# Análise Discriminante e Classificação de Imagens 2D de Ultrassonografia Mamária

Albert C. Xavier[1], João R. Sato[2], Gilson A. Giraldi[3] e Carlos E. Thomaz[1]

[1]FEI, São Bernardo do Campo, SP, Brasil
[2]UFABC, Santo André, SP, Brasil
[3]LNCC, Petrópolis, RJ, Brasil

## Abstract

*Research in detection and treatment of cancer using images, including breast cancer, has improved in the last years. Clinical experts say that breast cancer is the most common cause of cancer death among women worldwide. This paper approaches this problem by carrying out a statistical discriminant analysis of mammary tumors in 2D ultrasound images. To analyze the images, univariate and multivariate statistical methods have been investigated with the aim of extracting discriminant information for classification purposes. Our results show a good classification performance of the multivariate statistical methods when using preprocessed and segmented 2D ultrasound images as a whole. Additionally, our approach highlights the most statistically significant differences between the tumors, ranking in the original image space the simplest and most difficult cases of classification.*

## 1. Introdução

Na área médica, a classificação de imagens é utilizada para auxiliar no diagnóstico de vários tipos de doença, incluindo o câncer de mama. O câncer de mama tem registrado crescimento muito preocupante nos últimos anos e se posiciona como maior causador de morte por câncer na população feminina mundial. No Brasil, 33 mulheres morrem por dia em decorrência do câncer de mama [1].

Nesse contexto, os recursos computacionais atuais e o processamento de imagens, aliados à estatística, podem contribuir muito para o esclarecimento e um diagnóstico mais preciso do câncer de mama. Há um crescente avanço nos métodos de diagnóstico por imagem incluindo a utilização de ultrassom. Na imaginologia mamária, a ultrassonografia representa uma modalidade de diagnóstico muito significativa por ter menor custo e ainda evitar o contato prejudicial dos pacientes com radiação [7].

Este trabalho tem como objetivo principal analisar as alterações morfológicas estatisticamente relevantes para classificação das imagens de tumor de mama. Tal análise é essencial para o estudo das diferenças entre os grupos de tumores benignos e malignos. São realizadas análises no espaço de características, definido por quantidades geométricas e de textura, bem como no espaço de imagens, comparando-se os resultados.

## 2. Análise Discriminante Multivariada

Técnicas em estatísticas multivariadas como LDA (Linear Discriminant Analysis) e SVM (Support Vector Machine) podem ser utilizadas na etapa de classificação [2, 4, 5, 11, 14]. Nestas aplicações , cada amostra é representada por um ponto no espaço $n$-dimensional, onde $n$ é o número de variáveis do problema em questão. A denominação multivariada corresponde às técnicas que utilizam todas as características (variáveis) para interpretação do conjunto de dados.

### 2.1. LDA

A proposta do método LDA, também conhecido como método de Fisher [2], é encontrar o hiperplano de maior separação entre os grupos analisados. O cálculo desse hiperplano de separação considera o conhecimento prévio da classe ou grupo de cada amostra. A análise LDA é paramétrica, ou seja, considera que a distribuição de probabilidade das amostras é conhecida e pode ser representada pela média e dispersão das amostras.

O método baseia-se na diminuição do espalhamento das amostras com relação ao grupo a qual pertencem e, também, na maximização da distância da média entre estes grupos [2]. Em outras palavras, calcula-se as matrizes de espalhamento inter-classes e intra-classes com objetivo de discriminar os grupos de amostras pela maximização da separabilidade entre classes enquanto minimiza-se a variabilidade dentro das mesmas.

Matematicamente, as matrizes de espalhamento inter-classes ($S_b$) e intra-classes ($S_w$) são definidas como:

$$S_b = \sum_{i=1}^{g} N_i(\overline{\mathrm{x}}_i - \overline{\mathrm{x}})(\overline{\mathrm{x}}_i - \overline{\mathrm{x}})^T, \quad (1)$$

$$S_w = \sum_{i=1}^{g}\sum_{j=1}^{N_i}(\mathrm{x}_{i,j} - \overline{\mathrm{x}}_i)(\mathrm{x}_{i,j} - \overline{\mathrm{x}}_i)^T, \quad (2)$$

onde $g$ é o número de grupos analisados, $N_i$ a quantidade de amostras do grupo $i$, $\bar{\mathrm{x}}$ e $\bar{\mathrm{x}_i}$ são a média total e a média das amostras do classe $i$, respectivamente, e $\mathrm{x}_{i,j}$ é a amostra $j$ do grupo $i$. E a relação proposta por Fisher que deve ser maximizada é dada pela equação (3) a seguir:

O principal objetivo do método LDA é encontrar a matriz de projeção $P_{LDA}$ que maximiza a razão entre o determinante da matriz de espalhamento inter-classes $S_b$ e o determinante da matriz de espalhamento intra-classes $S_w$, conhecido como critério de Fisher e descrito matematicamente por

$$P_{LDA} = \arg\max \frac{|S_b|}{|S_w|}. \quad (3)$$

O critério de Fisher descrito pela Equação 3 é satisfeito quando a matriz de projeção $P_{LDA}$ é composta, no máximo, pelos $(g-1)$ autovetores de $S_w^{-1}S_b$, cujos autovalores correspondentes são não-nulos.

Na prática, essa razão somente pode ser calculada se a matriz de espalhamento intra-classe $S_w$ for não-singular. Exatamente nesse momento deve ser observado a proporção entre o número de amostras e de variáveis. Em cenários onde o número total de amostras $N$ é bem menor que o número de variáveis $n$, ocorre uma instabilidade no cálculo da matriz inversa de $S_w$ [3]. A quantidade de amostras necessárias para evitar essa instabilidade no cálculo da matriz inversa de $S_w$ deve ser igual ou superior a 5 vezes a quantidade de variáveis destas [6]. Considerando o cenário do trabalho desenvolvido aqui, certamente ocorreria esse problema, pois a quantidade de amostras é bem inferior a quantidade de variáveis de cada amostra, representada por imagens de tumores.

Portanto, para o tratamento do problema de instabilidade no cálculo da inversa da matriz $S_w$, utilizou-se o método denominado MLDA (*Maximum uncertainty Linear Discriminant Analysis*) [13]. Essa técnica consiste em substituir $S_w$ por outra matriz regularizada $S_w^*$, gerando um aumento no espalhamento dos dados e mantendo as variações mais relevantes existentes nas amostras. A nova matriz regularizada $S_w^*$ pode ser calculada por meio dos seguintes passos:

1. Selecionar os autovetores $\Phi$ e autovalores $\Lambda$ de $S_p$, onde $S_p = \frac{S_w}{N-g}$;
2. Calcular a média dos autovalores $\bar{\lambda}$;
3. Gerar uma nova matriz de autovalores baseada na dispersão dos maiores $\Lambda^* = diag[max(\lambda_1, \bar{\lambda}), ..., max(\lambda_n, \bar{\lambda})]$;
4. Calcular a matriz de espalhamento intra-classes regularizada $S_w^* = (\Phi\Lambda^*\Phi^T)(N-g)$.

Com a matriz $S_w^*$ calculada, substitui-se $S_w$ da equação (3) por $S_w^*$ e regulariza-se o critério de Fisher para problemas onde $N \ll n$.

## 2.2. SVM

Com o mesmo propósito final do LDA, o método SVM também visa encontrar o hiperplano de maior separação entre os grupos de amostras. De forma análoga ao LDA, o cálculo do hiperplano de separação considera o conhecimento prévio da classe de cada amostra $j$ investigada, ou seja, $y_j \in \{-1, 1\}$. Porém, o método SVM é não-paramétrico, ou seja, não considera a distribuição de probabilidade das amostras.

O SVM é uma técnica de reconhecimento de padrões com sólido embasamento teórico e que tem apresentado resultados muito satisfatórios mesmo quando comparado a métodos clássicos como redes neurais e árvores de decisão. Trata-se essencialmente de um classificador de duas classes mas que também pode ser estendido para tratamento de mais de duas classes [8]. Este método é baseado na teoria de aprendizado estatístico, pioneiramente descrita por Vapnik e colaboradores [14]. Como vantagens, o SVM apresenta boa capacidade de generalização, robustez em grandes dimensões e convexidade da função objetivo. Para encontrar a solução ótima do classificador é usada uma função quadrática, em que não há presença de vários mínimos locais, e sim apenas um mínimo global, o que facilita a obtenção do valor ótimo.

O hiperplano SVM pode ser definido resumidamente como:

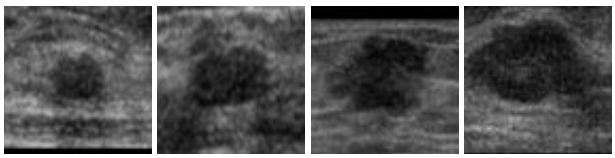
$$w_{SVM} = \sum_{j=1}^{N} \alpha_j y_j \mathrm{x}_j, \quad (4)$$

onde $\alpha_j$ são os coeficientes de Lagrange não-negativos obtidos pela solução de um problema de otimização quadrático com restrições de desigualdade linear [8]. As observações de treinamento $\mathrm{x}_j$, com $\alpha_j$ não-zero, ficam na fronteira da margem e são chamadas de vetores de suporte [11]. O SVM pode fazer uso de vários tipos de *kernel*, incluindo o polinomial, o Gaussiano-RBF e o linear. Para os experimentos desenvolvidos no presente trabalho, foi adotado somente o *kernel* linear.

## 3. Arcabouço Experimental

### 3.1. Banco de Imagens e Características

Foram utilizadas 250 imagens ultrassonográficas de tumores mamários e um conjunto de valores referentes às características de circularidade, sombra acústica e heterogeneidade destes tumores. As imagens de tumores benignos totalizam 100 e as de tumores malignos 150. Como existem 5 imagens diferentes do mesmo tumor, tem-se 20 diferentes tumores benignos e 30 diferentes tumores malignos. A Figura 1 apresenta exemplos das imagens investigadas aqui. As images foram adquiridas por meio de um equipamento modelo Voluson 730 (General Electric, USA) com um transdutor S-VNW5-10. As especificações técnicas deste são: frequência de varredura de 5-10 MHz, largura de varredura de 40 mm e ângulo de varredura de 20 a 30 graus.



(a) (b) (c) (d)

**Figura 1. Exemplos de imagens investigadas com tumores (a,b) benignos e (c,d) malignos.**

Os valores geométricos de circularidade foram obtidos a partir de um ponto central da região de interesse (tumor). Calculou-se a distância entre cada ponto da borda da lesão e o centro da imagem. Os valores também foram normalizados pela área total da imagem. Em geral, as lesões malignas apresentam valores mais altos de desvio padrão em relação a distância média quando comparadas às lesões benignas. As outras duas características, heterogeneidade e sombra acústica, são de textura. Os valores de heterogeneidade, utilizando imagens em escala de cinza, foram calculados por meio da entropia BGS (*Boltzman-Gibbs-Shannon*). Geralmente as lesões malignas são mais heterogêneas que as benignas. Já a sombra acústica tem relação direta com a região inferior da imagem. No caso das imagens de tumores mamários, a sombra acústica está extremamente relacionada ao tipo de tumor. Na maioria dos casos de tumor benigno ocorre a formação de um reforço acústico abaixo do região do tumor em decorrência da existência de muitas partículas de água. Os tumores malignos, que são geralmente mais sólidos, tendem a apresentar uma sombra acústica. Nos tumores malignos a sombra acústica é mais intensa (cor mais branca) que o reforço acústico presente nos tumores benignos. Então, para calcular os valores dessa característica foram comparados os histogramas da região da lesão e da região logo abaixo desta. Quanto mais escuro é a região abaixo da lesão, maior é a probabilidade desta ser benigna [10]. Os valores de pré-processamento das características foram atribuídos por radiologistas e calculados em estudos anteriores [9, 4].

### 3.2. Pré-processamento

Com o propósito de ajustar as diferenças de resolução, as imagens foram submetidas a uma etapa de pré-processamento. As imagens fornecidas, em escala de cinza, possuem resoluções diferentes que variam de 57 a 161 pixels na altura e de 75 a 199 pixels na largura. Independentemente da resolução, a maioria dos tumores está localizada no centro das imagens. Cada pixel (ponto) da imagem é representado dentro de uma escala de 0 a 255, que são as variações de tons de cinza em um sistema de representação de luminância com 8 bits de resolução. A grande maioria das imagens foi e recortadas para adequação ao processo de análise. A resolução de 70 x 70 pixels, escolhida após testes, visou evitar a perda de informações significativas no recorte das imagens. No entanto, 20 imagens de um total de 250 tiveram que ser redimensionadas pois possuiam resolução inferior a 70 pixels na altura. A maior diferença encontrada, em 5 dessas 20 imagens, foi de 13 pixels. Nas demais imagens redimensionadas o maior ajuste foi de 5 pixels.

### 3.3. Experimentos

As imagens foram submetidas primeiramente a um processo de redução de dimensionalidade utilizando a técnica PCA (*Principal Component Analysis*) [3], porém preservando todas as componentes com autovalores não-nulos [12]. Na sequência, as imagens projetadas no espaço do PCA foram então projetadas no espaço dos classificadores lineares MLDA e SVM para se obter a separação dos tumores malignos e benignos.

Neste processo de separação linear empregando os classificadores MLDA e SVM, foram adotados dois tipos de parâmetros de entrada. No primeiro a classificação das imagens de tumores mamários considera a intensidade dos pixels, ou seja, a análise é feita na imagem como um todo considerando todas as componentes do PCA com autovalores não-nulos. Já no segundo processo os classificadores são executados utilizando as características extraídas dos tumores nas imagens (circularidade, heterogeneidade e sombra acústica). Os valores atribuídos a cada característica são tratados separadamente nos classificadores.

O cálculo da acurácia dos classificadores foi baseado na abordagem de validação cruzada (*cross-validation*). Nos experimentos aqui realizados, foi adotada a forma extrema da abordagem de validação cruzada denominada *leave-one-out* [3]. Este método foi aplicado em decorrência do pequeno número de amostras rotuladas.

## 4. Resultados

### 4.1. Análise Estatística Univariada

A Figura 2[1] apresenta as imagens médias das amostras dos grupos benigno e maligno. Claramente observa-se uma diferença no formato e tamanho dos tumores nas imagens. Na imagem média das lesões benignas, vista na Figura 2(a), o tumor é menor e mais concentrado. Já o tumor das imagens de lesões malignas, visto na Figura 2(b), parece ser maior e mais espalhado. Na sequência, por meio de uma operação matemática simples, foi feita a subtração das imagens médias dos dois grupos com o propósito de visualizar possíveis regiões discriminantes. O resultado apresentado na Figura 2(c) indica que existe uma possível diferença entre as imagens na região inferior e também na região central direita. Provavelmente a sombra do tumor exerce influência sobre estas áreas. Em complemento às diferenças observadas na subtração das imagens médias, foi implementado o teste de hipóteses baseado na distribuição de probabilidade de *t-student*. Nesse experimento, a finalidade foi obter as variações mais relevantes estatisticamente entre as imagens dos grupos de tumores malignos e benignos. Na Figura 2(d), em cor azul, pode-se observar as regiões mais discriminantes selecionadas de acordo com um nível de confiança de 99.9% (ou $p < 0.001$) da tabela *t-student*. Do ponto de vista estatístico, estas são as regiões de maior discriminância entre os dois grupos de imagens de tumor mamário, avaliadas pixel-por-pixel. Essas regiões mais discriminantes foram projetadas sobre uma imagem de referência que representa aqui a subtração das imagens médias dos dois grupos.

### 4.2. Desempenho dos Classificadores

A Tabela 1 apresenta a acurácia dos classificadores nas análises das imagens como um todo e das características geométrica e de textura. O desempenho dos classificadores considerando a intensidade de todos os pixels da imagem simultaneamente foi notadamente superior ao desempenho desses mesmos classificadores analisando as características extraídas das imagens. Esses resultados indicam que os fatores determinantes na classificação dos grupos vão além de características como circularidade, heterogeneidade e sombra acústica, principalmente para separações lineares.

| | MLDA | | | SVM | | |
|---|---|---|---|---|---|---|
| | Total | Benignos | Malignos | Total | Benignos | Malignos |
| Imagem | 100% | 100% | 100% | 100% | 100% | 100% |
| Circularidade | 78% | 79% | 77% | 76% | 68% | 82% |
| Heterogeneidade | 64% | 63% | 65% | 60% | 46% | 70% |
| Sombra Acústica | 73% | 60% | 82% | 70% | 61% | 75% |

**Tabela 1. Taxas de classificação.**

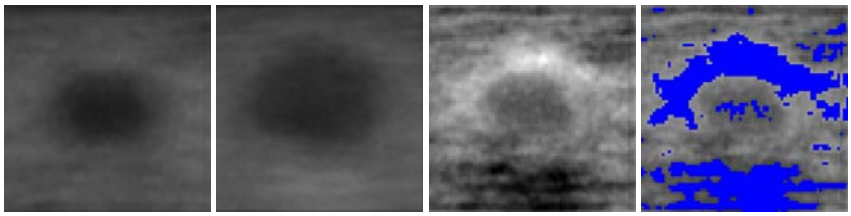
Pode-se constatar que os dois classificadores separaram muito bem os grupos na análise simultânea de todos os pixels da imagem. A taxa de acerto foi de 100% em ambos classificadores. Em contraste, a classificação considerando os valores das características apenas não apresentou resultado satisfatório. A maior taxa de acerto, de 78%, foi obtida na avaliação da característica circularidade pelo classificador MLDA.

A Figura 3 apresenta os hiperplanos descritos pelos dois classificadores com o intuito de observar que a ordem de classificação das amostras, no entanto, ficou diferente entre esses classificadores. A imagem 95, de tumor benigno, ficou mais próxima dos casos malignos no hiperplano classificador do SVM. Uma possível explicação para tal posicionamento seria a presença de características inerentes aos tumores malignos como a sombra acústica e o formato irregular. Já a imagem 93, também do grupo de tumores benignos (lado direito), foi ordenada como um dos extremos do grupo indicando uma classificação mais fácil e/ou mais simples. De forma análoga, também destaca-se a imagem 206 do grupo de tumores malignos. Esta ficou posicionada no extremo do grupo de tumores malignos (lado esquerdo). É valido destacar ainda a posição da imagem 166 no hiperplano gerado pelo classificador MLDA. Esta imagem foi ordenada como próxima aos tumores malignos, sugerindo uma maior complexidade de classificação. Observando a imagem 166, nota-se que esta apresenta características dos dois grupos de tumores, ou seja, esta apresenta características de tumor benigno como a ausência de sombra acústica, mas também apresenta formato irregular, geralmente característico dos tumores malignos. A proximidade da fronteira de decisão definida pela ordem das imagens pode ser interpretada como uma maior dificuldade de classificação. Já a maior distância desta pode indicar uma maior facilidade de classificação.

### 4.3. Navegação nos Hiperplanos

A Figura 4(a) apresenta uma padronização dos formatos de nódulos mamográficos proposta pelo BI-RADS (*Breast Image Reporting and Data System*). Nesta é possível observar a transição do nódulo benigno para o maligno. A

1 Para visualizar as imagens em cores, favor acessar a referência eletrônica do artigo.



(a) (b) (c) (d)

**Figura 2. Análise estatística univariada: (a) Imagem média de tumores benignos; (b) Imagem média de tumores malignos; (c) Imagem da subtração de (a) por (b); (d) Imagem das diferenças estatísticas ($p < 0.001$).**

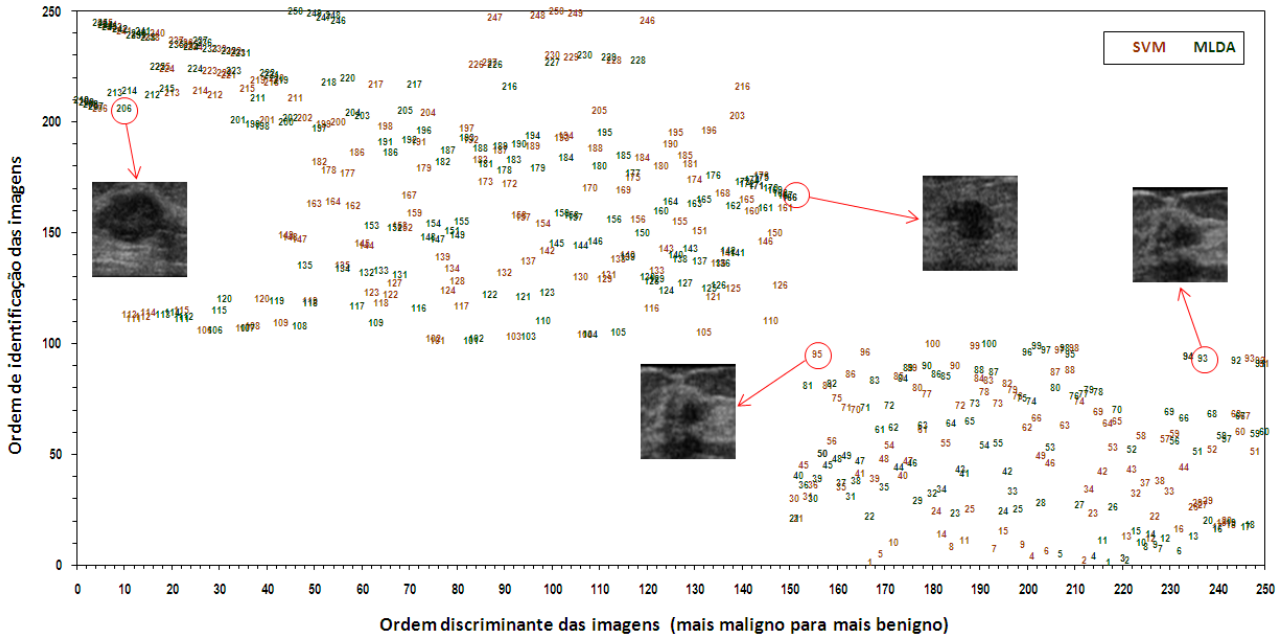
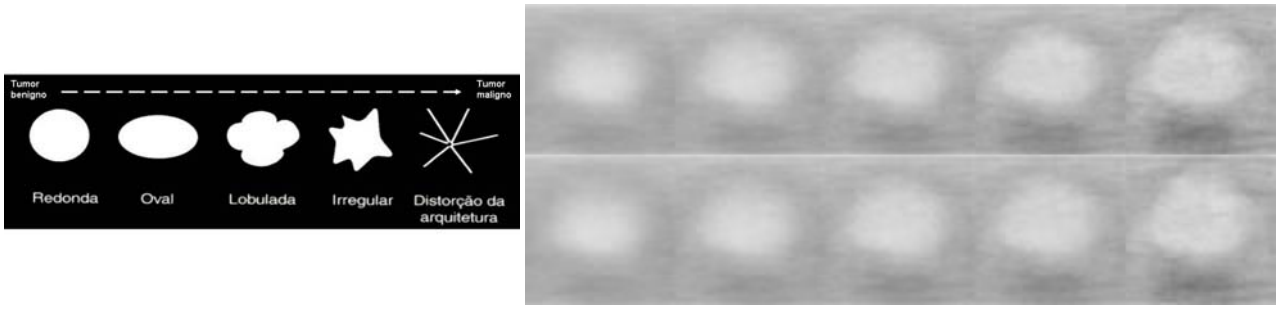



**Figura 3. Hiperplanos MLDA e SVM de separação: maligno a esquerda e benigno a direita.**

diferença no formato dos nódulos é bastante expressiva. Em âmbito geral, trata-se de um bom guia para o estudo dos tumores mamográficos em imagens de ultrassonográficas.

Para avaliar essa transição, apresentada pelo BI-RADS, foi gerada a navegação nos hiperplanos MLDA e SVM entre os grupos orientada a protótipos. Tal navegação foi construída com as imagens médias calculadas a partir de intervalos estabelecidos na ordenação das imagens, obtida pelos métodos MLDA e SVM. O intervalo adotado foi de cinquenta imagens, ou seja, a cada cinquenta imagens ordenadas calcula-se a imagem média. Como a ordenação é composta pelas 250 imagens amostrais, obtém-se então cinco imagens para navegação dos protótipos ordenados do mais benigno para o mais maligno.

A Figura 4(b), em modo negativo, ilustra a transição das imagens entre os grupos de tumores benignos e malignos para os classificadores MLDA e SVM. É possível notar, apesar do número limitado de imagens considerado, diferenças no formato dos tumores e também na região inferior destes. A região mais clara na parte inferior das imagens é mais intensa nas imagens que pertencem ao grupo dos tumores malignos (lado direito). O tamanho dos tumores também apresenta variação na transição entre os grupos. O tamanho dos tumores benignos (lado esquerdo) é menor e segue aumentando a medida que a navegação se aproxima dos tumores malignos (lado direito).




(a) (b)

**Figura 4. (a) Formato BI-RADS dos tumores mamários; (b) Navegação nos hiperplanos MLDA (linha superior) e SVM (linha inferior) por meio de protótipos (dos mais benignos para os mais malignos).**

## 5. Conclusão

O sucesso da aplicação de métodos estatísticos lineares é fortemente dependente das fases de pré-processamento e segmentação das imagens. Portanto, muitas vezes a aplicação de técnicas de melhoramento de imagens, na fase de pré-processamento, pode favorecer um determinado tipo de análise mas prejudicar outro. Os resultados obtidos na análise multivariada considerando a intensidade de todos os pixels da imagem simultaneamente, por meio dos classificadores MLDA e SVM, alcançaram desempenho superior perante a mesma análise utilizando características extraídas das imagens como circularidade e sombra acústica. Já a análise univariada indicou as regiões central e inferior da imagem como mais diferentes estatisticamente. Na navegação entre os grupos de tumores benignos e malignos, foi possível notar as diferenças entre os grupos justamente nessas regiões indicadas pela análise univariada. Muito provavelmente as características de circularidade e sombra acústica exerceram influência na discriminação dos grupos. Entretanto, a análise final dos resultados indicou que as diferenças entre as imagens de tumores benignos e malignos não são concentradas em determinadas regiões da imagem. Trata-se de uma ferramenta útil para o diagnóstico médico e identificação de casos mais simples e difíceis de classificação.

## Referências

[1] Instituto Nacional de Câncer. Estimativa 2010: incidência de câncer no Brasil. 2009.

[2] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(7):179–188, 1936.

[3] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 1990.

[4] G. A. Giraldi, P. S. Rodrigues, E. C. Kitani, J. R. Sato, and C. E. Thomaz. Statistical Learning Approaches for Discriminant Features Selection. *Journal of the Brazilian Computer Society*, 14(2):7–22, 2008.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, 2009.

[6] A. K. Jain and B. Chandrasekaran. Dimensionality and sample size considerations in pattern recognition practice. In P. R. Krishnaiah and L. N. Kanal, editors, *Handbook of Statistics*, volume 2, pages 835–855, The Netherlands, North-Holland, 1982. Eds. Amsterdam.

[7] D. B. Kopans. *Imagem da Mama*. Medsi, second edition, 2000.

[8] A. C. Lorena and A. C. P. L. F. Carvalho. Uma Introdução às Support Vector Machines. Technical report, 2007.

[9] P. S. Rodrigues, R.-F. Chang, and J. S. Suri. Non-Extensive Entropy for CAD Systems of Breast Cancer Images. *Computer Graphics and Image Processing, Brazilian Symposium on*, 0(3):121–128, 2006.

[10] P. S. Rodrigues, G. A. Giraldi, R.-F. Chang, and J. S. Suri. *Automatic Classification of Breast Lesions in 3-D Ultrasound Images*. Biomedical Imaging and Bioinformatics, October 2006.

[11] J. R. Sato, A. Fujita, C. E. Thomaz, M. da Graça Morais Martin, J. Mourão-Miranda, M. J. Brammer, and E. A. Junior. Evaluating SVM and MLDA in the extraction of discriminant regions for mental state prediction. *NeuroImage*, 46(1):105 – 114, 2009.

[12] C. E. Thomaz, J. P. Boardman, S. Counsell, D. L. G. Hill, J. V. Hajnal, A. D. Edwards, M. A. Rutherford, D. F. Gillies, and D. Rueckert. A Multivariate Statistical Analysis of the Developing Human Brain in Preterm Infants. *Image and Vision Computing*, 25(6):981–994, 2007.

[13] C. E. Thomaz, E. C. Kitani, and D. F. Gillies. A Maximum Uncertainty LDA-based approach for Limited Sample Size problems - with application to Face Recognition. *Journal of the Brazilian Computer Society*, 12:7–18, September 2006.

[14] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.