnicas de referência que se saibam serem válidas. Tal avaliação envolve, portanto, a comparação de medidas obtidas simultaneamente, utilizando o teste em estudo e um teste de referência. Os estudos de avaliação implicam que esse teste de referência seja o apropriado. Um dos grandes problemas inerentes a este tipo de estudo é o fato de, por

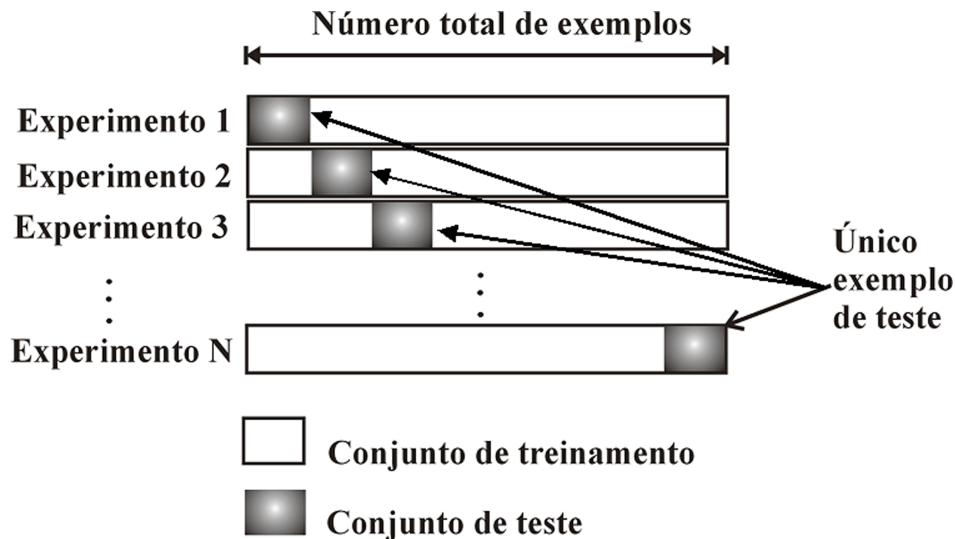


Figura 2.12: Exemplo da técnica *deixa um de fora*.

vezes, não existir uma referência, usando-se, então, o melhor procedimento disponível como procedimento de referência. Mais uma vez, é importante frisar que uma medida é válida se provém de um procedimento válido.

Os estudos de avaliação são freqüentemente descritos como testes de validade dos diagnósticos e são um dos mais importantes atos em Medicina. Para elaborar um diagnóstico, temos que utilizar métodos que permitam distinguir entre populações de doentes e de não doentes, ou seja, teste de diagnóstico.

Nos testes de diagnóstico o resultado é sempre dicotômico. Quando se avaliam esses testes, utilizamos um teste de referência cuja escala é também dicotômica. A validade de medidas dicotômicas pode ser avaliada construindo uma tabela de 2×2 (Tabela 2.3) [48].

		Doença	
		Presente	Ausente
Teste	Positivo	Verdadeiro Positivo	Falso Positivo
	Negativo	Falso Negativo	Verdadeiro Negativo

Tabela 2.3: Relação entre o resultado de um teste diagnóstico e o verdadeiro diagnóstico

A Tabela 2.3 evidencia que há dois tipos de conclusão errônea em um teste: Falso Positivo (indivíduo não doente é considerado como doente) e Falso Negativo (indivíduo doente é considerado normal).

2.5.1

Sensibilidade e Especificidade

O valor clínico de um teste está relacionado com a sua especificidade e sensibilidade. Ele deve fornecer uma boa indicação preliminar de quais indivíduos têm a doença e quais não têm, e isto só se consegue se os métodos utilizados forem válidos.

A sensibilidade é a proporção de indivíduos doentes que possuem um teste positivo, isto é, a probabilidade de, estando doente, um indivíduo ter um teste positivo (percentagem de vezes que o teste acerta). A especificidade é a proporção de indivíduos não doentes que possuem um teste negativo ou a probabilidade de, não estando doente, ter um teste negativo. A sensibilidade define-se, então, como sendo a capacidade de um teste para identificar corretamente aqueles indivíduos que possuem uma determinada doença, enquanto que a especificidade é definida como a capacidade do teste para identificar corretamente aqueles que não a possuem. Ambas são determinadas pela comparação dos resultados obtidos num determinado teste com os resultados de métodos de diagnóstico mais seguros (de referência). A extensão em que os resultados de um teste coincidem com o de referência dá uma medida da sensibilidade e especificidade desse teste [26], [8].

Quando indivíduos doentes são considerados negativos ou normais, os respectivos resultados deste teste são chamados “falsos negativos”. Por outro lado, quando indivíduos não doentes são considerados como doentes, os resultados deste teste são denominados “falsos positivos”. Note-se que a percentagem de falsos negativos é o complemento da sensibilidade e a percentagem de falsos positivos é o complemento da especificidade. Quando a sensibilidade é de 100%, temos a certeza que o teste nunca se engana nos falsos negativos.

A especificidade e a sensibilidade não provêem informação sobre os falsos positivos e os falsos negativos. São independentes da prevalência da doença (proporção de indivíduos doentes ou probabilidade de estar doente, independentemente do resultado do teste - probabilidade pré-teste) e esta é considerada a sua maior vantagem [48].

A Tabela 2.4 mostra a relação da especificidade e sensibilidade e como determinar seus valores [8].

Sensibilidade = $\frac{a}{a+b}$ = verdadeiros positivos / todos os doentes

Especificidade = $\frac{d}{c+d}$ = verdadeiros negativos / todos os não doentes

Precisão = $\frac{a+d}{a+b+c+d}$ = corretamente classificados / todos

		Doença		
		Presente	Ausente	Total
Teste	Positivo	a	c	a+c
	Negativo	b	d	b+d
	Total	a+b	c+d	N

Tabela 2.4: Cálculo da especificidade e sensibilidade para uma variável dicotômica

2.5.2

Cálculo da Curva ROC

Geralmente, a sensibilidade e a especificidade são características difíceis de conciliar, isto é, é complicado aumentar a sensibilidade e a especificidade de um teste ao mesmo tempo. As curvas ROC (*Receiver Operating Characteristic*) são uma forma de representar a relação, normalmente antagônica, entre a sensibilidade e a especificidade de um teste diagnóstico quantitativo ao longo de valores contínuos de ponto de corte [84].

Para construir uma curva ROC traça-se um diagrama que represente a sensibilidade em função da proporção de falsos positivos (1- especificidade) para um conjunto de valores de ponto de corte.

Quando se tem uma variável contínua, resultado da aplicação de um teste diagnóstico quantitativo, e se pretende transformá-la numa variável dicotômica, do tipo doente/não doente, temos que utilizar um determinado valor na escala contínua que discrimine entre essas duas classes. A esse valor dá-se o nome de “ponto de corte” (*cut off point*).

O valor escolhido como ponto de corte vai influenciar as características do teste, como exemplificado na Figura 2.13. Neste exemplo, quanto maior o ponto de corte, maior a especificidade do teste, mas a sensibilidade será menor; e quanto menor o ponto de corte, maior a sensibilidade, mas a especificidade será menor [48]. A Figura 2.14 representa graficamente a relação entre a sensibilidade e a especificidade para todos os possíveis pontos de corte da curva C_1 , C_2 e C_3 . Quanto maior for a sobreposição das curvas normais, menor será a área sob a curva ROC.

As curvas ROC descrevem a capacidade discriminativa de um teste diagnóstico para um determinado número de valores de ponto de corte. Isso permite colocar em evidência os valores para os quais existe uma maior otimização da sensibilidade em função da especificidade. O ponto numa curva ROC em que isso acontece é aquele que se encontra mais próximo do canto superior esquerdo do diagrama.

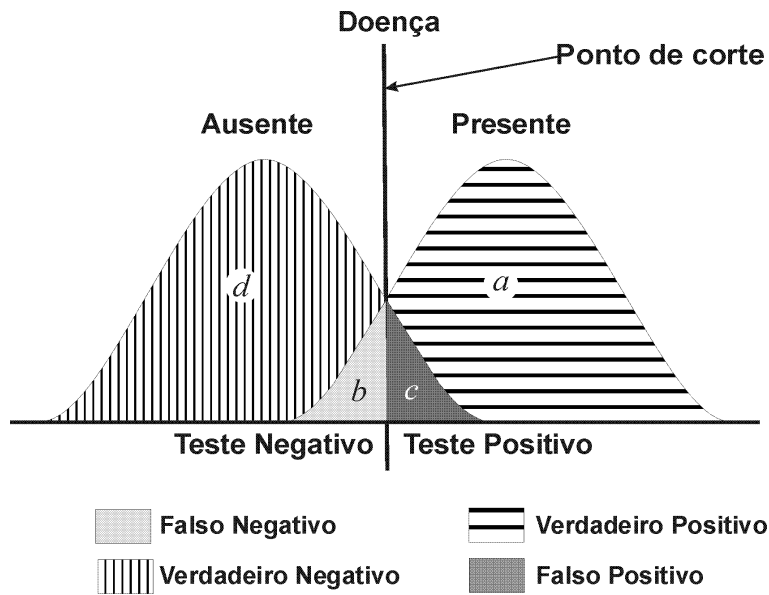


Figura 2.13: Ponto de corte.

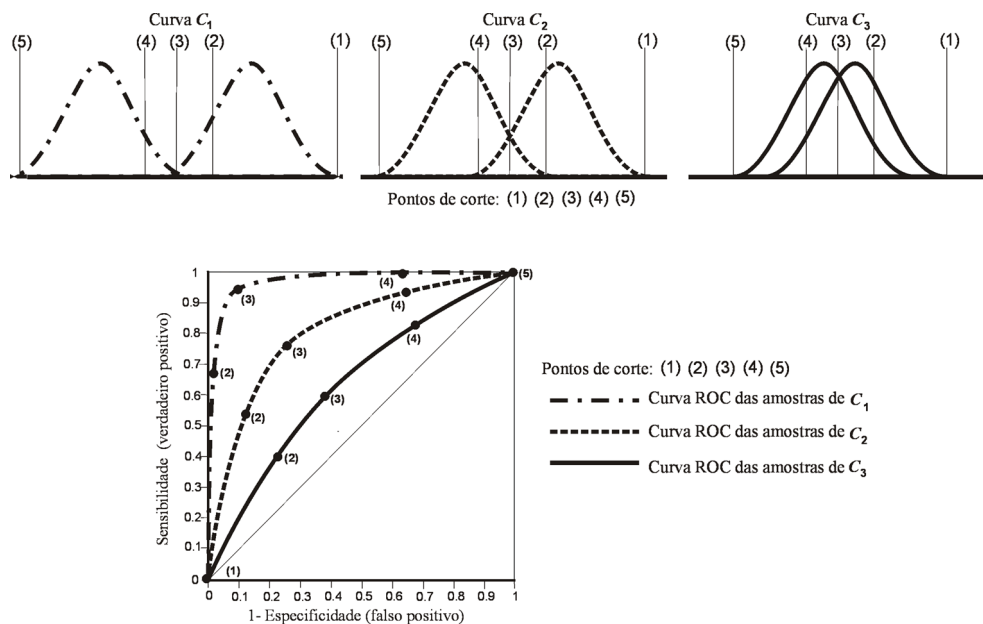


Figura 2.14: Relação entre a curva ROC e pontos de corte.

Por outro lado, as curvas ROC permitem quantificar a exatidão de um teste diagnóstico, já que esta é proporcional à área sob a curva ROC (AUC), isto é, ela será tanto maior quanto mais a curva se aproximar do canto superior esquerdo do diagrama. Em virtude disso, a curva será útil também na comparação de testes diagnósticos, que terá uma exatidão tanto maior quanto maior for a área sob a curva ROC. O valor da área igual a 1 representa um teste perfeito; a área igual a 0.5 representa um valor sem importância. A Figura 2.15 exemplifica várias curvas ROC e a Tabela 2.5 associa a qualidade do diagnóstico à área da curva ROC [74], [48], [10].

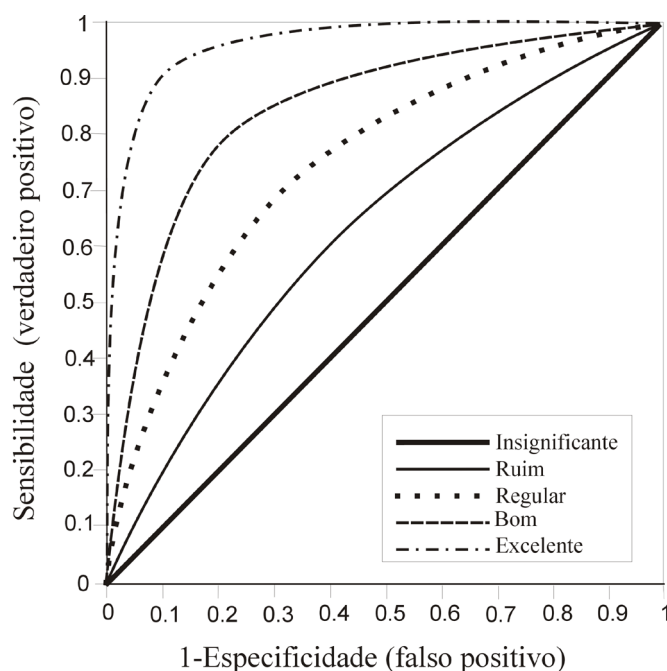


Figura 2.15: Curva ROC e a qualidade no diagnóstico.

Área (AUC)	Qualidade do diagnóstico
0.9 a 1.0	Excelente
0.8 a 0.9	Bom
0.7 a 0.8	Regular
0.6 a 0.7	Ruim
0.5 a 0.6	Insignificante

Tabela 2.5: Qualidade do diagnóstico em relação à área da curva ROC.

A área do curva ROC é comumente calculada através de dois métodos [84], [48], [26]:

1. Não paramétrico: se caracteriza por não fazer nenhuma suposição sobre as distribuições dos resultados do teste. Duas técnicas

geralmente utilizadas para o cálculo da área da curva são a regra do trapézio e a aproximação à estatística U de Wilcoxon-Mann-Whitney.

2. Paramétrico: se baseia em supor uma determinada distribuição para os resultados do teste. O modelo mais freqüentemente utilizado é o binormal, que supõe a normalidade das variáveis com probabilidade positiva e negativa. Utiliza o estimador de máxima verossimilhança para ajustar uma curva suave aos pontos.

Hanley e McNeil [10] descreveram um método não paramétrico para o cálculo da área da curva ROC (AUC), utilizando a aproximação à estatística U de Wilcoxon-Mann-Whitney. Com o resultado da área curva calculada por esse método, o erro padrão (SE) também pode ser estimado.

A estatística U de Wilcoxon-Mann-Whitney mede se as seqüências de casos normais e anormais podem ter vindo da mesma população ou não. Em relação à curva ROC, esse método testa se as distribuições são as mesmas ou diferentes.

Considere-se uma amostra de dimensão n_A para os indivíduos classificados como anormais, A , e outra de dimensão n_N para os indivíduos classificados como normais, N ; o procedimento de teste consiste em fazer todas as $n_A n_N$ comparações possíveis entre os valores x_A da amostra n_A e os valores x_N da amostra n_N , graduando cada comparação de acordo com a regra,

$$S(x_A, x_N) = \begin{cases} 1 & \text{se } x_A > x_N \\ 1/2 & \text{se } x_A = x_N \\ 0 & \text{se } x_A < x_N \end{cases}$$

e fazendo a média dos S' s para todas as $n_A n_N$ comparações, vem:

$$AUC = W = \frac{1}{n_A n_N} \sum_{i=1}^{n_A} \sum_{j=1}^{n_N} S(x_A, x_N) \quad (2-12)$$

que é uma estatística que não depende dos valores de x , mas apenas das graduações, designada como estatística de Wilcoxon-Mann-Whitney [10].

Como cada comparação é classificada por 1, 1/2 ou 0, o valor médio de W estará entre 0 e 1, e reflete, como não poderia deixar de ser, qual a proporção de x'_A s que são maiores que x_N .

Como nem todas as $n_A n_N$ comparações são independentes, incluir todas é mera conveniência, e o erro padrão de W tem em conta esta possível intercorrelação [10]. Assim, a probabilidade de atribuir uma classificação correta é igual à média ponderada de todas as combinações de pares de classificações possíveis.

As áreas das curvas ROC de dois ou mais procedimentos (métodos) são freqüentemente utilizadas para comparação e determinação de qual deles é mais preciso no diagnóstico. Essa comparação tem como objetivo verificar se existe diferença significativa entre as curvas, ou seja, entre os procedimentos.

O método proposto por Hanley e McNeil [11] para determinar a diferença (comparação) entre as curvas utiliza o valor crítico de z :

$$z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}} \quad (2-13)$$

onde A_1 e SE_1 referem-se a área observada e o erro padrão estimado da curva ROC do procedimento 1; A_2 e SE_2 referem-se a área observada e o erro padrão estimado da curva ROC do procedimento 2; e r representa correlação estimada entre A_1 e A_2 .

O valor de z é então verificado na tabela da distribuição normal, e o valor de z acima de algum limiar, por exemplo $z \geq 1.96$, é uma evidência que as áreas das curvas ROC são diferentes ($p < 0.05$). Desta forma, a hipótese nula de que não há diferença entre as áreas das curvas não é satisfeita.

2.6 Resumo

A Seção 2.1 deu uma visão geral de alguns conceitos importantes para a análise de uma imagem médica, como a aquisição da imagem, formas de tratamento de imagens em Computação Gráfica, o padrão DICOM e a técnica de interpolação linear.

Na Seção 2.2 foi dada uma visão geral do Nódulo Pulmonar Solitário (NPS) e foi mostrada a importância de se identificar e diagnosticar esses nódulos precocemente, para aumentar a chance de cura do paciente. Também foram abordados aspectos de textura e forma dos NPS que ajudam os médicos a diagnosticá-los como benignos ou malignos.

Na Seção 2.3, foram estudadas duas técnicas de classificação que determinarão a benignidade ou malignidade do NPS. A primeira técnica é chamada Análise Discriminante Linear de Fisher – ALDF. Para esta técnica foi apresentada a teoria básica para análise, aprendizagem e classificação entre dois grupos, como considerações iniciais para utilizar a ADLF, testes estatísticos necessários para as considerações a serem atendidas e a função discriminante de Fisher. A segunda técnica chama-se Rede Neural Perceptron de Múltiplas Camadas – MLP. Para a MLP foi apresentado o algoritmo de treinamento mais utilizado, retropropagação,

além da utilização da rede após o treinamento e suas limitações. Em seguida, foi feita uma breve comparação entre MLP e ADLF, mostrando aspectos similares entre ambas. Para finalizar, foi descrito um procedimento *passo a passo* que visa selecionar as medidas mais significativas para fazerem parte da Análise Discriminante Linear de Fisher e da Rede Neural Perceptron de Múltiplas Camadas.

Depois, na Seção 2.4, foi descrita uma técnica de validação do modelo, chamada *deixa um de fora*, que avalia mais realisticamente os modelos encontrados.

Por último, na Seção 2.5, foi abordada a Curva ROC (*Receiver Operating Characteristic*), que é uma técnica freqüentemente utilizada por médicos para avaliação de diagnósticos e algoritmos. Foi mostrado como se determina esta curva, assim como a importância da sua área (*AUC*) para a comparação entre diversos diagnósticos.