

**INSTITUTO DE CIÊNCIAS EXATAS  
CURSO DE PÓS-GRADUAÇÃO EM MODELAGEM  
MATEMÁTICA E COMPUTACIONAL**

**DISSERTAÇÃO**

**Inteligência Computacional Aplicada na Classificação de  
Microcalcificações Mamárias**

**Rúpila Rami da Silva Gôda**

**2016**



**UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO  
INSTITUTO DE CIÊNCIAS EXATAS  
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM  
MATEMÁTICA E COMPUTACIONAL**

**INTELIGÊNCIA COMPUTACIONAL APLICADA NA CLASSIFICAÇÃO  
DE MICROCALCIFICAÇÕES MAMÁRIAS**

**RÚPILA RAMI DA SILVA GÔDA**

*Sob a Orientação do Professor*  
**Robson Mariano da Silva**

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre em Ciências**, no Curso de Pós-Graduação em Modelagem Matemática e Computacional, Área de Concentração em Modelagem Matemática e Computacional.

Seropédica, RJ  
Julho de 2016

006.3

G577i

T

Gôda, Rúpila Rami da Silva, 1976-

Inteligência computacional aplicada na  
classificação de microcalcificações  
mamárias / Rúpila Rami da Silva Gôda -  
2016.

88 f.: il.

Orientador: Robson Mariano da Silva.

Dissertação (mestrado) - Universidade  
Federal Rural do Rio de Janeiro, Curso de  
Pós-Graduação em Modelagem Matemática e  
Computacional.

Bibliografia: f. 61-66.

1. Inteligência artificial - Teses.
2. Inteligência computacional - Teses.
3. Glândulas mamárias - Doenças - Teses. I. Silva, Robson Mariano da, 1963-. II. Universidade Federal Rural do Rio de Janeiro. Curso de Pós-Graduação em Modelagem Matemática e Computacional. III. Título.

**UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO**  
**INSTITUTO DE CIÊNCIAS EXATAS**  
**PROGRAMA DE PÓS-GRADUAÇÃO MODELAGEM MATEMÁTICA E**  
**COMPUTACIONAL**

**RÚPILA RAMI DA SILVA GÔDA**

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre em Ciências**, no Programa de Pós-Graduação em Modelagem Matemática e Computacional área de Concentração em Inteligência Computacional e Otimização.

DISSERTAÇÃO APROVADA EM 30/08/2016



**Robson Mariano da Silva Dr. UFFRJ (Orientador)**



**Angel Ramon Sanchez Delgado Dr. UFFRJ**



**Raquel Lima Oliveira Dra. CBNB**

## **Dedicatória**

Dedico este trabalho primeiramente a Deus por me ajudar a superar todas as dificuldades. Aos meus pais pelo incentivo e ajuda nas horas em que eu mais precisava. A todos meus amigos que me apoiaram. E ao amor da minha vida, meu filho Heitor.

## **Agradecimentos**

A Deus por ter proporcionado mais uma grande etapa na minha vida.

Aos meus pais Alamyrr Tavares Gôda e Eliana da Silva Gôda, por ser o motivo das minhas vitórias.

Aos meus irmãos Ramisés da Silva Gôda e Ramiele da Silva Gôda, por sempre me apoiar.

Ao meu Professor e orientador que me proporcionou a concretização de mais um sonho, por sempre me auxiliar dando todo o suporte diante de todas as adversidades.

A UFRRJ e ao seu Programa de Pós-Graduação em Modelagem Matemática e Computacional por ter possibilitado realizar este mestrado.

Aos meus amigos do Curso de Mestrado que sempre me apoiaram ajudando de todas as formas, que aqui gostaria de mencionar: Maurício de Carvalho, Keitilane Sany da Silveira, Marcelle de Oliveira Rosa Ribeiro, Felipe Augusto da Silva, Pablo Amauri Carvalho de Araújo, Bruno Areal de Santana, Ulisses Roque Tomas, Marlon Ferreira Corsi, Carolina Gurski Olanda, Werley de Oliveira Gonçalves, Eluã Ramos Coutinho e Kelly Harumi Fausta Kawase.

## RESUMO

GÔDA, Rúpila Rami da Silva. **Inteligência Computacional Aplicada em Microcalcificações Mamárias**. 2016. 68p. Dissertação (Mestrado em Modelagem Matemática e Computacional). Instituto de Ciências Exatas, Departamento de Matemática, Universidade Federal Rural do Rio de Janeiro, Seropédica, 2016.

O câncer de mama é a segunda neoplasia mais frequente no mundo. Segundo dados do Instituto Nacional de Câncer (INCA), no ano de 2014 foram diagnosticados 52.680 novos casos no Brasil, número este que corresponde a um aumento de 22% em relação ao ano de 2013. Sendo responsável por aproximadamente 39% dos óbitos das mulheres portadoras de câncer. Apesar da elevada taxa de incidência, a mortalidade causada por esta neoplasia tem diminuído desde o final dos anos oitenta, graças ao avanço das pesquisas em métodos para o diagnóstico precoce. No entanto, diagnosticar corretamente o câncer é um processo complexo e muito difícil em consequência das diversas variáveis envolvidas. Para um diagnóstico preciso, exige-se muita experiência e, principalmente, que a classificação do estadiamento clínico do tumor (estágio do câncer) esteja correta. Os tradicionais sistemas de classificação de patologias utilizados são complexos e em muitas vezes oferecem limitações. Como é o caso da técnica de mamografia, que amplamente utilizada, não é tão eficaz para mulheres com mamas densas, cirurgicamente alteradas ou com menos de 40 anos. Desta forma, torna-se necessário o desenvolvimento de sistemas integrados que combinados com a experiência dos profissionais da área, possibilite realizar o diagnóstico preciso na detecção do câncer de mama. O objetivo do presente trabalho é aplicar a técnica SVM (Máquina de Vetor de Suporte), de sorte a auxiliar na interpretação diagnóstica das microcalcificações detectadas em mamografia de rastreamento. O conjunto de dados utilizado consistiu de 961 amostras de exames mamográficos, obtidos junto ao Instituto de Radiologia da Universidade de Erlangen-Nuremberg. Neste conjunto possuímos informações referentes a idade da paciente, classificação BI-RADS (Breast Imaging Reporting and Data System), forma, massa, densidade e severidade (benigno|maligno) das microcalcificação. A SVM desenvolvida foi implementada utilizando-se o software R (R Development Core Team; [http:// www.R-project.org/](http://www.R-project.org/) ). Os dados foram divididos em dois grupos: o conjunto de treinamento composto por 80% das amostras de exames mamográficos, usado para estimar os parâmetros do modelo e o conjunto de teste independente, com 20% das amostras restantes, utilizado para mensurar a performance da SVM. Para avaliar o desempenho do modelo computacional proposto foram utilizados o valor da Precisão Total ou Acurácia (ACC), Sensibilidade (S) e Especificidade(E). Os resultados apresentados pela SVM na identificação das lesões malignas em pacientes portadoras de microcalcificações se mantiveram entre 72,7% e 100% o que demonstram que os mesmos alcançaram um grau satisfatório em relação com outras literaturas aplicadas.

**Palavras-chave:** SVM, S, ACC, E.

## ABSTRACT

Gôda, Rúpila Rami da Silva. Applied Computational Intelligence in microcalcifications Mammary. 2016. 68p. Dissertation (Masters in Mathematics and Computational Modelling). Institute of Exact Sciences, Department of Mathematics, Federal Rural University of Rio de Janeiro, Seropédica, 2016.

Breast cancer is the second most common cancer worldwide. According to the National Cancer Institute (INCA) in 2014 were diagnosed 52,680 new cases in Brazil, a number that corresponds to a 22% increase over the year 2013. Being responsible for approximately 39% of women's deaths cancer patients. Despite the high incidence rate, mortality from this cancer has declined since the late eighties, thanks to advances in research on methods for early diagnosis. However, correctly diagnosing cancer is a complex and difficult process as a result of the different variables involved. For an accurate diagnosis, a lot of experience and especially it is required, that the classification of clinical staging of tumor (cancer stage) is correct. The conditions used traditional classification systems are complex and often offer limitations. As is the case of mammography technique, widely used, it is not as effective for women with dense breasts, surgically altered, or under 40 years. Thus, it becomes necessary to develop integrated systems that combined with the professionals in the field experience, allows performing accurate diagnosis in detecting breast cancer. The objective of this study is to apply the technique SVM (Support Vector Machine), so as to assist in the diagnostic interpretation of microcalcifications detected on screening mammography. The data set used consisted of 961 samples of mammograms, obtained from the Radiology Institute of the University of Erlangen Nuremberg. In this set we have information on the age of the patient, BI-RADS (Breast Imaging Reporting and Data System), shape, mass, density and severity (benign | malignant) of microcalcifications. The SVM was developed implemented using the R software (R Development Core Team; <http://www.R-project.org/>). The data were divided into two groups: the training set consisting of 80% of the samples of mammographic, used to estimate the model parameters and the independent test set, with 20% of the remaining samples, used to measure the performance of SVM . To evaluate the performance of proposed computational model we used the value of the Total Precision or Accuracy (ACC), sensitivity (S) and specificity (E). The results presented by SVM in identifying malignant lesions in patients with calcifications remained between 72.7% and 100%, which shows that they achieved a satisfactory level in relation to other literatures applied.

*Keywords: SVM, S, ACC, E.*



## LISTA DE FIGURAS

Figura 1: O Câncer de Mama. Fonte: (MINISTÉRIO DA SAÚDE, 2002). .....	6
Figura 2: Distribuição proporcional dos dez tipos de câncer mais incidentes em mulheres brasileiras, estimados para 2013, exceto pele não melanoma. Fonte: INCA, 2011. ....	7
Figura 3: Representação espacial das taxas brutas de incidência de câncer de mama por 100 mil mulheres, estimadas para o ano de 2011. Fonte: INCA, 2011. ....	8
Figura 4: (a) Mamografia com nódulo benigno selecionado. (b) Mamografia com nódulo maligno selecionado. Fonte: (HEATH et al. 1998). ....	9
Figura 5: Mamógrafos (a) esquema. Fonte: (ACS 2007). (b) Mamógrafo Real. Fonte: (RSNA2007). ....	10
Figura 6: A forma da massa pode ser dividida em cinco formas. Fonte: Kopans (2000). ....	11
Figura 7: A Margem da massa pode ser dividida em cinco contornos. Fonte: Kopans (2000). ....	12
Figura 8: Punção Aspirativa por agulha fina (PAAF) Instituto de Mastologia, 2016. ....	16
Figura 9: Indução de classificador em aprendizado supervisionado. ....	18
Figura 10: Separação de duas classes através de hiperplanos. ....	19
Figura 11: Vetores de Suporte (destacados por círculos). ....	21
Figura 12: Hiperplano de separação (w,b) para um conjunto de treinamento bidimensional. ....	22
Figura 13: Hiperplano ótimo com máxima margem $\rho_0$ de separação dos padrões linearmente separáveis. ....	23
Figura 14: Interpretação gráfica da distância $x$ até o hiperplano ótimo. ....	24
Figura 15: (a) Conjunto de dados não linear; (b) Fronteira não linear no espaço de entradas; (c) Fronteira linear no espaço de características. ....	27
Figura 16: (a) o ponto $(x_i, d_i)$ se encontra na região de separação, mas do lado correto. (b) O ponto $(x_i, d_i)$ se encontra na região de separação, mas do lado incorreto. (c) O ponto $(x_i, d_i)$ se encontra fora da região de separação, mas do lado incorreto. ....	29
Figura 17: Transformação: problema não linearmente separável em um problema linearmente separável (Rebelo 2008). ....	31
Figura 18: Característica do Banco de Dados. ....	34

## LISTA DE TABELAS

Tabela 1: Distribuição proporcional dos dez tipos de câncer mais incidentes em mulheres brasileiras, estimados para 2013, exceto pele não melanoma. Fonte: INCA, 2011. ....	7
Tabela 2: Funções de kernel mais usadas. ....	32
Tabela 3: Classificação das variáveis do conjunto de dados ....	33
Tabela 4: Distribuição no conjunto treino. ....	36
Tabela 5: Distribuição no conjunto teste. ....	36
Tabela 6: Tabelas de parâmetros. ....	37
Tabela 7: Resultado das 50 simulações utilizando SMV com as variáveis: Birads, Margem, Forma e Densidade (BMFD). ....	39
Tabela 8: Resultado das 50 simulações utilizando SMV com as variáveis: Birads, Margem e Forma. (BFM). ....	41
Tabela 9: Resultado das 50 simulações utilizando SMV com as variáveis: Birads, Forma e Densidade (BFD). ....	43
Tabela 10: Resultado das 50 simulações utilizando SMV com as variáveis: Birads, Margem e Densidade (BMD). ....	45
Tabela 11: Resultado das 50 simulações utilizando SMV com as variáveis: Margem, Forma e Densidade (MFD). ....	47
Tabela 12: Resultado das médias de 50 simulações utilizando o modelo com as variáveis: Birads, Margem, Forma e Densidade (BMFD). ....	49
Tabela 13: (BFD) ....	49
Tabela 14: (FDM) ....	49
Tabela 15: BDM ....	50
Tabela 16: (BFM) ....	50
Tabela 17: Banco de dados coletados no Instituto de Radiologia da Universidade Erlangen-Nuremberg, entre 2003 e 2006. ....	57

# SUMÁRIO

<b>CAPÍTULO I</b> .....	<b>1</b>
<b>1. INTRODUÇÃO</b> .....	<b>1</b>
1.1. TRABALHOS RELACIONADOS .....	2
1.2. OBJETIVOS .....	3
1.2.1. Objetivo geral .....	3
1.2.2. Objetivos específicos .....	3
1.3. ESTRUTURA DO TRABALHO .....	4
<b>CAPÍTULO II</b> .....	<b>5</b>
<b>2. REVISÃO DE LITERATURA</b> .....	<b>5</b>
2.1. CÂNCER .....	5
2.1.1. Câncer de Mama .....	6
2.1.2. Mamografia .....	9
2.1.3. Descrição dos Dados Mamográficos .....	11
2.1.4. Biópsia .....	15
2.2. APRENDIZADO DE MÁQUINA .....	17
2.2.1. Máquinas de Vetores de Suporte .....	19
2.2.2. Máquinas de Vetores de Suporte Linearmente Separável .....	22
<b>2.2.2.1. Hiperplano Ótimo para Classes Linearmente Separáveis</b> .....	<b>25</b>
2.2.2. SVM Não Lineares .....	27
2.2.2.1. Hiperplano Ótimo para Classes não Linearmente Separáveis .....	28
2.2.2.2. Funções Kernel .....	31
<b>CAPÍTULO III</b> .....	<b>33</b>
<b>3. MATERIAIS E MÉTODOS</b> .....	<b>33</b>
3.1. BASE DE DADOS .....	33
3.3. MODELAGEM COMPUTACIONAL .....	35
3.3.1. Modelo .....	35
<b>CAPÍTULO IV</b> .....	<b>39</b>
<b>4. RESULTADOS</b> .....	<b>39</b>
<b>4.1. RESULTADOS</b> .....	<b>39</b>
4.1 MODELO COMPLETO (BMFD) .....	39
4.2 MODELO BFM. RETIRADA DA VARIÁVEL DENSIDADE .....	41
4.3 MODELO BFD. RETIRADA DA VARIÁVEL MARGEM .....	43
4.4 MODELO BMD. RETIRADA DA VARIÁVEL FORMA .....	45
4.5 MODELO FDM. RETIRADA DA VARIÁVEL BIRADS .....	47
<b>CAPÍTULO V</b> .....	<b>51</b>
<b>5.1. CONCLUSÕES</b> .....	<b>51</b>
<b>5.2. TRABALHOS FUTUROS</b> .....	<b>51</b>
<b>6. REFERÊNCIAS BIBLIOGRÁFICAS</b> .....	<b>52</b>

# CAPÍTULO I

## 1. INTRODUÇÃO

O câncer de mama é tipo de câncer que aparece com maior frequência entre as mulheres brasileiras (INCA 2013), nas quais a maior incidência ocorre em mulheres de 35 a 50 anos, a cada 8 mulheres uma é diagnosticada com câncer de mama.

Desde o início das pesquisas sobre o câncer de mama, a melhor maneira para cura da doença aproximadamente 95% é detecção rápida. A detecção pode ser conseguida através do exame da mamografia, cujo método é o mais eficaz para o rastreamento do câncer de mama disponível hoje. A mamografia é uma forma particular de radiografia capaz de registrar imagens da mama com a finalidade de diagnosticar a presença ou ausência de estruturas que possam indicar a doença. Com esse tipo de exame pode-se detectar o tumor antes que ele se torne palpável. No entanto, a avaliação do exame de mamografia e o diagnóstico, que é realizado por um radiologista, requer bastante habilidade, porém há limitações na predição primária do câncer de mama.

Estudos tem salientado que apenas 20% a 30% de nódulos mamográficos identificados pelo radiologista, considerados suspeitos e submetidos a exame de biópsia, são malignos, ou seja, o baixo valor preditivo positivo de câncer de mama na biópsia resultante da interpretação mamografia leva a cerca 70% de biópsias desnecessárias com resultados benignos (HERMANN et al. 1987; 1988; JACOBSON; EDEIKEN, 1990; ARAUJO, 2000).

O exame de biópsia resume-se em uma pequena cirurgia destinada a retirar a um pedaço no nódulo suspeito, para que este seja analisado. Um erro no diagnóstico (falso positivo que é submeter a biópsia em um nódulo suspeito de câncer, o qual não se constata ser câncer) o que acarreta um custo excessivo ao sistema de saúde e traumatismo ao paciente.

Por esses objetivos, vários pesquisadores estão desenvolvendo sistemas que auxiliam no diagnóstico do câncer evitando o exame desnecessário de biópsia. Estes sistemas são conhecidos como CAD (Diagnóstico Auxiliado por Computador em inglês, Computer-Aided-Diagnosis).

Sistemas CAD auxiliam no diagnóstico, atuando em uma região suspeita de câncer, pré-identificada no exame de mamografia. Assim ao identificar uma região suspeita é feita uma marcação (manualmente) no nódulo não palpável suspeito de câncer. A região que envolve a marcação é chamada de região de interesse, ROI (em inglês, Regions of Interest).

São utilizadas técnicas de processamento de imagens para indicar e remover determinadas características do nódulo (as ROIs). A partir dessas características extraídas, são aplicadas técnicas de inteligência computacional, que as classificam em benignas ou malignas.

Entre as inúmeras técnicas utilizadas na extração de características dos nódulos nas ROIs, algumas se destacam como os filtros de Wavelets (FERRARI et al., 2001), crescimento da região (SHEN et al., 1994), transformada H-dome (HALKIOTIS et al., 2002) e transformada de Watershed (SANTOS, 2002).

As Redes Neurais Artificiais (RNA) são um exemplo de técnica de inteligência artificial computacional que vem sendo utilizada para a classificação das características extraídas (HALKIOTS et al., 2002; YU; GUAN, 2000; ANDRE; RANGAYYAN, 2003). Para isso, diversas arquiteturas de RNA como Rede Perceptron de Múltiplas Camadas (MLP em inglês Multi-Layer Perceptron) são aplicadas numa melhor tentativa de classificação de nódulos.

## 1.1. Trabalhos Relacionados

A eficácia dos sistemas ou metodologias CAD/CADx é altamente dependente do método de extração de características (ZHANG e KUMAR, 2006), da seleção das características mais adequadas para uma maior discriminação e do classificador utilizado. A literatura disponível traz trabalhos reconhecidos que tratam do mesmo problema abordado pelo método apresentado, qual seja, desenvolver métodos computacionais que possam auxiliar o especialista na tarefa de análise das imagens médicas.

Em (BRAZ JÚNIOR et al ., 2007), é evidenciada uma metodologia para discriminação e classificação de regiões extraídas de mamografias em massas e não-massas através de estatísticas espaciais, como o Índice de Moran e o Coeficiente de Geary . O trabalho utiliza uma Máquina de Vetores de Suporte (MVS) para classificação das regiões, obtendo 99,64% de acurácia.

Outra estrutura que também utiliza MVS para a classificação de regiões da mama, com 89,30% de acurácia, é proposta em (MARTINS, 2007). Nela o algoritmo Growing Neural Gas é utilizado para segmentação dos candidatos a massa e a Função K de Ripley é utilizada para descrever a textura dos mesmos.

Em (COSTA et al. , 2007) é comparada a eficiência dos classificadores Máquina de Vetores de Suporte (MVS) e Análise Discriminante Linear (ADL). Os resultados da classificação alcançam 89,2% e 99,6% para ADL e MVS, respectivamente, no objetivo de classificação de regiões em massa e não-massa, comprovando a maior capacidade de generalização que o MVS é capaz de realizar sobre as amostras tratadas.

Empregando uma nova abordagem para classificação de massas em mamografias é proposta em (MOAYEDI et al ., 2007), onde são empregados vetores de suporte baseados em uma rede neural fuzzy (SVFNN – Support Vector Based Fuzzy Neural Network ) para desempenhar as tarefas do classificador. As características são extraídas a partir da representação das massas no domínio da frequência aplicando os coeficientes obtidos a partir de contourlet.

No (Timp et al . 2007) analisam o desempenho da metodologia de classificação de massas quando as características extraídas são realizadas em mamografias consecutivamente obtidas no tempo. O principal objetivo é melhorar a descrição de massas utilizando informações presentes em mais de uma mamografia, obtidas sucessivamente.

Em (OSTA et al. , 2008) é pesquisada a utilização de wavelets para a extração de características dos tecidos da mama. O estudo compara o desempenho da classificação das regiões de interesse em massas e não massas através de uma MVS e de uma rede neural RBF ( Radial-Basis-Function Neural Network - RBFNN ).

Em (CAMPOS et al., 2007), é proposta uma metodologia para discriminação e classificação de tecidos da mama nas classes benigno, maligno e normal, usando Independent Component Analysis (ICA) e redes neurais. O trabalho utiliza a base de dados MIAS, obtendo uma acurácia de 97%.

Os trabalhos relacionados acima indicam serem promissoras as pesquisas de novas técnicas para extração de características de imagens radiológicas, especialmente no que diz respeito à utilização de medidas geoestatísticas como descritores de textura, assim como a utilização do método de Máquina de Vetores de Suporte como classificador por apresentar resultados superiores durante a etapa de generalização de resultados.

## **1.2. Objetivos**

### **1.2.1. Objetivo geral**

Os processamentos de imagens juntamente com técnicas de diferentes áreas como estatística e inteligência artificial têm proposto a aplicação de uma vasta lista de técnicas. O objetivo principal, para estudos na área de câncer de mama, é criar uma metodologia capaz de prover um mecanismo que auxilie o especialista a acompanhar, delimitar e diagnosticar, através de imagens radiográficas, qualquer aparecimento de tecido que represente uma neoplasia numa mama. Elaborando um modelo computacional estruturado em SVM não linear, que possa auxiliar o médico na sua decisão de realizar uma biópsia de mama em uma suspeita lesão, no qual a prioridade está na diminuição da quantidade de exames invasivos. Exames que seriam realizados em situações de desconfiança pelo especialista, poderiam ser descartados pela aplicação que estaria baseando sua recomendação em muitas outras bases de informação. Evidente que a aplicação não tem o intuito de substituir o especialista, mas sim prover uma segunda opinião de maneira clara e rápida.

### **1.2.2. Objetivos específicos**

- Verificar a diferença estatística entre os resultados obtidos pelo modelo proposto.
- Encontrar relações entre os atributos do banco de dados.
- Realizar um estudo das variáveis envolvidas no processo
- Utilizar ferramentas computacionais na área de inteligência computacional, precisamente SVM;
- Verificar qual das variáveis teve o maior impacto no desempenho da rede.
- Avaliar método estruturado em Inteligência Computacional SVM não Linear comparando o impacto da ausência de dados na classificação de resultados de mamografia.

### 1.3. Estrutura do Trabalho

Este trabalho está organizado em capítulos da seguinte forma:

- Capítulo I: Traz uma breve introdução sobre Câncer de Mama e SVM, bem como a motivação de realizar este estudo juntamente com os objetivos e alguns trabalhos relacionados nesta área proposta.
- Capítulo II: Mostra a fundamentação teórica de Câncer, Câncer de Mama, Mamografia, Biópsia, Máquina de Vetor de Suporte, SVM Linearmente Separável, Hiperplano Ótimo para Classes Linearmente Separáveis, SVM Não Lineares, Hiperplano Ótimo para Classes Não Linearmente Separáveis, Funções Kernel, com o objetivo de um embasamento teórico bem esclarecido para um melhor entendimento da proposta deste estudo.
- Capítulo III: Apresenta Materiais e Métodos, o conjunto de dados utilizados, a forma de como os dados estão distribuídos, a Modelagem Computacional do método proposto.
- Capítulo IV: Este capítulo descreve o detalhamento dos resultados obtidos.
- Capítulo V: Considerações gerais e possibilidades de trabalhos futuros.

## CAPÍTULO II

### 2. REVISÃO DE LITERATURA

#### 2.1. Câncer

O câncer, termo derivado do latim *câncer* e do grego *karkínos* (caranguejo), inclui mais de 200 tipos conhecidos de doenças, cuja característica mais comum é a ocorrência de alterações nos processos de divisão das células do corpo. Tais alterações proporcionam um crescimento anormal e geralmente mais rápido de um conjunto de células dando origem ao que se denomina genericamente de tumor.

Os tumores podem originar-se em todo e qualquer tecido humano e em qualquer faixa etária, incluindo a capacidade de infiltrar-se através de barreiras de tecidos normais e invadir tecidos circunvizinhos por extensão direta ou por disseminação a regiões distantes, por meio de vias linfáticas ou sanguíneas.

É importante observar que o conceito de câncer, ainda sem uma definição unânime ou reconhecida mundialmente, não implica necessariamente a idéia de um tumor no sentido de uma massa sólida, sendo possível incluí-lo como uma das modalidades do grupo das patologias hemopáticas, como é o caso dos vários tipos de linfomas e leucemias que acometem crianças, adolescentes e adultos.

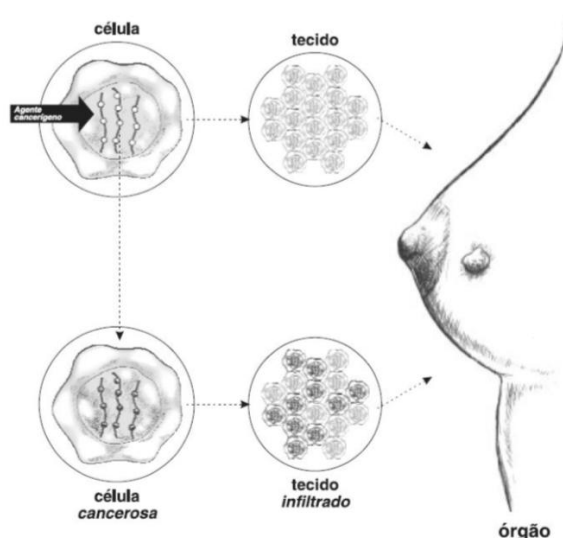
Câncer é o nome dado a um conjunto de mais de 100 doenças que têm em comum o crescimento desordenado (maligno) de células que invadem os tecidos e órgãos, podendo espalhar-se (metástase) para outras regiões do corpo. Dividindo-se rapidamente, estas células tendem a ser muito agressivas e incontroláveis, determinando a formação de tumores (acúmulo de células cancerosas) ou neoplasias malignas. Estas mutações levam a alteração na expressão ou função de genes-chave para a manutenção da homeostasia celular. Essas alterações genéticas podem converter uma célula normal em uma célula transformada, que se caracteriza por não mais responder aos sinais de controle de proliferação, morte e diferenciação que governam a comunidade celular (BELTRÃO-BRAGA, P.C.B.; VERÔNICA, R.T.; CHAMAS, R, 2004). Por outro lado, um tumor benigno significa simplesmente uma massa localizada de células que se multiplicam vagarosamente e se assemelham ao seu tecido original, raramente constituindo um risco de vida. (INSTITUTO NACIONAL DO CÂNCER, 2015).

Apesar de toda evolução tecnológica da área médica nas últimas décadas, o conhecimento acumulado sobre os fatores causadores do câncer ainda não foi suficientemente compreendido, sendo possível incluir causas tão diversas quanto a própria evolução da doença. Segundo dados do Instituto Nacional do Câncer (INCA), hábitos alimentares inadequados (35%), tabagismo (30%) e um conjunto complexo de fatores relacionados à exposição a radiações ionizantes, submissão a contextos ambientais estressantes, comportamentos de risco e fatores genéticos, étnicos e ocupacionais (35%) estão entre as principais variáveis geradoras de processos cancerígenos.



### 2.1.1. Câncer de Mama

Câncer é caracterizado por modificações que determinam um crescimento celular desenfreado. As células afetadas multiplicam-se de maneira descontrolada, invadindo tecidos vizinhos e formando tumores, como ilustrado na figura 1.



**Figura 1:** O Câncer de Mama. Fonte: (MINISTÉRIO DA SAÚDE, 2002).

O câncer decorre da reprodução constante das células dos diversos órgãos do corpo humano. Em determinadas ocasiões e por razões desconhecidas, certas células reproduzem-se ou degeneram-se com uma velocidade maior, desencadeando o aparecimento de massas celulares denominadas neoplasias (tumor), tecido de formação nova, mas de origem patológica. As neoplasias são classificadas como maligna ou benigna.

As células invadidas acabam perdendo a sua função especializada e, à medida que substituem as células normais, comprometem a função do órgão afetado. Podem adquirir a capacidade de se desprender do tumor e migrar, através da corrente sanguínea ou do sistema linfático, chegando a órgãos distantes, constituindo as metástases (MINISTÉRIO DA SAÚDE, 2002).


O processo de formação de câncer é em geral lento, podendo levar vários anos para que uma célula prolifere e dê origem a um tumor palpável. Esse processo é composto de vários estágios, quais sejam: estágio de iniciação, onde os genes sofrem ação de fatores cancerígenos; estágio de promoção, onde os agentes oncopromotores atuam na célula já alterada; e estágio de progressão, caracterizada pela multiplicação descontrolada e irreversível da célula (MINISTÉRIO DA SAÚDE, 2002).

Em 2008, a Agência Internacional para Pesquisa em Câncer (IARC)/OMS estimou que ocorreriam 12,4 milhões de casos novos e 7,6 milhões de óbitos por câncer no mundo. Destes, os mais incidentes foram o câncer de pulmão (1,52 milhões de casos novos), mama (1,29 milhões) e cólon e reto (1,15 milhões). Devido ao mau prognóstico, o câncer de pulmão foi a

principal causa de morte (1,31 milhões), seguido pelo câncer de estômago (780 mil óbitos) e pelo câncer de fígado (699 mil óbitos). Para América do Sul, Central e Caribe, estimou-se em 2008 cerca de um milhão de casos novos de câncer e 589 mil óbitos. Nas mulheres, o mais frequente foi o câncer de mama, seguido do colo do útero, cólon e reto, estômago e pulmão (INTERNATIONAL AGENCY FOR RESEARCH ON CANCER, 2008).

Atualmente, o câncer de mama é considerado como o segundo tipo de câncer mais comum no mundo, sendo o mais frequente entre as mulheres, como mostra a Tabela 1. Ao contrário do câncer de colo de útero, esta doença encontra-se relacionada ao processo de industrialização, com risco de adoecimento associado a elevado status sócio- econômico, além de outros fatores de risco clássicos descritos, tais como baixa paridade, idade precoce da menarca e tardia da menopausa, obesidade, altura e consumo de álcool (PARKIN DM, 2000).

**Tabela 1:** Distribuição proporcional dos dez tipos de câncer mais incidentes em mulheres brasileiras, estimados para 2013, exceto pele não melanoma. Fonte: INCA, 2011.

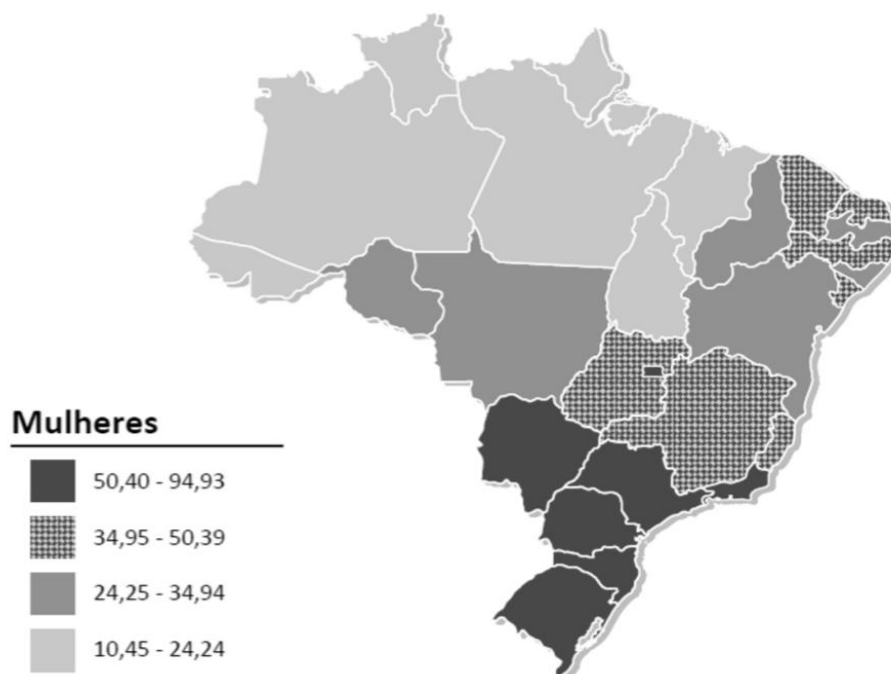
	<b>Localização primária</b>	<b>casos novos</b>	<b>percentual</b>
<b>Mulheres</b> 	Mama Feminina	52.680	27,9%
	Colo do Útero	17.540	9,3%
	Cólon e Reto	15.960	8,4%
	Glândula Tireoide	10.590	5,6%
	Traqueia, Brônquio e Pulmão	10.110	5,3%
	Estômago	7.420	3,9%
	Ovário	6.190	3,3%
	Corpo do Útero	4.520	2,4%
	Linfoma não Hodgkin	4.450	2,4%
	Sistema Nervoso Central	4.450	2,4%

No Brasil, as estimativas, para o ano de 2010, foram válidas também para o ano de 2011, e apontaram para a ocorrência de 489.270 casos novos de câncer. Sendo 236.240 casos novos para o sexo masculino e 253.030 para sexo feminino. Estima-se que o câncer de pele do tipo não melanoma (114 mil casos novos) será o mais incidente na população brasileira, seguido pelos tumores de próstata (52 mil), mama feminina (49 mil), cólon e reto (28 mil), pulmão (28 mil), estômago (21 mil) e colo do útero (18 mil), acompanhando o mesmo perfil da magnitude observada para a América Latina (INSTITUTO NACIONAL DE CÂNCER, 2010).

Na região Sudeste, o câncer de mama é o mais incidente entre as mulheres, com um risco estimado de 65 casos novos por 100 mil. Sem considerar os tumores de pele não melanoma, este tipo de câncer também é o mais frequente nas mulheres das regiões Sul (64/100.000), Centro-Oeste (38/100.000) e Nordeste (30/100.000). Na região Norte é o segundo tumor mais incidente (17/100.000).

Em 2020, o número de idosos no Brasil chegará a 25 milhões, dos quais 15 milhões serão de mulheres, numa população estimada em 219,1 milhões. Representarão 11,4 % da população e, segundo o IBGE, é irreversível o envelhecimento da população brasileira (IBGE

SÍNTESE DE INDICADORES SOCIAIS, 2006).



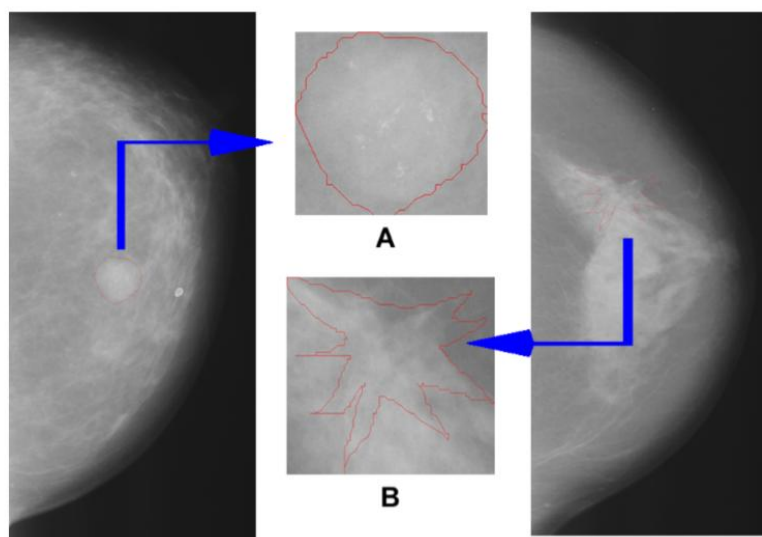
**Figura 2:** Representação espacial das taxas brutas de incidência de câncer de mama por 100 mil mulheres, estimadas para o ano de 2011. Fonte: INCA, 2011.

O câncer de mama é um tumor maligno que se desenvolve na mama como consequência de alterações genéticas em algum conjunto de células, que passam a se dividir descontroladamente. No Brasil, o Ministério da Saúde estima 52.680 novos casos em um ano, com um risco estimado de 52 casos a cada 100 mil mulheres (Figura 2). Os principais fatores que predispõe o câncer na mama são: histórico familiar, idade, menstruação precoce, menopausa tardia, obesidade, ausência de gravidez e reposição hormonal. Por isso é importante fazer os exames preventivos na idade adequada, antes do aparecimento de qualquer sintoma do câncer de mama, e ficar atento a todos os sinais, alguns como: vermelhidão na pele alterações no formato dos mamilos e das mamas, nódulos nas axilas, secreção escura saindo pelo mamilo, pele enrugada como casca de laranja e em estágios avançados a mama pode abrir uma ferida. A maioria dos tumores da mama, quando iniciais, não apresenta sintomas. Caso o tumor já esteja perceptível ao toque do dedo, é sinal de que ele tem cerca de 1 cm<sup>3</sup> - o que já é uma lesão muito grande( MINISTÉRIO DA SAÚDE (BR). POLÍTICA NACIONAL DE CÂNCER DE MAMA).

É necessário o diagnóstico, pois além da mamografia, ressonância magnética, ecografia existem outros exames de imagem que podem ser feitos para identificar uma alteração suspeita de câncer de mama, é necessário fazer uma biópsia do tecido coletado da mama. Nesse material da biópsia é que a equipe médica identifica se as células são tumorosas ou não. A prevenção do câncer de mama pode ser dividida em primária e secundária: a primeira envolve a adoção de hábitos saudáveis, e a segunda diz respeito à realização de exames de rastreamento, a fim de diagnosticar o câncer de mama em estágio precoce (HANSEL,2007).

## 2.1.2. Mamografia

O rastreamento mamográfico é o exame radiológico das mamas em mulheres assintomáticas, a intervalos regulares, com objetivo de detectar o câncer de mama precoce e, assim, interromper a progressão da doença (AGUILLAR VLN, BAUAB SP, 2003). Nos últimos anos, cada vez mais, a mamografia tem detectado casos de câncer de mama numa fase incipiente, não diagnosticados pelo exame clínico (QUEIROZ RY, HUMMELAC, 1995). Ela é considerada como o teste mais sensível para rastreamento do câncer de mama clinicamente oculto, e seu uso tem sido descrito como a estratégia mais eficiente para reduzir a mortalidade por esta doença, como podemos ver na Figura 3.



**Figura 3:** (a) Mamografia com nódulo benigno selecionado. (b) Mamografia com nódulo maligno selecionado. Fonte: (HEATH et al. 1998).

Diante das colocações precedentes, nota-se a necessidade de caracterizar adequadamente as microcalcificações detectadas pela mamografia, com uma integração entre os especialistas da área, a fim de uniformizar as ações para que a investigação do câncer da mama torne-se mais eficaz e menos agressiva.

Neste sentido, em 1993, o Colégio Americano de Radiologia (ACR) publicou a primeira edição do Breast Imaging Reporting and Data System (BI-RADS<sup>TM</sup>), que visou à uniformização da estrutura, terminologia, conclusões e recomendações de condutas nos laudos mamográficos.

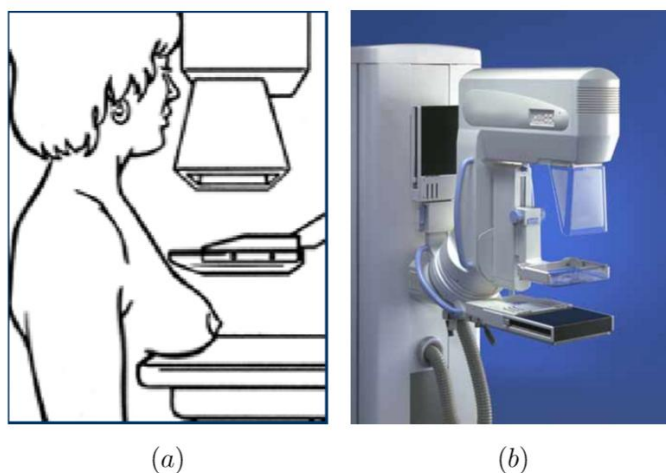
A padronização dos laudos mamográficos utilizando a classificação BI-RADS vem sendo uma importante ferramenta na orientação ao médico assistente quanto à conduta a ser tomada e como preditor de malignidade, afirmando-se como um método seguro em garantir benignidade quando nas Categorias 1 e 2, baixa chance de malignidade na Categoria 3 (<2%), sendo um indicador de estudo histopatológico nas Categorias 4 e 5, apresentando alta especificidade para malignidade na Categoria 5 (>95%) e variabilidade de malignidade quando na Categoria 4.

As calcificações são as menores estruturas observadas no exame mamográfico, sendo de difícil observação e interpretação, necessitando-se de exame com alta resolução, compressão mamária vigorosa e magnificação localizada da área a ser mais bem estudada para obter-se boa avaliação das imagens(SICKLES EA. , 1986).

A mamografia permite ao radiologista detectar as microcalcificações, estudar suas formas, distribuição, número e tamanho, Figura 4. Baseado no conhecimento das estruturas anatômicas normais e patológicas da mama pode-se ainda identificar as possíveis origens das calcificações e responder, na maioria dos casos, se apresentam características benignas, suspeitas ou reconhecer suas limitações e classificá-las como de potencial indeterminado (VIANNA AD, MARCHIORI E, 2002).

Estatísticas mostram que calcificações são encontradas em quase 50% das mulheres que foram submetidas à mamografia. A maioria das calcificações (cerca de 70-80%) é associada com alterações benignas, mas também podem corresponder a processos malignos (LIBERMAN L. et al., 1998).

O diagnóstico etiológico das calcificações mamárias vem se tornando mais importante a cada dia, pois quanto mais a tecnologia avança, aprimorando a qualidade das mamografias, mais e menores calcificações são detectadas, exigindo cada vez mais respostas precisas do radiologista, quanto ao seu significado, a fim de evitar biópsias desnecessárias nos casos benignos e permitir o diagnóstico precoce nos casos malignos (VIANNA AD, MARCHIORIE, 2002).



**Figura 4:** Mamógrafos (a) esquema. Fonte: (ACS 2007). (b) Mamógrafo Real. Fonte: (RSNA2007).

### 2.1.3. Descrição dos Dados Mamográficos

#### NÓDULOS

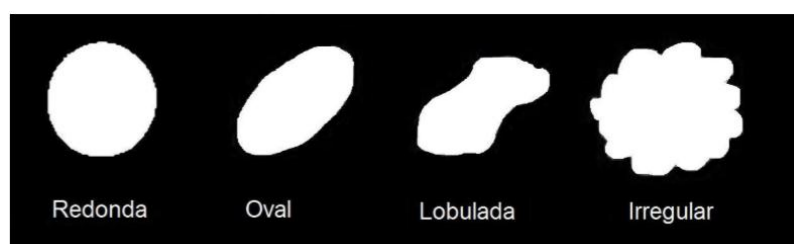
Definição:

- Um NÓDULO é tridimensional deve ser visto nas **2 incidências** mamográficas Médio Lateral Oblíqua (MLO) e Crânio Caudal (CC). Uma imagem compatível com um nódulo, mas visto em apenas uma incidência deve ser descrita como ASSIMETRIA.
- Um nódulo apresenta margens parciais ou completamente convexas.
- Os nódulos radiodensos são mais densos no centro que na periferia.

Características mamográficas:

As principais características utilizadas na descrição e classificação dos nódulos são:

• **Forma:** é o componente mais importante na avaliação das microcalcificações, sendo a heterogeneidade da morfologia e tamanho, parâmetros valiosos na determinação da causa provável do depósito das partículas. Quanto maior a variedade de formas, maior o grau de suspeição para malignidade (KOPANS , 2000). Como as calcificações malignas formam-se a partir de debris necróticos dentro dos ductos, elas tendem a ter a forma linear ou vermicular, podendo inclusive se ramificar e assumir a forma de letras (Y, X, V etc.). Esses tipos de calcificações têm um alto VPP ( Valor Preditivo Positivo) para detecção de câncer precoce, uma vez que cerca de 90% dos carcinomas se originam dos ductos, enquanto apenas 10% têm origem nos lóbulos. As calcificações com características malignas são as finas ou agrupadas ou de localização ductal, com 81% a 92% de malignidade em vários estudos. A forma de um nódulo deve ser descrita como: redonda, oval, oval lobulada, irregular e distorção de arquitetura (Figura 5).



**Figura 5:** A forma da massa pode ser dividida em cinco formas. Fonte: Kopans (2000).

• **Margem:** A margem é o limite ou a fronteira da lesão, ela representa a relação do nódulo com as estruturas vizinhas; portanto, limites mal definidos são mais sugestivos para malignidade (MUDIGONDA et al., 2001). Em um nódulo mamográfico a margem deve ser descrita como: circunscrita, obscurecida, microlobulada, mal definida e espiculada (Figura 6).



**Figura 6:** A Margem da massa pode ser dividida em cinco contornos. Fonte: Kopans (2000).

• **Densidade:** as microcalcificações malignas geralmente são hiperatenuantes e têm importante variação de densidade dentro das partículas e entre elas, porém a densidade não foi considerada como um confiável preditor de malignidade. Calcificações que não são consideradas tipicamente benignas ou tipicamente malignas são consideradas indeterminadas e devem ser tratadas como malignas até que se prove o contrário (TSE GM et al., 2008), sendo abordadas com estudo anatomopatológico. Uma das dificuldades da análise do exame com microcalcificações é que grande parte das lesões diagnosticadas não apresenta características patognomônicas de benignidade ou malignidade (KESTELMAN FP et al., 2007). A revisão da literatura mostrou muita dificuldade na descrição das formas e na classificação dessas lesões devido à grande superposição das características benignas e malignas. Os nódulos mais densos tendem a ser mais suspeitos. A densidade deve ser caracterizada como:

1. **ALTA DENSIDADE.** A atenuação do nódulo com alta densidade é maior do que o esperado em um volume equivalente de tecido fibroglandular mamário.
2. **DENSIDADE IGUAL AO PARÊNQUIMA (MÉDIA).** Antigamente descrito como isodenso, a atenuação do nódulo com igual densidade é semelhante à atenuação esperada em um volume equivalente de tecido fibroglandular mamário.
3. **BAIXA DENSIDADE.** A atenuação do nódulo com baixa densidade é menor do que a esperada para um volume equivalente de tecido fibroglandular mamário.
4. **CONTENDO-GORDURA.** Nessa categoria devem ser incluídos todos os nódulos contendo gordura, como cistos oleosos, lipomas, galactoceles e hamartomas. Os nódulos contendo gordura são quase sempre benignos.

• **BI-RADS** (Breast Image Reporting and Data System) é o nome de um sistema padronizado, criado nos Estados Unidos, utilizado para uniformizar os relatos de radiologia quando se analisam as imagens de mamografia. O BI-RADS foi criado pelo trabalho conjunto entre o Colégio Americano de Radiologia, o NCI (National Cancer Institute), o CDC (Centers for Disease Control and Prevention), a FDA (Food and Drug Administration), o Colégio Americano de Cirurgiões e o Colégio Americano de Patologistas.

A intenção por trás da criação e do uso do BI-RADS é uniformização e a padronização dos laudos de mamografia, sujeitos a confusão na interpretação de resultados quando se utilizam critérios puramente descritivos. Por exemplo, no relato de uma mamografia em que apareçam microcalcificações agrupadas, o uso do BI-RADS pode facilitar a compreensão, entre os não-radiologistas e mastologistas, do significado clínico do achado radiológico. Na prática clínica diária, o BI-RADS fornecido pelo radiologista é uma ferramenta importante para separar as lesões consideradas benignas, as malignas e as suspeitas.

Além dessa função de utilidade na prática clínica imediata, o BI-RADS serve como um controle de qualidade dos dados provenientes da radiografia. Ao se acompanhar a evolução das pacientes por tempo suficiente após feita a mamografia, existe a possibilidade de revisar a adequação do diagnóstico radiológico prévio (ONBENAUER e HERMANN; 2005).

A Classificação de BI-RADS é dividida em um grupo de avaliações completas (1, 2, 3, 4, 5 e 6) e uma avaliação incompleta (zero). A avaliação incompleta requer exames adicionais, com incidências mamográficas diferentes das duas incidências padrão (crânio-caudal e oblíqua-medio-lateral). Podem ser necessários também filmes anteriores para comparação, ultrassom ou ressonância nuclear magnética. Ainda em relação à classificação, alguns profissionais têm como hábito atribuir um BI-RADS a cada uma das mamas. Por exemplo, a mama direita pode ter um BI-RADS 2 e a mama esquerda um BI-RADS 4. A recomendação do Colégio Americano de Radiologia é de que, nesses casos, o laudo do exame contenha uma avaliação final global e categorize o exame com uma classe BI-RADS somente. No exemplo acima, o laudo mamográfico considera o exame como BI-RADS 4. Da mesma forma, se houve uma avaliação incompleta para uma das mamas (BI-RADS zero), e a mama contralateral é BI-RADS 2, o exame tem classificação final BI-RADS zero.

### **Categoria zero**

A categoria BI-RADS zero é utilizada em mamografias de rastreamento quando imagens adicionais são necessárias ou quando é necessária a comparação a exames prévios.

### **Categoria 1**

A categoria BI-RADS 1 significa mamografia negativa. Não são necessários comentários adicionais. As mamas são simétricas, sem massas, distorções de arquitetura ou calcificações suspeitas.

### **Categoria 2**

A categoria BI-RADS 2 significa mamografia negativa, com achados benignos. Em relação ao risco de câncer, é idêntica à categoria BI-RADS 1, mas nela, o radiologista opta por descrever achados benignos característicos, cujo grau de precisão de diagnóstico através da mamografia é grande. Nela, incluem-se fibroadenomas calcificados, calcificações múltiplas de origem secretória, cistos oleosos, lipomas, galactoceles e hamartomas de densidade mista.

### **Categoria 3**

A categoria BI-RADS 3 é utilizada nas avaliações cujo resultado é “provavelmente benigna”. Essa categoria não deve ser utilizada como um sinônimo de exame indeterminado. As lesões



que fazem parte dessa categoria mamográfica devem ter, no máximo, 2% de risco de malignidade. Nessa categoria incluem-se muitas das lesões encontradas em primeiras mamografias ou em mamografias que não têm exame prévio para comparação. As lesões que podem ser corretamente classificadas como BI-RADS 3 incluem massas circunscritas e não-palpáveis em uma mamografia inicial (com exceção de cistos, linfonodos intramamários e achados benignos), assimetrias focais que diminuem ou desaparecem à compressão, e grupamentos de calcificações puntiformes.

A mamografia BI-RADS 3 é um critério para que a paciente seja acompanhada por novo exame mamográfico num período de seis meses. Caso esse novo exame, que não necessariamente é bilateral, seja novamente codificado como BI-RADS 3, e a lesão seja estável, repete-se o acompanhamento novamente em seis meses, totalizando doze meses após o exame inicial. Se nesse décimo segundo mês, o exame mamográfico permanece sendo BI-RADS 3, pode-se repetir o exame doze meses depois, totalizando vinte e quatro meses do exame inicial. Caso a achado mamográfico persista, a mamografia pode ser caracterizada como, a critério do radiologista, como BI-RADS 2 – benigna – ou BI-RADS 3 – provavelmente benigna.

Lesões ultrassonográficas classificadas como BI-RADS 3 incluem cistos não-palpáveis complicados, ou massas ovaladas sólidas, hipoecóicas e circunscritas que não são facilmente distinguíveis de cistos complicados. A incidência de neoplasia maligna nessas lesões é menor do que 2%. Microcistos agrupados sem um componente sólido também podem ser classificadas como BI-RADS 3.

#### **Categoria 4**

A categoria BI-RADS 4, lesão suspeita, inclui lesões na mama que necessitam de avaliação histológica ou citológica adicional. Com o grande número de lesões que é abrigada nessa categoria, foram desenvolvidas subcategorias adicionais.

**Categoria 4A:** nessa categoria incluem-se lesões que necessitam de intervenção mas cujo grau de suspeição é baixo. Aí estão os cistos complicados que necessitam de aspiração, as lesões palpáveis sólidas, parcialmente circunscritas, e que o ultrassom sugere tratarem-se de fibroadenomas, ou um abscesso mamário. O seguimento dessas lesões pode mostrar um diagnóstico anátomo-patológico adicional comprovando malignidade, ou um seguimento semestral benigno.

**Categoria 4B:** nessa categoria estão as lesões de grau intermediário de suspeição. As lesões nessa categoria necessitam de correlação histopatológica. Caso o resultado seja benigno, é necessária a concordância entre os membros da equipe envolvidos no diagnóstico. Uma massa de margens indistintas, com algumas áreas circunscritas, cujo resultado seja necrose gordurosa ou fibroadenoma é um achado aceitável. No entanto, um diagnóstico histopatológico de papiloma pode sugerir que se prossiga a investigação com biópsia excisional.

**Categoria 4C:** nessa categoria estão os achados de grau moderado, mas não clássicos (BI-RADS 5) de neoplasia. Nessa categoria, encontram-se as massas irregulares e mal-definidas, ou novos grupamentos de calcificações pleomórficas. O resultado anátomo-patológico das lesões BI-RADS 4 esperado é o de neoplasia maligna.

#### **Categoria 5**

A categoria BI-RADS 5 é representada por lesões cujo resultado anátomo-patológico, salvo exceções, é o de carcinoma de mama. Nessa categoria, mais do que 95% das lesões

representam câncer de mama, e os achados radiológicos são os característicos da descrição clássica do câncer de mama. Massas espiculadas, irregulares, de alta densidade, ou massas espiculadas de alta densidade associadas a microcalcificações pleomórficas, ou calcificações lineares finas dispostas num segmento ou linearmente estão incluídas na categoria BI-RADS 5. Todas as lesões de alto grau de suspeição, mas que não preenchem os critérios clássicos de câncer de mama, devem ser classificadas como BI-RADS 4.

### **Categoria 6**

A categoria BI-RADS 6 é definida para achados mamográficos já biopsiados cujo diagnóstico anátomo-patológico é de câncer de mama, antes da terapia definitiva. Por exemplo, ela pode ser usada para classificação dos achados de uma mamografia de monitoramento após quimioterapia neoadjuvante, ou para revisões diagnósticas de achados biopsiados.

## **2.1.4. Biópsia**

Biópsia (ou biopsia) é o procedimento cirúrgico no qual se colhe células ou um pequeno fragmento de tecido orgânico para serem posteriormente submetidos a estudo em laboratório, visando determinar a natureza e o grau da lesão estudada. Também podem ser examinados líquidos, secreções, esfregaços e outros materiais orgânicos. Praticamente todos os órgãos e componentes corporais podem ser biopsiados: músculos, pele, ossos, líquidos, secreções, etc. O termo biópsia vem do grego, bios = vida e oipsis = aparência, visão.

Os exames de imagem fornecem uma visão da morfologia ou funcionalidade dos órgãos ou de partes do corpo e os exames bioquímicos oferecem algumas comprovações indiretas do funcionamento intrínseco deles, no entanto, a morfologia das células e tecidos depende de uma análise microscópica de amostras retiradas das pessoas. Assim, o mais comum é proceder-se à biópsia naquelas pessoas com suspeitas diagnósticas de doenças que podem provocar alterações morfológicas nas células e tecidos, como os tumores, por exemplo, ou para estabelecer um diagnóstico diferencial entre enfermidades assemelhadas. Esse exame diagnóstico é indicado tanto em enfermidades simples, como as verrugas, como nas mais graves, como o câncer. Embora o termo biópsia sempre desperte certa apreensão nas pessoas, a maioria delas revela situações simples e benignas. Em doenças infecciosas a biópsia pode ajudar a determinar o agente causal. Em doenças autoimunes uma biópsia ajuda a confirmar ou informar as alterações esperadas em órgãos ou tecidos. Uma biópsia também pode ajudar a avaliar a gravidade da lesão e a evolução do tratamento. Em lesões de malignidade suspeita ou confirmada, as biópsias ajudam a estabelecer o grau histológico de neoplasia e a determinar a natureza, taxa de crescimento e agressividade do tumor, ajudando a elaborar o plano do tratamento e a prever o prognóstico da doença.

De uma maneira geral as biópsias são procedimentos simples, realizados em ambulatório, mas algumas podem demandar internações. As complicações dependem do tipo de intervenção, mas num sentido geral pode ocorrer agravamento de lesões por excesso de manipulação, hemorragias, infecções e/ou formação de fístulas.

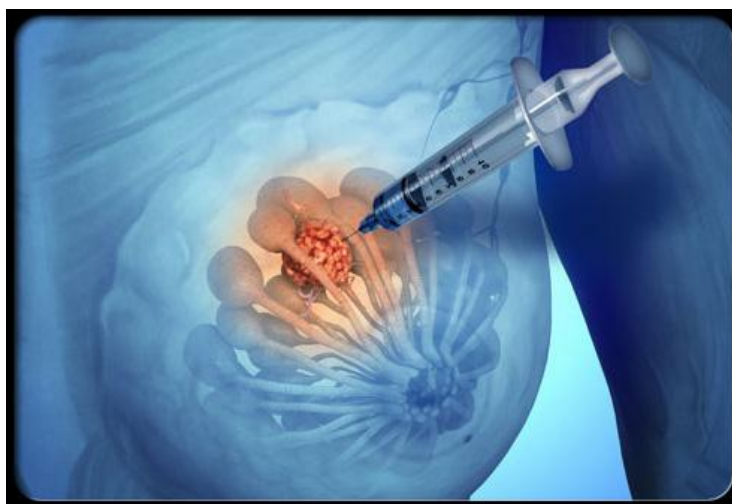
Existem vários tipos de biópsia, dos quais aqui se descrevem os mais comuns. O tipo de biópsia a ser realizado depende do tipo de lesão, do órgão a ser estudado, da hipótese diagnóstica e de condições pessoais do paciente (ABCMED, 2013):

- Biópsias externas: quando feitas em lesões superficiais, geralmente salientes na pele ou mucosas.
- Biópsias internas: feitas por incisão ou punção às cegas ou ecoguiadas (guiadas por ultrassonografia) ou por endoscopia.
- Biópsias perioperatórias: feitas durante uma cirurgia.
- Biópsias incisionais: quando é retirada apenas uma parte da lesão.
- Biópsias excisionais: quando é retirada a lesão inteira.
- Biópsias por aspiração: quando o material a ser examinado é aspirado por uma seringa ou instrumento semelhante ( figura 7).

Em geral as biópsias são realizadas sem necessidade de internação. Apenas em poucos casos a hospitalização pode se impor. Uma biópsia bem feita começa com uma adequada coleta do material. O profissional deve escolher a melhor área da lesão a ser biopsiada, a extensão correta de coleta e o material a ser colhido. O material colhido deverá ser conservado em solução de formol e posteriormente enviado a um laboratório de patologia, para avaliação e emissão de um laudo histopatológico. No caso de tecidos, o patologista fará fatias microscópicas deles, as tingirá com tinturas especiais e as fixará em uma placa de vidro a ser examinada ao microscópio. Os prazos necessários para que se possa produzir esses laudos variam de acordo com o tipo dele são, do material a ser analisado e o procedimento técnico adotado. O prazo médio oscila entre sete e quatorze dias, podendo chegar a um mês em casos de exames mais sofisticados. Nas biópsias feitas durante uma cirurgia, o cirurgião costuma ter um resultado provisório, durante o próprio ato cirúrgico, para orientar o prosseguimento da cirurgia, mas um resultado definitivo só é possível alguns dias depois. As biópsias podem ser orientadas por imagens de ultrassonografia, radiografia, tomografia computadorizada ou ressonância magnética, que ajudam a determinar o local exato em que o material deve ser colhido.

Uma biópsia serve para esclarecer um diagnóstico se ainda restarem dúvidas depois da história clínica, exame físico, dados bioquímicos e de imagens.

Em casos de tumores, além de definir o diagnóstico, podem estabelecer a malignidade ou não deles e o grau histológico em que se encontram.



**Figura 7:** Punção Aspirativa por agulha fina (PAAF) Instituto de Mastologia, 2016.

## 2.2. Aprendizado de Máquina (AM)

As técnicas de AM empregam um princípio de inferência denominado indução, no aprendizado indutivo pode ser dividido em dois tipos principais: supervisionado e não-supervisionado.

No aprendizado supervisionado tem-se a figura de um professor externo, o qual apresenta o conhecimento do ambiente por conjuntos de exemplos na forma: entrada, saída desejada (S. HAYKI, 1999). O algoritmo de AM extrai a representação do conhecimento a partir desses exemplos. O objetivo é que a representação gerada seja capaz de produzir saídas corretas para novas entradas não apresentadas previamente, visto na figura 8.

No aprendizado não-supervisionado não há a presença de um professor, ou seja, não existem exemplos rotulados. O algoritmo de AM aprende a representar (ou agrupar) as entradas submetidas segundo uma medida de qualidade. Essas técnicas são utilizadas principalmente quando o objetivo for encontrar padrões ou tendências que auxiliem no entendimento dos dados (M. C. P. SOUTO et al. 2003).

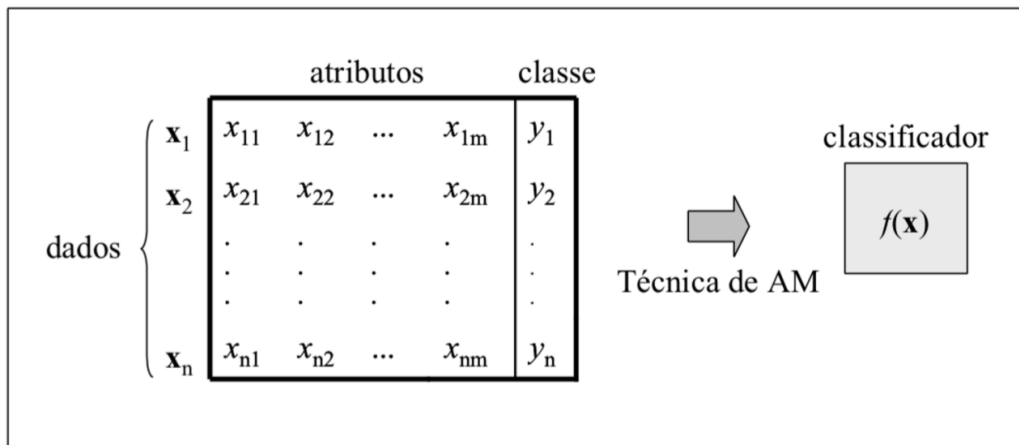
O tipo de aprendizado abordado neste trabalho é o supervisionado. Neste caso, dado um conjunto de exemplos rotulados na forma  $(x_i, y_i)$ , em que  $x_i$  representa um exemplo e  $y_i$  denota o seu rótulo, deve-se produzir um classificador, também denominado modelo, preditor ou hipótese, capaz de predizer precisamente o rótulo de novos dados. Esse processo de indução de um classificador a partir de uma amostra de dados é denominado treinamento. O classificador obtido também pode ser visto como uma função  $f$ , a qual recebe um dado  $x$  e fornece uma predição  $y$ .

Os rótulos ou classes representam o fenômeno de interesse sobre o qual se deseja fazer previsões. Neste trabalho, considera-se o caso em que os rótulos assumem valores discretos  $1, \dots, k$ . Tem-se então um problema de classificação. Caso os rótulos possuam valores contínuos, tem-se uma regressão (T. MITCHELL, 1997). Um problema de classificação no qual  $k = 2$  é denominado binário. Para  $k > 2$ , configurará-se um problema multiclases.

Cada exemplo, também referenciado por dado ou caso, é tipicamente representado por um vetor de características. Cada característica, também denominada atributo, expressa um determinado aspecto do exemplo (M. C. MONARD AND J. A. BARANAUSKAS, 2003). Normalmente, há dois tipos básicos de atributos: nominal e contínuo. Um atributo é definido como nominal (ou categórico) quando não existe uma ordem entre os valores que ele pode assumir (por exemplo, entre cores). No caso de atributos contínuos, é possível definir uma ordem linear nos valores assumidos (por exemplo, entre pesos  $\in \mathbb{R}$ ).

Um requisito importante para as técnicas de AM é que elas sejam capazes de lidar com dados imperfeitos, denominados ruídos. Muitos conjuntos de dados apresentam esse tipo de caso, sendo alguns erros comuns a presença de dados com rótulos e/ou atributos incorretos. A técnica de AM deve idealmente ser robusta a ruídos presentes nos dados, procurando não fixar a obtenção dos classificadores sobre esse tipo de caso. Deve-se também minimizar a influência de outliers no processo de indução. Os outliers são exemplos muito distintos dos demais presentes no conjunto de dados. Esses dados podem ser ruídos ou casos muito particulares, raramente presentes no domínio.

Os conceitos referentes à geração de um classificador a partir do aprendizado supervisionado são representados de forma simplificada na Figura 8. Tem-se nessa figura um conjunto com  $n$  dados. Cada dado  $x_i$  possui  $m$  atributos, ou seja,  $x_i = (x_{i1}, \dots, x_{im})$ . As variáveis  $y_i$  representam as classes. A partir dos exemplos e as suas respectivas classes, o algoritmo de AM extrai um classificador. Pode-se considerar que o modelo gerado fornece uma descrição compacta dos dados fornecidos (J. A. BARANAUSKAS AND M. C. MONARD, 2000).



**Figura 8:** Indução de classificador em aprendizado supervisionado.

A obtenção de um classificador por um algoritmo de AM a partir de uma amostra de dados também pode ser considerada um processo de busca (T. MITCHELL, MCGRAW HILL, 1997). Procura-se, entre todas as hipóteses que o algoritmo é capaz de gerar a partir dos dados, aquela com melhor capacidade de descrever o domínio em que ocorre o aprendizado.

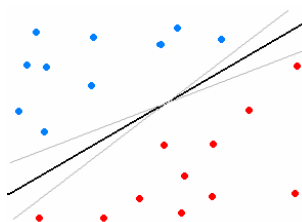
Para estimar a taxa de predições corretas ou incorretas (também denominadas taxa de acerto e taxa de erro, respectivamente) obtidas por um classificador sobre novos dados, o conjunto de exemplos é, em geral, dividido em dois subconjuntos disjuntos: de treinamento e de teste. O subconjunto de treinamento é utilizado no aprendizado do conceito e o subconjunto de teste é utilizado para medir o grau de efetividade do conceito aprendido na predição da classe de novos dados.

Um conceito comumente empregado em AM é o de generalização de um classificador, definida como a sua capacidade de prever corretamente a classe de novos dados. No caso em que o modelo se especializa nos dados utilizados em seu treinamento, apresentando uma baixa taxa de acerto quando confrontado com novos dados, tem-se a ocorrência de um superajustamento (overfitting). É também possível induzir hipóteses que apresentem uma baixa taxa de acerto mesmo no subconjunto de treinamento, configurando uma condição de subajustamento (underfitting). Essa situação pode ocorrer, por exemplo, quando os exemplos de treinamento disponíveis são pouco representativos ou quando o modelo obtido é muito simples (M. C. MONARD AND J. A. BARANAUSKAS, 2003). Na Seção 3, esses conceitos são ilustrados e discutidos novamente. São feitas então considerações e motivações sobre a escolha de classificadores com boa capacidade de generalização.

### 2.2.1. Máquinas de Vetores de Suporte

A Máquina de Vetores de Suporte (SVM) introduzida por (VAPNIK, 1998) é um método de aprendizagem supervisionada usado para estimar uma função que classifique dados de entrada em duas classes. A ideia básica por trás da SVM é construir um hiperplano com superfície de decisão, de tal maneira que a margem de separação entre as classes seja a máxima. O objetivo do treinamento através de SVM é a obtenção de hiperplanos que dividam as amostras de tal maneira que sejam otimizados os limites de generalização.

As SVM são consideradas sistemas de aprendizagem que utilizam um espaço de hipóteses de funções lineares em um espaço de muitas dimensões. Os algoritmos de treinamento das SVM possuem uma forte influência da teoria da otimização e de aprendizagem estatística. Essa teoria estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida como a sua capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu. Em poucos anos, as SVM vem demonstrando sua superioridade frente a outros classificadores em uma grande variedade de aplicações (CRISTIANINI and SHAWE-TAYLOR 2000).



**Figura 9:** Separação de duas classes através de hiperplanos.

Em casos em que o conjunto de amostras é composto por duas classes separáveis, um classificador SVM é capaz de encontrar um hiperplano baseado em um conjunto de pontos denominados “vetores de suporte”, o qual maximiza a margem de separação entre as classes. Por hiperplano entende-se uma superfície de separação de duas regiões num espaço multidimensional, onde o número de dimensões possíveis pode ser, até, infinito. 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. A Figura 10 mostra em duas dimensões para melhor visualização, hiperplanos de separação entre duas classes linearmente separáveis. O hiperplano ótimo (linha central), não somente separa as duas classes, mas mantém a maior distância possível com relação aos pontos da amostra.

Seja conjunto de amostras de treinamento  $(x_i, y_i)$  sendo  $x_i \in \mathbb{R}^n$  o vetor de entrada, a  $y_i$  a classificação correta das amostras e  $i = 1, \dots, n$  o índice de cada ponto amostral. O objetivo da classificação é estimar a função  $f: \mathbb{R}^n \rightarrow \{\pm 1\}$ , que separe corretamente os exemplos de testes em classes distintas.

A etapa de treinamento estima a função  $f(x) = (w \cdot x) + b$ , procurando por valores de  $w$  e  $b$  tais a seguinte relação seja satisfeita:

$$y_i((w \cdot x_i) + b) \geq 1 \quad (1)$$

sendo  $w$  o vetor normal ao hiperplano de decisão e  $b$  o corte ou distância da função  $f$  em relação a origem. Os valores ótimos de  $w$  e  $b$  serão encontrados ao minimizar a seguinte equação, de acordo com a restrição dada pela Equação 1 (CHAVES, 2006).

$$\phi(w) = \frac{w^2}{2} \quad (2)$$

SVM ainda possibilita encontrar um hiperplano que minimize a ocorrência de erros de classificação nos casos em que uma perfeita separação entre as duas classes não for possível. Isso graças a inclusão de variáveis de folga, que permitem que as restrições presentes na Equação 1 sejam quebradas.

O problema de otimização passa a ser então minimização da Equação 3, de acordo com a restrição imposta pela Equação 2.1  $C$  é um parâmetro de treinamento que estabelece um equilíbrio entre a complexidade do modelo e o erro de treinamento e deve ser selecionado pelo o usuário.

$$\phi(w, \epsilon) = \frac{w^2}{2} + C \sum_{i=1}^N \epsilon_i \quad (3)$$

Para

$$y_i((w \cdot x_i) + b) + \epsilon_i \geq 1 \quad (4)$$

Através da teoria dos multiplicadores de Lagrange, chega-se a Equação 5. O objetivo então passa a ser encontrar os multiplicadores de Lagrange  $\alpha_i$  ótimos que satisfaça a Equação 6 (CHAVES, 2006).

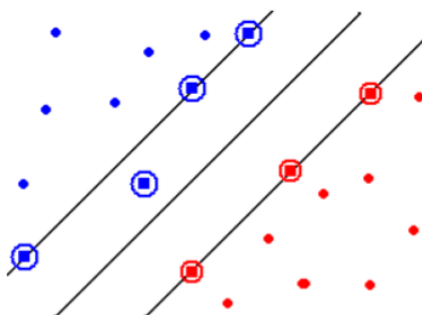
$$w(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j (x_i x_j) \quad (5)$$

$$\sum_{i=1}^N \alpha_i y_i = 0, 0 \leq \alpha_i \leq C \quad (6)$$

Somente os pontos onde a restrição 1 seja exatamente igual a unidade tem

correspondentes  $\alpha \neq 0$ . Esses pontos são chamados de vetores de suporte, pois se localizam geometricamente sobre as margens. Tais pontos tem fundamental importância na definição do hiperplano ótimo, pois os mesmos delimitam a margem do conjunto de treinamento.

Na Figura 10 destacam os pontos que representam os vetores de suporte. Os pontos além da margem não influenciam decisivamente na determinação do hiperplano, enquanto que os vetores de suporte, por terem pesos não nulos, são decisivos.



**Figura 10:** Vetores de Suporte (destacados por círculos).

Para que a SVM possa classificar amostras que não são linearmente separáveis, é necessário uma transformação não linear que transforme o espaço entrada (dados) para um novo espaço (espaço de características). Esse espaço deve apresentar dimensão suficientemente grande, e através dele, a amostra pode ser linearmente separável. Dessa forma, o hiperplano de separação é definido como uma função linear de vetores retirados do espaço de característica ao invés do espaço de entrada original. Essa construção depende do cálculo da função K de núcleo de um produto interno (HAYKIN and ENGEL, 2001). A função K pode realizar o mapeamento das amostras para um espaço de dimensão muito elevada sem aumentar a complexidade dos cálculos.

A Equação 7 mostra o resultado da Equação 5 com a utilização de um núcleo.

$$w(x) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j k(x_i, x_j) \quad (7)$$

Uma importante família de funções de núcleo é a função de base radial, muito utilizada em reconhecimento de padrões.



## 2.2.2. Máquinas de Vetores de Suporte Linearmente Separável

O problema de classificação binária, problema de classificação inicial tratado pela SVM, trata da classificação de duas classes, sem perda de generalidade, através de um hiperplano ótimo à partir de conjunto de treinamento linearmente separável se for possível separar os padrões de classes diferentes contidos no mesmo por pelo menos um hiperplano (Haykin 2001, Semolini 2002).

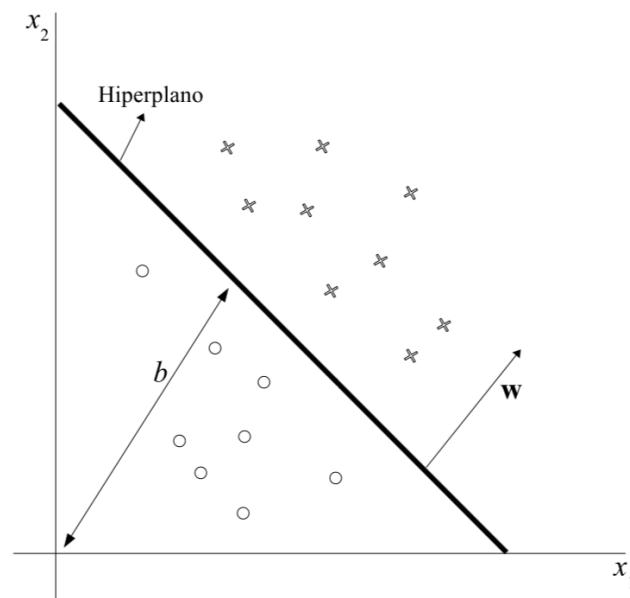
Considere o conjunto de treinamento  $\{(x_i, d_i)\}_{i=1}^N$ , onde  $x_i$  é o padrão de entrada o  $i$ -ésimo exemplo e  $d_i$  é a resposta desejada,  $d_i = \{+1, -1\}$  que representa as classes linearmente separáveis.

A equação que separa os padrões através do hiperplano pode ser definida por:

$$\mathbf{w}^T \cdot \mathbf{x} + b = 0 \quad (8)$$

onde,  $\mathbf{w}^T \cdot \mathbf{x}$  é produto escalar entre os vetores  $\mathbf{w}$  e  $\mathbf{x}$ , em  $\mathbf{x}$  é o vetor de entrada que representa os padrões de entrada do conjunto treinamento,  $\mathbf{w}$  é vetor de pesos ajustáveis e  $b$  é um limiar conhecido como bias.

A Figura 11 mostra o hiperplano de separação  $(\mathbf{w}, b)$  um espaço bidimensional para um conjunto de treinamento linearmente separável.

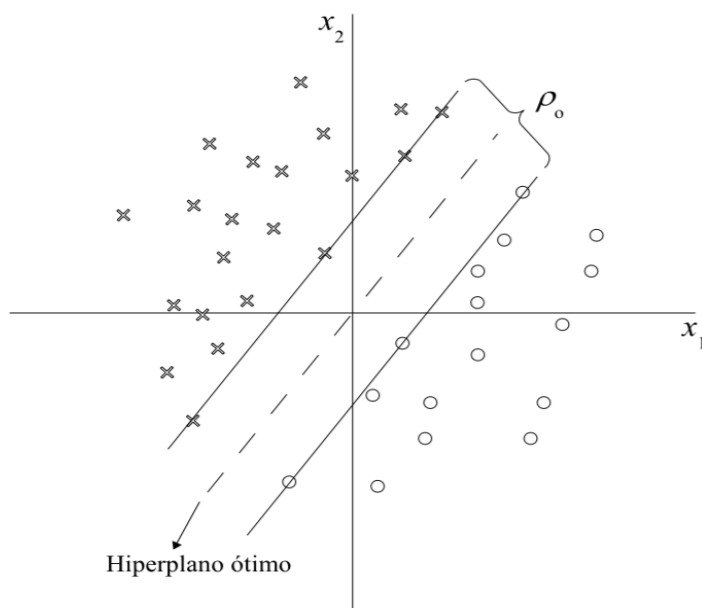


**Figura 11:** Hiperplano de separação  $(\mathbf{w}, b)$  para um conjunto de treinamento bidimensional.

A figura equação (8) pode ser reescrita por:

$$\begin{aligned} w^T \cdot x + b &\geq 0, \text{ se } d_i = +1 \\ w^T \cdot x + b &< 0, \text{ se } d_i = -1 \end{aligned} \quad (9)$$

A margem de separação, distância entre o hiperplano definido na equação 1 e o ponto mais próximo de ambas as classes é representado por  $\rho$ . O objetivo de uma SVM é encontrar um hiperplano que separe o conjunto de treinamento sem erro e maximize a margem de separação, sobre essa condição, o hiperplano é referido como hiperplano ótimo. A Figura 12 ilustra o hiperplano ótimo para um espaço de entrada bidimensional.



**Figura 12:** Hiperplano ótimo com máxima margem  $\rho_0$  de separação dos padrões linearmente separáveis.

Considerando que  $w_0$  e  $b_0$  representam os valores ótimos do vetor peso e do bias, respectivamente a equação 10 pode ser reescrita para hiperplano como:

$$w_0^T \cdot x_0 + b_0 = 0 \quad (10)$$

A função discriminante

$$g(x) = w_0^T \cdot x_0 \quad (11)$$

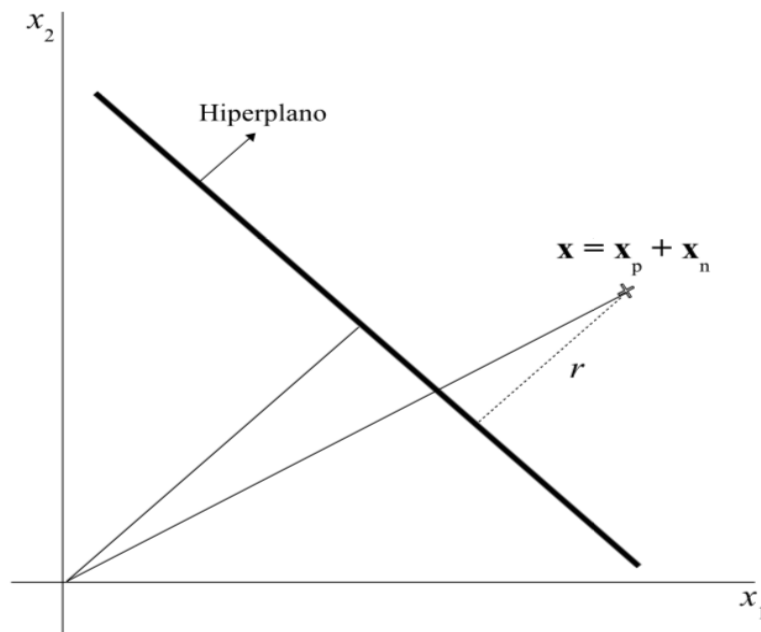
fornece uma medida algébrica de distância  $r$  entre  $x$  e o hiperplano  $(w_0, b_0)$  que pode ser representada por:

$$x = x_p + x_n \quad (12)$$

onde  $x_p$  é a projeção normal de  $x$  sobre o hiperplano ótimo e  $x_n$  é o vetor normal com a distância  $r$ , onde,

$$x_n = r \cdot \frac{w_0}{\|w_0\|} \quad (13)$$

Na Figura 13 ilustra a distância  $r$  entre  $x$  e o hiperplano  $(w_0, b_0)$ , onde  $r$  é positivo se  $x$  estiver no lado positivo do hiperplano ótimo caso contrário será negativo.



**Figura 13:** Interpretação gráfica da distância  $x$  até o hiperplano ótimo.

### 2.2.1.1. Hiperplano Ótimo para Classes Linearmente Separáveis

O hiperplano ótimo definido para parâmetros os parâmetros  $\mathbf{w}$  e  $\mathbf{b}$  que satisfaçam as desigualdades da equação 14.

$$d_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 \quad (14)$$

O objetivo da SVM é encontrar um procedimento computacional que, utilizando o conjunto de treinamento  $\{(x_i, d_i)\}_{i=1}^N$  encontra o hiperplano ótimo sujeito as restrições da equação 14. Sendo resolvido através do problema de otimização com restrições minimizando a função custo  $\phi$  em relação ao vetor  $\mathbf{w}$  e satisfazendo a restrição da equação 14.

$$\phi(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \cdot \mathbf{w} \quad (15)$$

A partir da função custo da equação 15 pode ser formulado o problema de otimização com restrições, chamado de problema primal:

$$\begin{array}{ll} \text{Minimizar:} & \frac{1}{2} \mathbf{w}^T \cdot \mathbf{w} \\ \text{Restrições:} & d_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1, \text{ para } i = 1, \dots, n \end{array} \quad (16)$$

Este é um problema clássico de otimização de programação quadrática (HEARST, 1998) sob aspecto de aprendizado de máquina. O problema de otimização analisado sob o ponto de vista de otimização quadrática pode ser resolvido induzindo uma função langrangiana, definida em termos de  $\mathbf{w}$  e  $\mathbf{b}$ .

$$J(\mathbf{w}, \mathbf{b}, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^N \alpha_i (d_i(\mathbf{w}^T \cdot \mathbf{x}_i + b)) \quad (17)$$

onde, os  $\alpha_i$  são denominados multiplicadores de Lagrange não-negativos.

O problema passa a ser a minimização da equação 17 em relação a  $\mathbf{w}$  e  $\mathbf{b}$  e maximização de  $\alpha_i$ , com  $\alpha_i \geq 0$ . Os pontos ótimos dessa equação são obtidos diferenciando a equação (2.17) em relação  $\mathbf{w}$  e  $\mathbf{b}$  e igualando os resultados obtidos a zero, obtendo as condições de otimização:

$$\begin{aligned}
\text{Condição 1: } & \frac{\partial(w,b,\alpha)}{\rho(w)} = 0 \\
\text{Condição 2: } & \frac{\partial(w,b,\alpha)}{\partial(b)} = 0
\end{aligned} \tag{18}$$

A aplicação das condições de 18 à função lagrangiana da equação 17 leva ao resultado:

$$\begin{aligned}
w &= \sum_{i=1}^N \alpha_i d_i x_i \\
\sum_{i=1}^N \alpha_i d_i &= 0
\end{aligned} \tag{19}$$

Substituindo a equação 19 em 17, obtém-se o problema dual de otimização:

$$\begin{aligned}
\text{Maximizar: } & \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j x_i^T \cdot x_j \\
\text{Restrições: } & \begin{cases} (1) \alpha_i \geq 0, \quad i = 1, \dots, N \\ (2) \sum_{i=1}^N \alpha_i d_i = 0 \end{cases}
\end{aligned} \tag{20}$$

Tendo encontrado os multiplicadores de Lagrange ótimos, representados por  $w_{0i}$ , pode-se calcular o vetor de peso ótimo  $w_0$  através da equação 19:

$$w_0 = \sum_{i=1}^N \alpha_{0i} d_i x_i \tag{21}$$

O valor do bias ótimo  $b_0$  é encontrado encontrando os pesos ótimos  $w_0$  encontrados na equação 21 e descrito como:

$$b_0 = 1 - w_0^t \cdot x^{(s)} \text{ para } d^{(s)} = 1 \tag{22}$$

O problema dual (20) é formulado totalmente em termos dos padrões de treinamento, além disso, a equação a ser maximizada depende totalmente dos padrões de entrada. O hiperplano ótimo depende somente dos vetores de suporte, considerados os padrões mais significativos do conjunto de treinamento. Os multiplicadores de Lagrange  $\alpha_0 > 0$

(diferentes de zero) são justamente os padrões de entrada com margem igual a 1, chamados vetores de suporte.

O hiperplano ótimo é expresso em termos do conjunto de vetores de suporte descrito pela função sinal como:

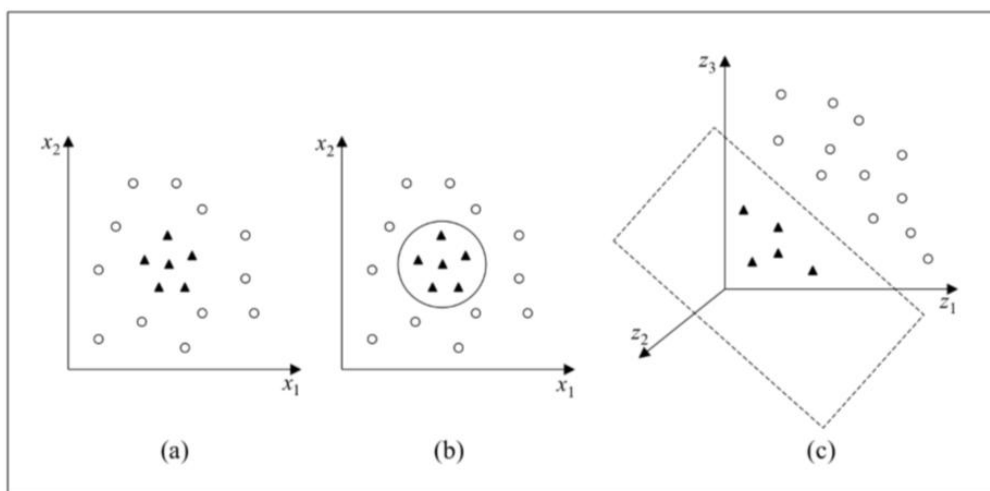
$$f(x) = \text{sgn} \left( \sum_{i=1}^{N_{SV}} d_i \alpha_{0i} x^T \cdot x + b_0 \right) \quad (23)$$

Os padrões de entrada que não vetores de suporte não influenciam na função de decisão da escolha do hiperplano ótimo pela SVM.

Um ponto importante para este problema de otimização é que fornece solução única encontrado eficientemente, diferente de outras técnicas.

### 2.2.2. SVM Não Lineares

As SVMs lineares são eficazes na classificação de conjuntos de dados linearmente separáveis ou que possuam uma distribuição aproximadamente linear, sendo que a versão de margens suaves tolera a presença de alguns ruídos e outliers. Porém, há muitos casos em que não é possível dividir satisfatoriamente os dados de treinamento por um hiperplano. Um exemplo é apresentado na Figura 14, em que o uso de uma fronteira curva seria mais adequada na separação das classes.



**Figura 14:** (a) Conjunto de dados não linear; (b) Fronteira não linear no espaço de entradas; (c) Fronteira linear no espaço de características.

As SVMs lidam com problemas não lineares mapeando o conjunto de treinamento de seu espaço original, referenciado como de entradas, para um novo espaço de maior dimensão,

denominado espaço de características (feature space) (M. A. HEARST et al., 1998). Seja  $\Phi: X \rightarrow \mathcal{F}$  um mapeamento, em que  $X$  é o espaço de entradas e  $\mathcal{F}$  denota o espaço de características. A escolha apropriada de  $\Phi$  faz com que o conjunto de treinamento mapeado em  $\mathcal{F}$  possa ser separado por uma SVM linear.

O uso desse procedimento é motivado pelo teorema de Cover (S. HAYKIN, 1999). Dado um conjunto de dados não linear no espaço de entradas  $X$ , esse teorema afirma que  $X$  pode ser transformado em um espaço de características  $\mathcal{F}$  no qual com alta probabilidade os dados são linearmente separáveis. Para isso duas condições devem ser satisfeitas. A primeira é que a transformação seja não linear, enquanto a segunda é que a dimensão do espaço de características seja suficientemente alta.

Para ilustrar esses conceitos, considere o conjunto de dados apresentado na Figura 15(a) (K. R. Müller, S. Mika, G. Rätsch, K. Tsuda and B. Schölkopf, 2001). Transformando os dados de  $\mathbb{R}^2$  para  $\mathbb{R}^3$  com o mapeamento representado na Equação 34, o conjunto de dados não linear em  $\mathbb{R}^2$  torna-se linearmente separável em  $\mathbb{R}^3$  (Figura 15c). É possível então encontrar um hiperplano capaz de separar esses dados, descrito na Equação 35. Pode-se verificar que a função apresentada, embora linear em  $\mathbb{R}^3$  (Figura 15c), corresponde a uma fronteira não linear em  $\mathbb{R}^2$  (Figura 15b).

$$\Phi(x) = \Phi(x_1, x_2) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2) \quad (24)$$

$$f(x) = w \cdot \Phi(x) + b = w_1 x_1^2 + w_2 \sqrt{2} x_1 x_2 + w_3 x_2^2 + b = 0 \quad (25)$$

Logo, mapeiam-se inicialmente os dados para um espaço de maior dimensão utilizando  $\Phi$  e aplica-se a SVM linear sobre este espaço. Essa encontra o hiperplano com maior margem de separação, garantindo assim uma boa generalização.

### 2.2.2.1. Hiperplano Ótimo para Classes não Linearmente Separáveis

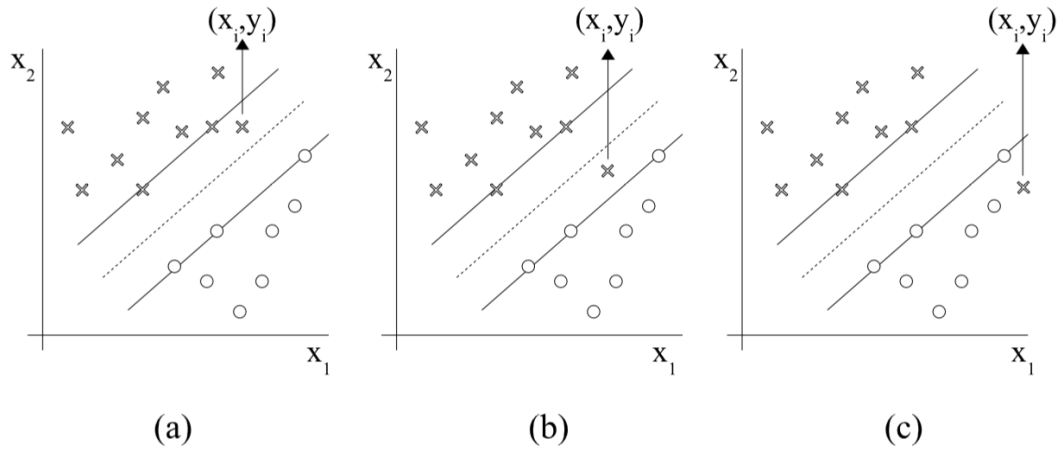
Considerando, em uma visão geral, o problema de classificação, onde, a classe não linearmente separável, a construção de hiperplano de separação, dado os padrões de treinamento possivelmente gerará erros de classificação. O objetivo da SVM é encontrar um hiperplano que minimiza a probabilidade de erro de classificação junto ao conjunto de treinamento.

Existem alguns casos, onde não é necessário fazer um mapeamento de características no conjunto de treinamento. Esses casos tratados pela SVM linear com margens de separação suaves ou flexíveis, pois poderão existir pontos  $(x_i, d_i)$  que violaram a equação 14.

Esta violação pode ocorrer em três diferentes situações:

- O ponto  $(x_i, d_i)$  se encontra dentro da região de separação e no lado correto da superfície de decisão, como mostra na Figura 16 (a). Neste caso, houve uma escolha incorreta do hiperplano.
- O ponto  $(x_i, d_i)$  se encontra dentro da região de separação e no lado incorreto da superfície de decisão, como mostra na Figura 16 (b). Neste caso, houve uma escolha incorreta do hiperplano de margem maior.

- O ponto  $(x_i, d_i)$  se encontra fora da região de separação e no lado incorreto da superfície de decisão, como mostra a Figura 16 (c).



**Figura 15:** (a) o ponto  $(x_i, d_i)$  se encontra na região de separação, mas do lado correto. (b) O ponto  $(x_i, d_i)$  se encontra na região de separação, mas do lado incorreto. (c) O ponto  $(x_i, d_i)$  se encontra fora da região de separação, mas do lado incorreto.

Para tratar desses problemas introduz-se uma variável negativa  $\{\xi_i\}_{1 \leq i \leq N}$  na definição do hiperplano de separação.

$$d_i(w^T \cdot x_i + b) \geq 1 + \xi_i \quad (26)$$

As variáveis  $\xi_i$  são denominadas variáveis soltas, e medem os desvios dos pontos  $(x_i, d_i)$  para a condição ideal de separação de classes. Quando  $\xi_i$  satisfizer  $0 \leq \xi_i \leq 1$  o ponto se encontra-se dentro da região de separação mas do lado correto da superfície de decisão. Quando  $\xi_i > 1$  o ponto encontra-se do lado incorreto do hiperplano de separação.

Os vetores-suporte são pontos que o resultado da equação (2.24) é igual a  $1 - \xi_i$  mesmo que  $\xi_i > 0$ . Ao retirar um padrão do conjunto treinamento em que  $\xi_i > 0$  a superfície de decisão tem possibilidade de mudança, porém, ao retirar um padrão em que  $\xi_i = 0$  e o resultado da equação 24 for maior que 1 a sua superfície de decisão permanecerá inalterada.

O objetivo é encontrar um hiperplano de separação onde o erro de classificação incorreta seja mínimo perante o conjunto de treinamento, podendo ser feito minimizando a equação:

$$\phi(\xi) = \sum_{i=1}^N (\xi_i - 1) \quad (27)$$

em relação ao vetor peso  $w$ , sujeito a restrição, sujeito a restrição da equação do hiperplano de separação da equação 24 e a restrição sobre  $w^T w$ . A função  $I(\xi - 1)$  é uma função indicadora, definida por:



$$I(\xi - 1) = \begin{cases} 0 & \text{se } \xi \leq 0 \\ 0 & \text{se } \xi > 0 \end{cases} \quad (28)$$

A minimização de  $\phi(\xi)$  é um problema de otimização não convexo de classe NP-completo não determinístico em tempo polinomial. Para fazer este problema de otimização matematicamente tratável, aproxima-se a função  $\phi(\xi)$  por:

$$\phi(\xi) = \sum_{i=1}^N \xi_i \quad (29)$$

Para a simplificação de cálculos computacionais a função a ser minimizada em relação ao vetor peso  $w$  segue:

$$\phi(w, \xi) = \frac{1}{2} w^T \cdot w + C \sum_{i=1}^N \xi_i \quad (30)$$

onde, o parâmetro  $C$  controla a entre a complexidade do algoritmo e o número de amostras do conjunto de treinamento classificados incorretamente, sendo denominado *parâmetro de penalização*.

A minimização do primeiro termo da equação 30 está relacionada á minimização da dimensão VC da SVM. O segundo termo pode ser visto como um limitante superior para o número de erros no padrão treinamento apresentados a máquina. Logo, a equação 2.30 satisfaz os princípios de minimização do risco estrutural.

O problema de otimização em sua representação primal para encontrar o hiperplano ótimo de separação para classes não linearmente separáveis pode ser escrito como:

$$\text{Minimizar : } \frac{1}{2} w^T \cdot w + C \sum_{i=1}^N \xi_i \quad (31)$$

$$\text{Sujeito as restrições: } \begin{cases} (1) d_i(w^T \cdot x_i + b) \geq 1 - \xi_i, \text{ para } i = 1, \dots, N \\ (2) \xi_i \geq 0, \forall i = 1, \dots, N \end{cases}$$

Utilizando o método dos multiplicadores de Lagrange, pode-se formular o problema de otimização primal em seu correspondente problema dual de maneira similar a descrita na seção 2.1.

$$\text{Minimizar: } \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j x_i^T \cdot x_j$$

$$\text{Sujeito as restrições: } \begin{cases} (1) \sum_{i=1}^N d_i \alpha_i \\ (2) 0 \leq \alpha_i \leq C \end{cases} \quad (32)$$

onde,  $C > 0$  é especificado pelo usuário.

A principal diferença entre o caso de classes linearmente separáveis, e o caso não linearmente separáveis é que a restrição  $\alpha_i \geq 0$  é substituída por uma mais forte  $0 \leq \alpha_i \leq C$ .

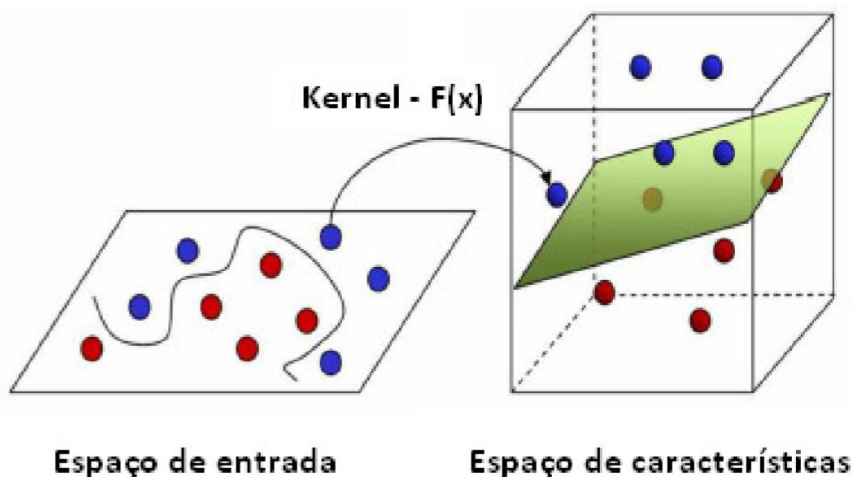
O vetor pesos ótimos  $\mathbf{w}_0$  é calculado da mesma maneira do caso de classes linearmente separáveis, equação 21. O bias ótimo  $\mathbf{b}$  também segue o procedimento similar descrito anteriormente, equação 22.

Existem casos também onde, há a necessidade de mapear o espaço de entrada não linear para um espaço de características. Para realizar esse mapeamento, as funções Kernel ou produto do núcleo interno são utilizados.

### 2.2.2.2. Funções Kernel

As funções de kernel têm a finalidade de projetar os vetores de características de entrada em um espaço de características de alta dimensão para classificação de problemas que se encontram em espaços não linearmente separáveis. Isso é feito, pois a medida que se aumenta o espaço da dimensão do problema, aumenta também a probabilidade desse problema se tornar linearmente separável em relação a um espaço de baixa dimensão. Entretanto, para obter uma boa distribuição para esse tipo de problema é necessário um conjunto de treinamento com um elevado número de instâncias (GONÇALVES, 2010).

A Figura 16 mostra o processo de transformação de um domínio não linearmente separável, em um problema linearmente separável através do aumento da dimensão, onde é feito um mapeamento por uma função de kernel  $F(x)$ .



**Figura 16:** Transformação: problema não linearmente separável em um problema linearmente separável (Rebelo 2008).

Uma função é definida como sendo uma função de kernel, se obedecer à seguinte teoria de Hilbert (RKHS – Reproducing Kernel Hilbert Space ):

$$k(x_i, x_j) = \{\phi(x_i) \cdot \phi(x_j)\} \quad (33)$$

onde  $\phi(\cdot)$  deve pertencer a um domínio onde seja possível o calculo do produto interno. Essas funções também satisfazem as condições do Teorema de Mercer (Gonçalves 2010).

**Teorema de Mercer:** uma determinada função é definida como sendo de kernel, se a matriz K é positivamente definida (autovalores maiores que zero), onde K é obtida por:

$$k = k_{i,j} = k(x_i, x_j) \quad (34)$$

Essas funções de kernel também são conhecidas como Kernels de Mercer . Existem vários tipos de kernels que podem ser usados, porém as mais conhecidas são mostradas na Tabela 1.

**Tabela 2:** Funções de kernel mais usadas.

Tipo de Kernel	Função $k(x_i, x_j)$	Tipo de Classificador
Polinomial	$((x_i \cdot x_j) + 1)^p$	Máquina de aprendizagem polinomial
Gaussiano ou RBF	$\exp - \left( \frac{\ x_i - x_j\ ^2}{2\sigma^2} \right)$	Rede RBF
Sigmoidal	$\tanh(\beta_0(x_i \cdot x_j)) + \beta_1$	Perceptron de duas camadas

## CAPÍTULO III

### 3. MATERIAIS E MÉTODOS

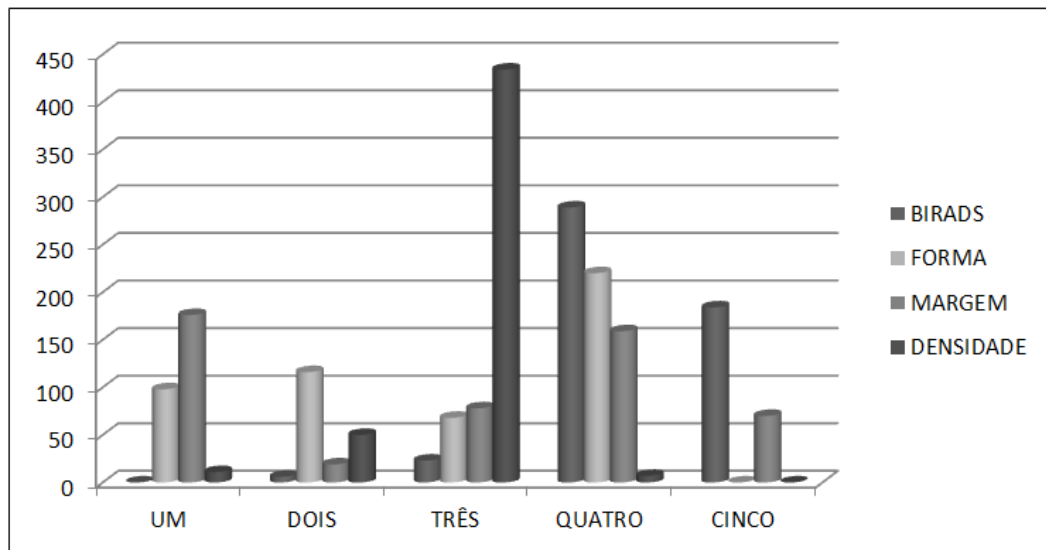
#### 3.1. Base de Dados

O conjunto de dados utilizado nesse estudo consiste de 961 sequencias que depurados passaram a ser 503, proveniente de pacientes com suspeita de câncer de mama obtidos junto ao Instituto de Radiologia da Universidade Erlangen-Nuremberg, no período de 2003 a 2006.

O banco de dados possui informações clínicas sobre Birads, Forma, Margem, Densidade e Diagnóstico. Um resumo das variáveis e características do conjunto de dados é demonstrado na tabela 3 e figura 17.

**Tabela 3:** Classificação das variáveis do conjunto de dados

VARIÁVEIS	CATEGORIAS
<b>BIRADS</b>	<b>Varia na escala de 1 a 5</b>
<b>FORMA</b> (forma de massa mamográfica)	redonda = 1 oval = 2 oval lobulada= 3 lobular irregular = 4
<b>MARGEM</b> (margem de massa mamográfica)	circunscrito = 1 microlobulada = 2 obscurecida =3 espiculado =4 mal definida = 5
<b>DENSIDADE</b> (a densidade de massa mamográfica)	alta = 1 médio=2 baixo = 3 contendo gordura = 4
<b>DIAGNÓSTICO</b>	Maligno=1 Benigno=0



**Figura 17:** Distribuição das característica do banco de dados.

### 3.3. Modelagem Computacional

#### 3.3.1. Modelo

O método proposto neste estudo, consiste em um modelo computacional baseado em máquina de vetor de suporte não linear (SVM) na classificação de malignidade de massa mamária obtidas em dados mamográficos. O fluxograma do modelo proposto é exemplificado na figura 18 e suas etapas são descritas a seguir:

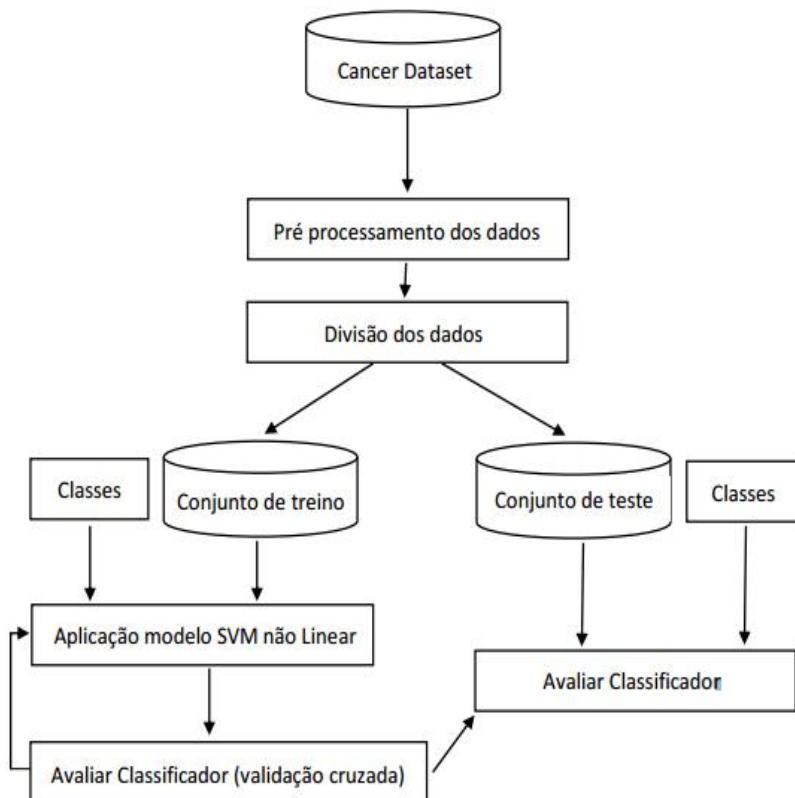


Figura 18: Fluxograma do modelo proposto.

O pré processamento dos dados consistiu em eliminar na base de dados as amostras com dados faltantes. Sendo então a amostra dividida em dois grupos independentes, o conjunto de treinamento e o conjunto de teste. Um resumo das características desses conjuntos de dados pode ser vista nas tabelas 4 e 5.

**Tabela 4:** Distribuição no conjunto treino.

BIRADS	FORMA	MARGEM	DENSIDADE	MALIGNIDADE
2:4	1:77	1:136	1:9	benigno: 203
3:20	2:89	2:12	2:40	maligno: 199
4:227	3:54	3:67	3:347	
5:151	4:182	4:130	4:6	
	5:57			

**Tabela 5:** Distribuição no conjunto teste.

BIRADS	FORMA	MARGEM	DENSIDADE	MALIGNIDADE
2:2	1:21	1:40	1:2	benigno: 58
3:3	2:27	2:7	2:10	maligno: 43
4:63	3:14	3:11	3:88	
5:33	4:39	3:30	4:1	
	5:13			

O conjunto de treinamento composto por 80% dos dados e o conjunto teste composto com os 20% restantes obtidos de forma aleatória.

O modelo SVM não linear é um algoritmo de classificação, que envolve então a escolha de uma função kernel, além de parâmetros dessa função e do valor da constante de regularização  $C$ . A escolha do kernel e dos parâmetros considerados afeta o desempenho do classificador obtido, pois eles definem a fronteira de decisão induzida. ((HSU et al., 2003).

A função de kernel utilizada e os valores dos parâmetros  $C$  e  $\sigma$  foram estimados através da avaliação dos resultados de simulações com amostras de treinamento utilizando validação cruzada. O objetivo dessa técnica é treinar da melhor forma possível a máquina para que ela possa generalizar sobre as futuras entradas. A desvantagem dessa técnica é que o modelo tem que ser treinado  $K$  vezes, demandando certo tempo, afetando de forma negativa o desempenho.

Na etapa de treinamento gera-se o modelo com os vetores de suporte que é utilizado pelo modelo proposto (SVM não linear) na etapa de testes. A etapa de treinamento desconhece totalmente as amostras contidas no conjunto de teste (conjuntos independentes). Após o modelo gerado, é aplicado no conjunto de teste de sorte a verificar sua eficácia na classificação das microcalcificações mamárias.

A performance do modelo SVM é avaliado pela acurácia de classificação ou precisão total (ACC), sensibilidade (S) e especificidade (E), definidos respectivamente por:

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \quad (35)$$

$$S = \frac{VP}{VP + FN} \quad (36)$$

$$E = \frac{VN}{VN + FP} \quad (37)$$

A acurácia ou precisão total define a taxa de casos identificados corretamente sobre o total de casos.

A sensibilidade define a proporção de verdadeiros-positivos identificados no teste. Indica o quão eficaz é o teste para identificar os indivíduos doentes.

A especificidade define a proporção de verdadeiros-negativos identificados no teste. Indica o quão bom é o teste para identificar indivíduos não doentes.

- VP- Verdadeiro Positivo: o teste é positivo e o paciente tem a doença.
- FP- Falso Positivo: o teste é positivo, mas paciente não tem a doença.
- VN- Verdadeiro Negativo: o teste é negativo e o paciente não tem a doença.
- FN- Falso Negativo: o teste é negativo, mas paciente tem a doença.

Ao final das 50 simulações, são selecionadas as que apresentarem os melhores resultados em termos do erro de treino. Para tal, os valores do erro de treino são ordenados em ordem crescente e as simulações que apresentarem resultado superior ao limite de distribuição do primeiro quartil serão selecionadas.

Após essa seleção, será calculado a média nesse conjunto de simulações, visando obter um resultado mais robusto do modelo SVM, na classificação das microcalcificações mamárias.

Os parâmetros utilizados no modelo SVM na classificação foram obtidos de forma experimentalmente. A relação dos parâmetros está sumarizada na tabela 5.

**Tabela 6:** Tabelas de parâmetros.

Parâmetros	Valor
Simulações	50
Tipo do classificador	C-svc
Função de kernel	rbf
Variância da função de kernel ( $\sigma$ )	[0,5; 0,25; 0,125; 0,05]
Parâmetro de regularização (C)	[ 0,1; 1,0; 10; 100]
Critério de parada (tol)	0,001
Erro validação cruzada	[1, 10, 100]



O modelo computacional proposto nesse estudo utilizou-se do pacote Kernel-Based Machine Learning Lab (*kernelab*) , (KARATZOGLOU et al, 2016), na implementação do classificador SVM não linear e o script foi desenvolvido utilizando o software R (<http://www.R-project.org/>).

## CAPÍTULO IV

### 4. RESULTADOS

#### 4.1. Resultados

A utilização o modelo computacional SVM não linear proposto nesse trabalho foi avaliado com a incorporação de todas as variáveis (Birads, Margem, Forma e Densidade) e excluindo uma variável por vez no conjunto da dados de pacientes portadores de microcalcificação mamárias obtidas por meio de mamografia.

##### 4.1 Modelo completo (BMFD)

O desempenho obtido no conjunto das 50 simulações realizadas pelo modelo SVM não linear na categorização de malignidade de microcalcificação mamária , ordenado em função do erro de treinamento, número de vetores de suporte, erro de validação cruzada (*leave-one-out*), precisão total, sensibilidade e especificidade no conjunto de teste, são apresentados na tabela 7.

**Tabela 7:** Resultado das 50 simulações utilizando SVM com as variáveis: Birads, Margem, Forma e Densidade (BMFD).

erro_treino	n_vetor	erro_vc	Precisao_Total	Sensibilidade	Especificidade
0,180	209	0,215	0,724	0,694	0,750
0,183	212	0,195	0,697	0,722	0,675
0,185	217	0,218	0,684	0,579	0,789
0,185	212	0,196	0,697	0,725	0,666
0,187	216	0,219	0,724	0,741	0,714
0,187	210	0,210	0,724	0,706	0,738
0,190	214	0,241	0,724	0,609	0,857
0,190	218	0,223	0,737	0,771	0,707
0,190	220	0,210	0,750	0,825	0,667
0,192	212	0,243	0,737	0,567	0,897
0,192	212	0,209	0,750	0,604	0,939
0,192	214	0,208	0,750	0,806	0,711
0,194	218	0,243	0,737	0,743	0,732
0,194	219	0,236	0,737	0,871	0,644
0,194	213	0,207	0,737	0,538	0,945
0,194	216	0,230	0,750	0,621	0,872
0,194	211	0,202	0,750	0,781	0,727
0,194	214	0,227	0,763	0,816	0,710
0,197	218	0,208	0,737	0,567	0,897
0,197	214	0,205	0,737	0,567	0,897

0,197	215	0,239	0,763	0,731	0,800
0,197	211	0,237	0,763	0,611	0,900
0,197	217	0,239	0,789	0,763	0,816
0,199	208	0,218	0,737	0,659	0,842
0,199	222	0,237	0,750	0,555	0,925
0,199	212	0,209	0,750	0,619	0,912
0,199	216	0,229	0,763	0,714	0,805
0,199	227	0,229	0,803	0,880	0,706
0,199	220	0,263	0,816	0,750	0,875
0,199	223	0,220	0,816	0,839	0,800
0,201	222	0,235	0,737	0,588	0,857
0,201	211	0,213	0,763	0,617	0,881
0,201	217	0,244	0,776	0,725	0,833
0,201	218	0,241	0,776	0,727	0,843
0,201	220	0,244	0,789	0,743	0,829
0,201	227	0,239	0,803	0,725	0,889
0,201	226	0,225	0,816	0,765	0,857
0,204	221	0,244	0,763	0,821	0,729
0,204	211	0,260	0,789	0,676	0,881
0,204	224	0,221	0,789	0,674	0,939
0,206	212	0,261	0,789	0,757	0,820
0,206	219	0,249	0,803	0,738	0,882
0,206	216	0,269	0,829	0,848	0,814
0,208	226	0,265	0,816	0,684	0,947
0,208	219	0,222	0,816	0,656	0,932
0,211	227	0,230	0,816	0,790	0,848
0,213	230	0,222	0,816	0,636	0,953
0,213	219	0,239	0,842	0,882	0,809
0,215	220	0,254	0,816	0,806	0,822
0,218	223	0,224	0,882	0,828	0,927

A simulação que apresentou melhor desempenho no conjunto de validação em termos da precisão total obteve valor de 88,2% na caracterização referente a malignidade de microcalcificação mamária, com número de vetores de suporte igual a 223. No que tange a especificidade, classificação referente a não malignidade da microcalcificação mamária o melhor valor obtido foi 95,35% como maior número de vetor de suporte 230. Indicando que para alcançar esse valor o modelo apresentou a maior complexidade, visto que a mesma é função do número de vetor de suporte utilizada pelo modelo.

#### 4.2 Modelo BFM. Retirada da variável Densidade.

**Tabela 8:** Resultado das 50 simulações utilizando SVM com as variáveis: Birads, Margem e Forma. (BFM).

erro_treino	n_vetor	erro_vc	Precisao_Total	Sensibilidade	Especificidade
0,215	216	0,221	0,816	0,656	0,932
0,206	209	0,243	0,802	0,771	0,829
0,211	213	0,254	0,789	0,879	0,721
0,208	212	0,261	0,829	0,805	0,850
0,204	205	0,221	0,763	0,727	0,812
0,201	199	0,200	0,737	0,604	0,909
0,215	211	0,217	0,816	0,666	0,930
0,194	206	0,199	0,750	0,828	0,682
0,206	212	0,206	0,763	0,666	0,882
0,213	215	0,265	0,842	0,903	0,800
0,199	207	0,226	0,763	0,794	0,738
0,203	207	0,207	0,763	0,674	0,879
0,211	205	0,236	0,789	0,676	0,880
0,189	202	0,252	0,684	0,590	0,783
0,201	210	0,241	0,763	0,800	0,722
0,199	203	0,223	0,750	0,771	0,731
0,199	212	0,218	0,776	0,833	0,706
0,204	206	0,205	0,750	0,594	0,897
0,211	218	0,220	0,842	0,886	0,805
0,204	210	0,210	0,763	0,732	0,800
0,192	195	0,211	0,684	0,579	0,789
0,211	213	0,218	0,802	0,790	0,818
0,192	204	0,208	0,737	0,694	0,775
0,194	207	0,200	0,750	0,825	0,667
0,201	201	0,243	0,750	0,621	0,872
0,194	201	0,193	0,737	0,706	0,762
0,208	215	0,219	0,816	0,794	0,833
0,213	215	0,227	0,842	0,882	0,809
0,197	201	0,211	0,723	0,609	0,857
0,185	200	0,185	0,697	0,722	0,675
0,187	195	0,187	0,697	0,750	0,639
0,208	209	0,233	0,776	0,588	0,928
0,208	218	0,227	0,776	0,857	0,729
0,197	205	0,213	0,723	0,682	0,7812
0,197	203	0,195	0,750	0,839	0,689
0,206	204	0,225	0,763	0,714	0,804
0,211	216	0,253	0,816	0,737	0,895
0,206	208	0,2115	0,802	0,839	0,778
0,208	211	0,2325	0,802	0,725	0,889

0,199	206	0,248	0,776	0,838	0,718
0,204	209	0,2025	0,737	0,567	0,897
0,185	199	0,222	0,697	0,694	0,700
0,204	206	0,256	0,763	0,583	0,925
0,204	206	0,2425	0,776	0,763	0,789
0,197	205	0,2165	0,737	0,567	0,897
0,204	211	0,2045	0,776	0,880	0,647
0,197	208	0,216	0,750	0,842	0,658
0,197	210	0,208	0,750	0,778	0,734
0,199	208	0,21	0,763	0,812	0,727
0,201	208	0,214	0,763	0,871	0,689

Para o modelo que considerou as variáveis Birads, Forma e Margem (BFM) a simulação que apresentou maior valor na precisão total obteve 84,21%. Na análise da especificidade o modelo obteve o valor máximo de 93,18% com número de vetor de suporte 216. Fato que corrobora com a premissa que a retirada de uma variável no modelo implica na diminuição da complexidade do mesmo.

### 4.3 Modelo BFD. Retirada da variável Margem.

**Tabela 9:** Resultado das 50 simulações utilizando SVM com as variáveis: Birads, Forma e Densidade (BFD).

<b>erro_treino</b>	<b>n_vetor</b>	<b>erro_vc</b>	<b>Precisão_total</b>	<b>Sensibilidade</b>	<b>Especificidade</b>
0,192	196	0,206	0,711	0,553	0,868
0,194	199	0,207	0,750	0,564	0,946
0,194	194	0,206	0,737	0,556	0,900
0,197	197	0,207	0,750	0,605	0,939
0,197	200	0,206	0,750	0,568	0,923
0,197	199	0,202	0,750	0,585	0,943
0,197	203	0,216	0,750	0,588	0,881
0,199	201	0,209	0,763	0,643	0,912
0,199	206	0,209	0,776	0,647	0,881
0,199	198	0,253	0,776	0,667	0,875
0,199	197	0,218	0,750	0,675	0,833
0,199	200	0,206	0,763	0,659	0,906
0,199	199	0,214	0,750	0,568	0,923
0,201	204	0,208	0,776	0,618	0,905
0,201	204	0,207	0,776	0,743	0,805
0,201	196	0,210	0,750	0,556	0,925
0,201	201	0,215	0,776	0,622	0,923
0,201	203	0,218	0,776	0,618	0,905
0,201	201	0,212	0,776	0,714	0,829
0,201	204	0,208	0,776	0,568	0,974
0,201	203	0,284	0,737	0,816	0,658
0,204	198	0,201	0,763	0,629	0,878
0,204	203	0,209	0,776	0,677	0,844
0,204	207	0,225	0,803	0,730	0,872
0,204	197	0,211	0,776	0,583	0,950
0,206	202	0,212	0,776	0,682	0,906
0,206	205	0,216	0,789	0,674	0,939
0,206	208	0,219	0,789	0,675	0,917
0,206	207	0,218	0,803	0,714	0,912
0,206	208	0,217	0,803	0,732	0,886
0,206	207	0,224	0,750	0,821	0,708
0,206	204	0,273	0,803	0,688	0,886
0,208	212	0,226	0,816	0,742	0,867
0,208	206	0,245	0,803	0,725	0,889
0,208	203	0,219	0,789	0,684	0,895
0,208	204	0,227	0,803	0,667	0,878
0,208	210	0,292	0,763	0,871	0,689
0,211	209	0,224	0,829	0,656	0,955

0,211	212	0,212	0,816	0,684	0,947
0,211	211	0,224	0,829	0,675	1,000
0,211	213	0,226	0,842	0,786	0,912
0,213	213	0,226	0,842	0,686	0,976
0,213	212	0,229	0,842	0,667	0,977
0,213	207	0,221	0,816	0,750	0,875
0,213	212	0,220	0,829	0,636	0,977
0,213	208	0,225	0,816	0,791	0,848
0,213	213	0,223	0,842	0,710	0,933
0,218	216	0,221	0,855	0,676	1,000
0,220	219	0,227	0,882	0,771	0,976
0,222	213	0,227	0,882	0,824	0,929

Este modelo utilizou as variáveis Birads, Forma e Densidade, no que tange a especificidade este alcançou o melhor o resultado de 100%. Comparando os parâmetros com os demais modelos, notamos uma igualdade de valores.

#### 4.4 Modelo BMD. Retirada da variável Forma.

**Tabela 10:** Resultado das 50 simulações utilizando SVM com as variáveis: Birads, Margem e Densidade (BMD).

<b>erro_treino</b>	<b>n_vetor</b>	<b>erro_vc</b>	<b>Precisao_Total</b>	<b>Sensibilidade</b>	<b>Especificidade</b>
0,210	207	0,222	0,816	0,656	0,932
0,206	207	0,238	0,789	0,657	0,902
0,211	211	0,229	0,829	0,667	0,953
0,211	205	0,242	0,816	0,750	0,875
0,204	201	0,210	0,776	0,682	0,906
0,194	197	0,202	0,737	0,604	0,909
0,213	211	0,217	0,816	0,636	0,953
0,194	198	0,223	0,737	0,657	0,805
0,201	194	0,204	0,737	0,619	0,882
0,215	208	0,235	0,802	0,742	0,844
0,203	202	0,279	0,789	0,676	0,881
0,204	202	0,217	0,776	0,674	0,909
0,204	199	0,203	0,763	0,617	0,881
0,197	200	0,205	0,723	0,538	0,919
0,206	203	0,235	0,776	0,675	0,889
0,192	199	0,192	0,723	0,743	0,707
0,206	203	0,216	0,789	0,714	0,882
0,197	198	0,203	0,737	0,567	0,897
0,220	214	0,230	0,855	0,743	0,951
0,201	205	0,212	0,763	0,732	0,800
0,189	192	0,213	0,697	0,552	0,842
0,211	205	0,225	0,816	0,791	0,848
0,199	198	0,212	0,750	0,555	0,925
0,204	204	0,229	0,776	0,800	0,750
0,201	198	0,230	0,763	0,621	0,897
0,197	196	0,217	0,750	0,676	0,809
0,215	211	0,236	0,842	0,705	0,952
0,218	211	0,222	0,842	0,823	0,857
0,194	196	0,216	0,737	0,585	0,914
0,192	198	0,256	0,697	0,750	0,650
0,194	196	0,228	0,724	0,675	0,778
0,204	201	0,225	0,789	0,588	0,952
0,208	212	0,236	0,776	0,893	0,708
0,197	195	0,206	0,737	0,659	0,844
0,204	201	0,216	0,763	0,677	0,822
0,199	201	0,227	0,763	0,714	0,805
0,213	210	0,217	0,802	0,684	0,921
0,201	206	0,210	0,802	0,742	0,844



0,208	210	0,229	0,802	0,650	0,972
0,208	202	0,232	0,789	0,729	0,846
0,197	200	0,202	0,737	0,567	0,897
0,192	193	0,224	0,723	0,555	0,875
0,201	197	0,211	0,750	0,555	0,925
0,204	205	0,229	0,789	0,763	0,816
0,194	199	0,222	0,737	0,567	0,897
0,213	209	0,234	0,829	0,785	0,882
0,208	202	0,248	0,802	0,7367	0,868
0,192	200	0,206	0,684	0,741	0,653
0,204	204	0,226	0,789	0,719	0,841
0,213	210	0,268	0,829	0,839	0,822

Neste modelo ocorreu a retirada da variável Forma, observamos que o melhor desempenho foi no parâmetro especificidade chegando a ultrapassar os 95%. Comparando a quantidade de vetor de suporte com os demais modelos notamos uma igualdade, ou seja, a retirada da variável Forma não impactou na complexidade do mesmo.

#### 4.5 Modelo FDM. Retirada da variável Birads.

**Tabela 11:** Resultado das 50 simulações utilizando SVM com as variáveis: Margem, Forma e Densidade (FDM).

<b>erro_treino</b>	<b>n_vetor</b>	<b>erro_vc</b>	<b>Precisao_Total</b>	<b>Sensibilidade</b>	<b>Especificidade</b>
0,234	250	0,251	0,658	0,871	0,511
0,237	246	0,252	0,684	0,805	0,575
0,237	243	0,270	0,671	0,805	0,550
0,237	242	0,253	0,644	0,741	0,592
0,239	249	0,249	0,684	0,765	0,619
0,241	253	0,268	0,710	0,818	0,628
0,241	251	0,280	0,697	0,886	0,536
0,241	254	0,260	0,697	0,853	0,571
0,241	253	0,261	0,710	0,846	0,567
0,241	253	0,273	0,684	0,778	0,600
0,241	249	0,293	0,710	0,868	0,552
0,244	253	0,260	0,710	0,828	0,609
0,244	254	0,275	0,710	0,853	0,543
0,244	249	0,280	0,724	0,805	0,650
0,244	255	0,262	0,724	0,875	0,556
0,244	250	0,271	0,684	0,928	0,542
0,246	250	0,279	0,737	0,861	0,625
0,246	254	0,260	0,710	0,882	0,571
0,246	256	0,267	0,684	0,794	0,595
0,246	250	0,262	0,697	0,727	0,656
0,248	257	0,279	0,750	0,784	0,718
0,248	257	0,282	0,750	0,829	0,657
0,248	255	0,264	0,737	0,850	0,611
0,248	256	0,265	0,723	0,839	0,644
0,248	251	0,277	0,710	0,892	0,538
0,248	252	0,275	0,724	0,895	0,552
0,248	251	0,283	0,724	0,784	0,667
0,251	255	0,278	0,724	0,814	0,606
0,251	256	0,273	0,750	0,875	0,611
0,251	255	0,291	0,724	0,786	0,647
0,251	259	0,272	0,763	0,828	0,707
0,251	259	0,278	0,763	0,886	0,658
0,253	259	0,261	0,750	0,882	0,643
0,253	257	0,293	0,710	0,814	0,576
0,253	260	0,275	0,763	0,875	0,639
0,255	262	0,299	0,763	0,800	0,732
0,255	261	0,278	0,763	0,841	0,656
0,255	262	0,285	0,724	0,763	0,684

0,255	253	0,294	0,750	0,906	0,636
0,258	262	0,288	0,776	0,857	0,676
0,258	267	0,272	0,763	0,784	0,743
0,258	263	0,287	0,802	1,000	0,667
0,258	259	0,302	0,776	0,881	0,647
0,260	265	0,272	0,789	0,818	0,767
0,260	261	0,306	0,750	0,968	0,600
0,260	259	0,304	0,763	0,837	0,667
0,262	266	0,285	0,802	0,789	0,816
0,262	268	0,302	0,789	0,729	0,847
0,265	265	0,294	0,818	0,852	0,786
0,272	271	0,309	0,855	0,875	0,841

O modelo que incorporou as variáveis Forma, Densidade e Margem (FDM) obteve o melhor resultado no parâmetro sensibilidade (100%), ou seja na classificação positiva no que tange a malignidade das microcalcificações mamárias. Na análise da quantidade de do número de variáveis utilizada (n=3) apresentou valores maiores que os demais modelos, indicando que a retirada da variável Birads aumentou a complexidade do modelo.

Nas tabelas (12,13,14,15 e 16) apresentamos a comparação entre os resultados médio obtidos nas 50 simulações dos modelos (BFDM, BFM, BFD, BMD e MFD).

**Tabela 12:** Resultado das médias de 50 simulações utilizando o modelo com as variáveis: Birads, Margem, Forma e Densidade (BMFD).

	ACC (%)	S (%)	E (%)	Erro Treino	Erro VC
[Média ±sd]	79,10±0,04	71,70±0,09	84,40±0,08	0,203±0,01	0,236±0,02
[Mediana±irq]	78,90±0,050	72,6±0,14	85,7±0,08	0,201±0,07	0,237±0,02
[Max-Min]	88,20-73,70	88,2-56,8	95,3-70,5	0,218-19,7	0,269-20,5
<b>Melhor simulação</b>	88,16	82,29	92,68	20,30	20,10

**Tabela 13:** Resultado das médias de 50 simulações utilizando o modelo com as variáveis: Birads, Forma e Densidade (BFD).

	ACC (%)	S (%)	E (%)	Erro Treino	Erro VC
[Média ±sd]	80,2±0,035	69,9±0,073	89,6±0,079	0,205±0,005	0,225±0,019
[Mediana±irq]	80,3±0,053	68,4±0,081	90,6±0,076	0,206±0,008	0,221±0,012
[Max-Min]	88,2-75,0	82,4-56,8	100-68,9	0,222-0,201	0,29-0,201
<b>Melhor simulação</b>	88,20	77,10	97,60	22,00	22,70

**Tabela 14:** Resultado das médias de 50 simulações utilizando o modelo com as variáveis: Forma, Densidade e Margem (FDM).

	ACC (%)	S (%)	E (%)	Erro Treino	Erro VC
[Média ±sd]	75±0,036	82,8±0,059	65,7±0,077	0,254±0,006	0,282±0,014
[Mediana±irq]	75±0,043	84±0,088	65,7±0,099	0,253±0,009	0,279±0,021
[Max-Min]	85,5-68,4	100-72,7	84,6-53,8	0,272-0,244	0,272-0,244
<b>Melhor simulação</b>	85,52	87,50	85,52	27,16	30,90

**Tabela 15:** Resultado das médias de 50 simulações utilizando o modelo com as variáveis: Birads, Densidade e Margem (BDM).

	ACC(%)	S (%)	E(%)	Erro Treino	Erro VC
[Média ±sd]	79,3±0,029	70,33±0,077	87,72±0,059	0,207±0,005	0,277±0,016
[Mediana±irq]	78,9±0,039	70,59±0,087	88,24±0,081	0,206±0,007	0,0227±0,019
[Max-Min]	85,5-73,7	89,3-55,5	95,5-7,8	0,220-0,199	0,268±0,203
<b>Melhor simulação</b>	85,52	74,30	95,12	22,01	23,00

**Tabela 16:** Resultado das médias de 50 simulações utilizando o modelo com as variáveis: Birads, Forma e Margem (BFM).

	ACC (%)	S (%)	E (%)	Erro Treino	Erro VC
[Média ±sd]	78,35±0,029	75,16±0,100	81,65±0,080	0,202±0,008	0,221±0,019
[Mediana±irq]	77,63±0,039	77,14±0,17	81,25±0,16	0,206±0,009	0,227±0,031
[Max-Min]	84,21-73,68	90,32-56,76	93,18-68,89	0,2155-0,185	0,2655-0,1855
<b>Melhor simulação</b>	84,21	90,32	80,0	21,31	26,55

Pode-se observar nas tabelas acima que os modelos que incorporou as variáveis Birads (B) e Densidade (D) apresentou melhor desempenho na classificação de microcalcificações benignas. Indicando que esses parâmetros podem ser bons marcadores para esse tipo de classificação. Entretanto a incorporação das variáveis Forma (F) e Margem (M) possibilitou um melhor desempenho na classificação da sensibilidade (malignidade da lesão).

No que tange a melhor simulação, o modelo completo com a incorporação das variáveis Birads, Forma, Margem e Densidade obteve o melhor resultado, sinalizando a importância desses parâmetros na determinação da malignidade da microcalcificação.

## CAPITULO V

### 5.1. Conclusões

A elevada taxa de incidência e mortes causadas pelo câncer de mama, atualmente, no Brasil e no mundo, justifica o desenvolvimento de pesquisas científicas voltadas para estratégias de auxílio na detecção precoce da doença, fator determinante para o sucesso do tratamento. Dentro deste contexto, o presente trabalho teve como objetivo principal aplicar e analisar o modelo SVM na predição de neoplasias mamárias.

De acordo com a análise dos resultados foi possível evidenciar o desempenho promissor da SVM, que obteve uma precisão total acima de 88% , sensibilidade entre 72,7% a 100% e especificidade ente 65,7 a 100%.

Contudo, este modelo é bastante sensível aos parâmetros de entrada o que demanda de um longo período de ajuste o que dificulta a criação do mesmo.

A acurácia obtida pelo modelo na classificação de microcalcificação mamária, encontra-se próxima aos valores obtidos na literatura com a utilização de técnicas baseada em inteligência computacional (SHANTHV,BHASKARAN, 2013).

A incorporação da variáveis forma e margem no modelo possibilitou um desempenho superior , principalmente ao nível de sensibilidade, onde o valor médio obtido no conjunto das simulações realizadas foi maior que 75,00%.

Houve diferença estatística significativa a nível de 1% (p-valor <0,001) no desempenho do modelo SVM no parâmetro sensibilidade entre os FDM e BMFD, FDM e BFD, FDM e BDM. Indicando que as variáveis forma e margem tem grande impacto na classificação da microcalcificação maligna.

Não houve diferença significativa no parâmetro sensibilidade a nível de 1% (p-valor < 0,001) entre os modelos BMFD e BFD.

Apesar dos bons resultados obtidos com a aplicação do modelo SVM na classificação de microcalcificações mamárias com dados de mamografias, há necessidade de aumentar o número de amostras.

Entretanto, apesar do modelo estruturado em SVM não ter alcançado em todos casos 100% de precisão em suas predições, a mesmo obteve resultados satisfatórios. Por isso, pode-se concluir que a SVM é ferramenta útil na previsão de microcalcificações mamárias.

### 5.2. Trabalhos Futuros

Como trabalhos futuros, pretende-se verificar: O desempenho do modelo baseado em SVM incorporando a dimensão fractal nos dados referentes a forma da lesão; Aplicação da técnica de fuzzificação na codificação dos dados; Aplicação de outras ferramentas baseadas em inteligência computacional de sorte a comparar os resultados com o modelo proposto.

## 6. REFERÊNCIAS BIBLIOGRÁFICAS

ABCMED, 2013. Biópsia. Disponível em: <<http://www.abc.med.br/p/exames-e-procedimentos/513419/biopsia-o-que-e-quando-e-indicada-quais-os-tipos-quais-as-possiveis-complicacoes-como-ela-e-realizada.htm>>.

Aguillar VLN, Bauab SP. Rastreamento mamográfico para detecção precoce do câncer de mama. Revista brasileira de mastologia 2003.

Andre, T.; Rangayyan, R. M. Classification of tumors and masses in mammograms using neural networks with shape and texture features. 25th Annual International Conference of the IEEE Transactions on Engineering in Medicine and Biology Society.

Almeida F. F. M. (2007). Relatório técnico: Support Vector Machine. Universidade Federal de Campina Grande, Centro de Ciência e Tecnologia.

Araújo, S. D. T. de. Mortalidade por câncer de mama, com idade igual ou superior a 50 anos, estado de São Paulo 1979-1997. Tese de mestrado apresentada à Universidade de São Paulo, Dezembro 2000.

Braz Júnior, G.; Silva, E. C.; Paiva, A. C.; Silva, A. C.; Gattass, M. Breast Tissues Mammograms Images Classification using Moran's Index, Geary's Coefficient and SVM . In: International Conference on Neural Information Processing, 2007, Kitakyushu. Lecture Notes in Computer Science, LNCS, 2007.

Campos, L. F. A., Silva, A. C., Barros, A. K., 2007. Independent Component Analysis and Neural Networks Applied for Classification of Malignant, Benign and Normal Tissue in Digital Mammography . Methods of Information in Medicine, 2007.

Chang, C. E Lin, C., 2003. Libsvm – A Library for Support Vector Machines.

Costa, D. D., Barros, A. K., E Silva, A. C., 2007. Independent Component Analysis in Breast Tissues Mammograms Images Classification using LDA and SVM . Information Technology Applications in Biomedicine - ITAB2007 - Tokyo. Conference on 6th International Special Topic .

Cristianini, N. E Shawe-Taylor, J. , 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods . Cambridge University Press.

C. J. C. Burges. A tutorial on support vector machines for pattern recognition. Knowledge Discovery and Data Mining, 2(2):1–43, 1998.

Chaves, A. C. F., 2006. Extração de Regras Fuzzy para Máquinas de Vetor de Suporte (SVM) para Classificação em Múltiplas Classes . PhD Thesis. Pontifícia Universidade Católica do Rio de Janeiro.

DDSM, 2001. The Digital Database for Screening Mammography , Michael Heath, Kevin

Bowyer, Daniel Kopans, Richard Moore and W. Philip Kegelmeyer, in Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed., Medical Physics Publishing. Disponível em <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.

Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D'orsi, C., Berns, E. A., Cutter, G., Hendrick, R. E., Barlow, W. E., ELMORE, J. G., 2007. Influence of Computer-Aided Detection on Performance of Screening Mammography . Breast Diseases: A Year Book Quarterly.

Ferrari, R. J.; Rangayan, R. M., Desautels, J. E. L.; Freire, A. F. Analysis of asymmetry in mammograms via directional filtering with gaborwavelets. IEEE Transactions on Medical Imaging, v. 20 , September 2001.

Freer, T. W. E Ulissey, M. J., 2001. Screening Mammography with Computer-Aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center . Radiology.

Giger, M. L., 2000. Computer-aided diagnosis of breast lesions in medical images. Computing in Science & Engineering.

Goldberd, D., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. EUA: Addison-Wesley.

Gonçalves A. R. Máquinas de Vetores suporte, 2010.

Gonzalez, R. C. E Woods, R. E., 2007. Digital Image Processing . 3rd Edition. Prentice Hal

Gunn S. R. (1998). Support Vector Machine for Classification and Regression. Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science.

Halkiotis, S.; Mantas, J.; Botsis, T. Computer-aided detection of clustered microcalcifications in digital mammograms. Special Issue: Proceedings of the fifth European Systems Science Congress, V. 2, October 2002.

H. Nassif, D. Page, M. Ayvaci, J. Shavlik, and E. S. Burnside. Uncovering age-specific invasive and dcis breast cancer rules using inductive logic programming . Proceedings of 2010 ACM International Health Informatics Symposium (IHI 2010), ACM Digital Library, 2010.

Hansel, D . Fundamentos da Patologia, Câncer de Mama, pág. 553 - 557. Rio de Janeiro Guanabara Koogan, 2007.

Haykin, S. e Engel, P. M., 2001. Redes Neurais: Principios e Pratica . Bookman.

Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, W.P., 1998. Current Status of the Digital Database for Screening Mammography. Digital Mammography . iCad (2008). iCad Solutions. Disponível em: <http://www.icadmed.com/>.



Hermann, G.; Janus, C.; Schwartz, I S.; Krivisky, B.; Bier, S.; Rabinowitz, J.G. Nonpalpable breast lesions: Accuracy of prebiopsy mammographic diagnosis, 1987.

INCA, 2015. Instituto Nacional do Câncer. Estimativas 2008: Incidência de Câncer no Brasil . Available at <http://www.inca.gov.br>.

Instituto Nacional de Câncer (Brasil). Estimativas 2010. Incidência de câncer no Brasil. Rio de Janeiro: INCA; 2009.

INTEL, 2008. Opencv, Open Computer Vision Library . Intel Technology and Research. Disponível em <http://sourceforge.net/projects/opencvlibrary/>.

J. A. Baranauskas and M. C. Monard. Reviewing some machine learning concepts and methods. Technical Report 102, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos. Disponível em: [ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/rel\\_tec/RT\\_102.ps.zip](ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/rel_tec/RT_102.ps.zip), Fevereiro 2000.

Jacobson, H. G.; Edeiken, J. Biopsy of occult breast lesions: Analysis of 1261 abnormalities. J. Am. Math. Assoc., 1990.

J. L. Castro, L. D. Flores-Hidalgo, C. J. Mantas, and J. M. Puche. Extraction of fuzzy rules from support vector machines. Fuzzy Sets and Systems archive, 2007.

Jain, A. K., Duin, R. P. W., Mao , J., 2000. Statistical pattern recognition: A review . IEEE Transactions on Pattern Analysis e Machine Intelligence.

Karatzoglou, A., Smola, A., Hornik, K., & Karatzoglou, M. A. (2016). Package ‘kernlab’. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Kestelman FP, Sousa GA, Thuler LC, Martins G, Freitas VAR, Canella EO. Breast Imaging Reporting and Data System - BI-RADS®: valor preditivo positivo das Categorias 3, 4 e 5. Revisão sistemática da literatura. Radiologia brasileira 2007.

Kopans, DB. Imagem da mama. 2ª ed. Rio de Janeiro: Revinter Medsi; 2000.

Liberman L, Abramson AF, Squires FB, Glassman JR, Morris EA, Dershaw DD. The Breast imaging report and data system: positive predict value of mammographic features and final assement categories. AJR 1998.

Lorena A. C. e Carvalho, A. C. P. L. F. (2007) Uma Introdução às Support Vector Machines.

M. A. Hearst, B. Schölkopf, S. Dumais, E. Osuna, and J. Platt. Trends and controversies - support vector machines. IEEE Intelligent System , 1998.

M. C. Monard and J. A. Baranauskas. Conceitos de aprendizado de máquina. In S. O. Rezende, editor, Sistemas Inteligentes - Fundamentos e Aplicações. Editora Manole, 2003.

M. C. P. Souto, A. C. Lorena, A. C. B. Delbem, and A. C. P. L. F. Carvalho. Técnicas de Aprendizado de Máquina para problemas de Biologia Molecular, pages 103–152. Minicursos de Inteligência Artificial, Jornada de Atualização Científica em Inteligência Artificial, XXIII Congresso da Sociedade Brasileira de Computação, 2003.

Martins, L. O., 2007. Detecção de Massas em Imagens Mamográficas Através do Algoritmo Growing Neural Gas e da Função K De Ripley. Dissertação de mestrado. Universidade Federal do Maranhão, Departamento de Engenharia de Eletricidade, Programa de Pós-Graduação em Engenharia de Eletricidade. São Luís, 2007.

Ministério da Saúde, 2002 . Falando Sobre Câncer de Mama. Instituto Nacional de Câncer, Coordenação de Prevenção e Vigilância. Rio de Janeiro.

Ministério da Saúde (BR). Política Nacional de Câncer de Mama. Brasília: Ministério da Saúde; 2008.

Moayed, F., Boostani, R. Azimifar, Z. Katebi, S., 2007. A Support Vector Based Fuzzy Neural Network Approach for Mass Classification in 81 Mammography . Digital Signal Processing, 2007. 15th International Conference.

Mudigonda, N. R; Rangayyan, R. M; Desautels, J. E. L. Detection of the Breast Masses in Mammograms by Density Slicing and Texture Flow-Field Analysis. **IEEE** Transaction on Medical Imaging, 2001.

Obenauer, S.; Hermann, K. P.; Grabbe, E. Applicatios and and literature review of the BI-RADS classification. European Radiology, 2005.

Osta, H., Qahwaji, R. E Ipson, S., 2008. Wavelet-based Feature Extraction and Classification for Mammogram Images using RBF and SVM. In Proceedings of International Conference on Visualization, Imaging, and Image Processing (VIIP). Palma de Mallorca, Spain.

Padwal, M., 2007. Elements of breasting imaging basics . Disponível em: [http://www.gehealthcare.com/usen/ultrasound/education/products/cme\\_breast.html](http://www.gehealthcare.com/usen/ultrasound/education/products/cme_breast.html).

Parkin DM, Bray FI, Devesa SS. Cancer burden in the year 2000. The global picture. Eur J Cancer. 2001 Oct.

Queiroz RY, Hummel AC. O valor das microcalcificações agrupadas no diagnóstico precoce do câncer de mama. Revista brasileira de cancerologia 1995.

R. Herbrich. Learning Kernel Classifiers: Theory and Algorithms. MIT Press, 2001.

R. Woods, L. Oliphant, K. Shinki, D. Page, J. Shavlik, and E. Burnside. Validation of results from knowledge discovery: Mass density as a predictor of breast cancer . J Digit Imaging, 2009.

Ryan Woods and Elizabeth Burnside. The mammographic density of a mass is a significant predictor of breast cancer . Radiology, USA, 2010.

S. Haykin. Neural Networks - A Comprehensive Foundation. Prentice-Hall, New Jersey, 2nd edition, 1999.

Sampat, M. P., Markey, M. K., Bovik, A. C., 2005. Computer-Aided Detection and Diagnosis in Mammography . Handbook of Image and Video Processing.

Santos, V. T. Segmentação de imagens mamográficas para detecção de nódulos em mamas densas. Tese de mestrado apresentada na Universidade de São Paulo, novembro de 2002.

Shanthu S, Bhaskaran AV. A Novel Approach for detecting and Classifying Breast Cancer. In: International Journal of Intelligent Information Technologies, 2013.

Shen, L.; Rangayyan, R. M.; Desautels, J. E. L.; Frere, A. F. Application of shape analysis to mammographic calcifications . IEEE Transactions on Medical Imaging v. 13.

Sickles EA. Breast calcifications: mammographic evaluation. Radiology 1986.

Sousa, J. R., Silva, A. C., Paiva, A. C., 2007. Lung Structures Classification Using 3D Geometric Measurements and SVM . In: 12th Iberoamerican Congress on Pattern Recognition - CIARP 2007, Valparaiso. Lecture Notes Computer Science - LNCS. Berlin: Springer-Verla.

T. Mitchell. Machine Learning. McGraw Hill, 1997.

T. Ayer, O. Alagoz, J. Chhatwal, J. W. Shavlik, C. E. J. Kahn, and E. S. Burnside. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration . Vol. Cancer. 2010.

Tse GM, Tan PH, Cheung HS, Chu WCW, Lam WWM. Intermediate to highly suspicious calcification in breast lesions: a radio-pathologic correlation. Breast Cancer Res Treat 2008.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications, 1971.

V. N. Vapnik. Statistical Learning Theory. John Wiley and Sons, 1998.

Vianna AD, Marchiori E. Calcificações malignas da mama: correlação mamografia anatomia patológica. Radiologia brasileira 2002.

Zhang, P. e Kumar, K., 2006. Analyzing Feature Significance from Various Systems for Mass Diagnosis . Proceedings of the International Conference 83 on Computational Intelligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce.

Zurrida, S. et al. A dissecação axilar no carcinoma de mama. In: VERONESI, U. et al. Mastologia Oncológica . Rio de Janeiro: MEDSI, 2002.

Yu, S.; Guan, L. A cad system for the automatic detection of clustered microcalcifications in digitized mammogram films. IEEE Transaction on the Medical Imaging , v 20.

## ANEXO

**Tabela 17:** Banco de dados coletados no Instituto de Radiologia da Universidade Erlangen-Nuremberg, entre 2003 e 2006.

Birads	Idade	Forma	Margem	Densidade	Diagnóstico
5	67	3	5	3	1
4	43	1	1	?	1
5	58	4	5	3	1
4	28	1	1	3	0
5	74	1	5	?	1
4	65	1	?	3	0
4	70	?	?	3	0
5	42	1	?	3	0
5	57	1	5	3	1
5	60	?	5	1	1
5	76	1	4	3	1
3	42	2	1	3	1
4	64	1	?	3	0
4	36	3	1	2	0
4	60	2	1	2	0
4	54	1	1	3	0
3	52	3	4	3	0
4	59	2	1	3	1
4	54	1	1	3	1
4	40	1	?	?	0
?	66	?	?	1	1
5	56	4	3	1	1
4	43	1	?	?	0
5	42	4	4	3	1
4	59	2	4	3	1
5	75	4	5	3	1
2	66	1	1	?	0

5	63	3	?	3	0
5	45	4	5	3	1
5	55	4	4	3	0
4	46	1	5	2	0
5	54	4	4	3	1
5	57	4	4	3	1
4	39	1	1	2	0
4	81	1	1	3	0
4	77	3	?	?	0
4	60	2	1	3	0
5	67	3	4	2	1
4	48	4	5	?	1
4	55	3	4	2	0
4	59	2	1	?	0
4	78	1	1	1	0
4	50	1	1	3	0
4	61	2	1	?	0
5	62	3	5	2	1
5	44	2	4	?	1
5	64	4	5	3	1
4	23	1	1	?	0
2	42	?	?	4	0
5	67	4	5	3	1
4	74	2	1	2	0
5	80	3	5	3	1
4	23	1	1	?	0
4	63	2	1	?	0
4	53	?	5	3	1
4	43	3	4	?	0
4	49	2	1	1	0
5	51	2	4	?	0
4	45	2	1	?	0
5	59	2	?	?	1
5	52	4	3	3	1
5	60	4	3	3	1
4	57	2	5	3	0
3	57	2	1	?	0
5	74	4	4	3	1
4	25	2	1	?	0
4	49	1	1	3	0
5	72	4	3	?	1
4	45	2	1	3	0
4	64	2	1	3	0
4	73	2	1	2	0

5	68	4	3	3	1
5	52	4	5	3	0
5	66	4	4	3	1
5	70	?	4	?	1
4	25	1	1	3	0
5	74	1	1	2	1
4	64	1	1	3	0
5	60	4	3	2	1
5	67	2	4	1	0
4	67	4	5	3	0
5	44	4	4	2	1
3	68	1	1	3	1
4	57	?	4	1	0
5	51	4	?	?	1
4	33	1	?	?	0
5	58	4	4	3	1
5	36	1	?	?	0
4	63	1	1	?	0
5	62	1	5	3	1
4	73	3	4	3	1
4	80	4	4	3	1
4	67	1	1	?	0
5	59	2	1	3	1
5	60	1	?	3	0
5	54	4	4	3	1
4	40	1	1	?	0
4	47	2	1	?	0
5	62	4	4	3	0
4	33	2	1	3	0
5	59	2	?	?	0
4	65	2	?	?	0
4	58	4	4	?	0
4	29	2	?	?	0
4	58	1	1	?	0
4	54	1	1	?	0
4	44	1	1	?	1
3	34	2	1	?	0
4	57	1	1	3	0
5	33	4	4	?	1
4	45	4	4	3	0
5	71	4	4	3	1
5	59	4	4	2	0
4	56	2	1	?	0
4	40	3	4	?	0

4	56	1	1	3	0
4	45	2	1	?	0
4	57	2	1	2	0
5	55	3	4	3	1
5	84	4	5	3	0
5	51	4	4	3	1
4	43	1	1	?	0
4	24	2	1	2	0
4	66	1	1	3	0
5	33	4	4	3	0
4	59	4	3	2	0
4	76	2	3	?	0
4	40	1	1	?	0
4	52	?	4	?	0
5	40	4	5	3	1
5	67	4	4	3	1
5	75	4	3	3	1
5	86	4	4	3	0
4	60	2	?	?	0
5	66	4	4	3	1
5	46	4	5	3	1
4	59	4	4	3	1
5	65	4	4	3	1
4	53	1	1	3	0
5	67	3	5	3	1
5	80	4	5	3	1
4	55	2	1	3	0
4	48	1	1	?	0
4	47	1	1	2	0
4	50	2	1	?	0
5	62	4	5	3	1
5	63	4	4	3	1
4	63	4	?	3	1
4	71	4	4	3	1
4	41	1	1	3	0
5	57	4	4	4	1
5	71	4	4	4	1
4	66	1	1	3	0
4	47	2	4	2	0
3	34	4	4	3	0
4	59	3	4	3	0
5	55	2	?	?	1
4	51	?	?	3	0
4	62	2	1	?	0

4	58	4	?	3	1
5	67	4	4	3	1
4	41	2	1	3	0
4	23	3	1	3	0
4	53	?	4	3	0
4	42	2	1	3	0
5	87	4	5	3	1
4	68	1	1	3	1
4	64	1	1	3	0
5	54	3	5	3	1
5	86	4	5	3	1
4	21	2	1	3	0
4	39	1	1	?	0
4	53	4	4	3	0
4	44	4	4	3	0
4	54	1	1	3	0
5	63	4	5	3	1
4	62	2	1	?	0
4	45	2	1	2	0
5	71	4	5	3	0
5	49	4	4	3	1
4	49	4	4	3	0
5	66	4	4	4	0
4	19	1	1	3	0
4	35	1	1	2	0
4	71	3	3	?	1
5	74	4	5	3	1
5	37	4	4	3	1
4	67	1	?	3	0
5	81	3	4	3	1
5	59	4	4	3	1
4	34	1	1	3	0
5	79	4	3	3	1
5	60	3	1	3	0
4	41	1	1	3	1
4	50	1	1	3	0
5	85	4	4	3	1
4	46	1	1	3	0
5	66	4	4	3	1
4	73	3	1	2	0
4	55	1	1	3	0
4	49	2	1	3	0
3	49	4	4	3	0
4	51	4	5	3	1



2	48	4	4	3	0
4	58	4	5	3	0
5	72	4	5	3	1
4	46	2	3	3	0
4	43	4	3	3	1
?	52	4	4	3	0
4	66	2	1	?	0
4	46	1	1	1	0
4	69	3	1	3	0
2	59	1	1	?	1
5	43	2	1	3	1
5	76	4	5	3	1
4	46	1	1	3	0
4	59	2	4	3	0
4	57	1	1	3	0
5	43	4	5	?	0
3	45	2	1	3	0
3	43	2	1	3	0
4	45	2	1	3	0
5	57	4	5	3	1
5	79	4	4	3	1
5	54	2	1	3	1
4	40	3	4	3	0
5	63	4	4	3	1
2	55	1	?	1	0
4	52	2	1	3	0
4	38	1	1	3	0
3	72	4	3	3	0
5	80	4	3	3	1
5	76	4	3	3	1
4	62	3	1	3	0
5	64	4	5	3	1
5	42	4	5	3	0
3	60	?	3	1	0
4	64	4	5	3	0
4	63	4	4	3	1
4	24	2	1	2	0
5	72	4	4	3	1
4	63	2	1	3	0
4	46	1	1	3	0
3	33	1	1	3	0
5	76	4	4	3	1
4	36	2	3	3	0
4	40	2	1	3	0

5	58	1	5	3	1
4	43	2	1	3	0
3	42	1	1	3	0
4	32	1	1	3	0
5	57	4	4	2	1
4	37	1	1	3	0
4	70	4	4	3	1
5	56	4	2	3	1
3	76	?	3	2	0
5	73	4	4	3	1
5	77	4	5	3	1
5	67	4	4	1	1
5	71	4	3	3	1
5	65	4	4	3	1
4	43	1	1	3	0
4	40	2	1	?	0
4	49	2	1	3	0
5	76	4	2	3	1
4	55	4	4	3	0
5	72	4	5	3	1
3	53	4	3	3	0
5	75	4	4	3	1
5	61	4	5	3	1
5	67	4	4	3	1
5	55	4	2	3	1
5	66	4	4	3	1
2	76	1	1	2	0
4	57	4	4	3	1
5	71	3	1	3	0
5	70	4	5	3	1
4	35	4	2	?	0
5	79	1	?	3	1
4	63	2	1	3	0
5	40	1	4	3	1
4	41	1	1	3	0
4	47	2	1	2	0
4	68	1	1	3	1
4	64	4	3	3	1
4	65	4	4	?	1
4	73	4	3	3	0
4	39	4	3	3	0
5	55	4	5	4	1
5	53	3	4	4	0
5	66	4	4	3	1

4	43	3	1	2	0
5	44	4	5	3	1
4	77	4	4	3	1
4	62	2	4	3	0
5	80	4	4	3	1
4	33	4	4	3	0
4	50	4	5	3	1
4	71	1	?	3	0
5	46	4	4	3	1
5	49	4	5	3	1
4	53	1	1	3	0
3	46	2	1	2	0
4	57	1	1	3	0
4	54	3	1	3	0
4	54	1	?	?	0
2	49	2	1	2	0
4	47	3	1	3	0
4	40	1	1	3	0
4	45	1	1	3	0
4	50	4	5	3	1
5	54	4	4	3	1
4	67	4	1	3	1
4	77	4	4	3	1
4	66	4	3	3	0
4	71	2	?	3	1
4	36	2	3	3	0
4	69	4	4	3	0
4	48	1	1	3	0
4	64	4	4	3	1
4	71	4	2	3	1
5	60	4	3	3	1
4	24	1	1	3	0
5	34	4	5	2	1
4	79	1	1	2	0
4	45	1	1	3	0
4	37	2	1	2	0
4	42	1	1	2	0
4	72	4	4	3	1
5	60	4	5	3	1
5	85	3	5	3	1
4	51	1	1	3	0
5	54	4	5	3	1
5	55	4	3	3	1
4	64	4	4	3	0

5	67	4	5	3	1
5	75	4	3	3	1
5	87	4	4	3	1
4	46	4	4	3	1
4	59	2	1	?	0
55	46	4	3	3	1
5	61	1	1	3	1
4	44	1	4	3	0
4	32	1	1	3	0
4	62	1	1	3	0
5	59	4	5	3	1
4	61	4	1	3	0
5	78	4	4	3	1
5	42	4	5	3	0
4	45	1	2	3	0
5	34	2	1	3	1
5	39	4	3	?	1
4	27	3	1	3	0
4	43	1	1	3	0
5	83	4	4	3	1
4	36	2	1	3	0
4	37	2	1	3	0
4	56	3	1	3	1
5	55	4	4	3	1
5	46	3	?	3	0
4	88	4	4	3	1
5	71	4	4	3	1
4	41	2	1	3	0
5	49	4	4	3	1
3	51	1	1	4	0
4	39	1	3	3	0
4	46	2	1	3	0
5	52	4	4	3	1
5	58	4	4	3	1
4	67	4	5	3	1
5	80	4	4	3	1
3	46	1	?	?	0
3	43	1	?	?	0
4	45	1	1	3	0
5	68	4	4	3	1
4	54	4	4	?	1
4	44	2	3	3	0
5	74	4	3	3	1
5	55	4	5	3	0

4	49	4	4	3	1
4	49	1	1	3	0
5	50	4	3	3	1
5	52	3	5	3	1
4	45	1	1	3	0
4	66	1	1	3	0
4	68	4	4	3	1
4	72	2	1	3	0
5	64	?	?	3	0
2	49	?	3	3	0
3	44	?	4	3	0
5	74	4	4	3	1
5	58	4	4	3	1
4	77	2	3	3	0
4	49	3	1	3	0
4	34	?	?	4	0
5	60	4	3	3	1
5	69	4	3	3	1
4	53	2	1	3	0
3	46	3	4	3	0
5	74	4	4	3	1
4	58	1	1	3	0
5	68	4	4	3	1
5	46	4	3	3	0
5	61	2	4	3	1
5	70	4	3	3	1
5	37	4	4	3	1
3	65	4	5	3	1
4	67	4	4	3	0
5	69	3	4	3	0
5	76	4	4	3	1
4	65	4	3	3	0
5	72	4	2	3	1
4	62	4	2	3	0
5	42	4	4	3	1
5	66	4	3	3	1
5	48	4	4	3	1
4	35	1	1	3	0
5	60	4	4	3	1
5	67	4	2	3	1
5	78	4	4	3	1
4	66	1	1	3	1
4	26	1	1	?	0
4	48	1	1	3	0

4	31	1	1	3	0
5	43	4	3	3	1
5	72	2	4	3	0
5	66	1	1	3	1
4	56	4	4	3	0
5	58	4	5	3	1
5	33	2	4	3	1
4	37	1	1	3	0
5	36	4	3	3	1
4	39	2	3	3	0
4	39	4	4	3	1
5	83	4	4	3	1
4	68	4	5	3	1
5	63	3	4	3	1
5	78	4	4	3	1
4	38	2	3	3	0
5	46	4	3	3	1
5	60	4	4	3	1
5	56	2	3	3	1
4	33	1	1	3	0
4	?	4	5	3	1
4	69	1	5	3	1
5	66	1	4	3	1
4	72	1	3	3	0
4	29	1	1	3	0
5	54	4	5	3	1
5	80	4	4	3	1
5	68	4	3	3	1
4	35	2	1	3	0
4	57	3	?	3	0
5	?	4	4	3	1
4	50	1	1	3	0
4	32	4	3	3	0
0	69	4	5	3	1
4	71	4	5	3	1
5	87	4	5	3	1
3	40	2	?	3	0
4	31	1	1	?	0
4	64	1	1	3	0
5	55	4	5	3	1
4	18	1	1	3	0
3	50	2	1	?	0
4	53	1	1	3	0
5	84	4	5	3	1

5	80	4	3	3	1
4	32	1	1	3	0
5	77	3	4	3	1
4	38	1	1	3	0
5	54	4	5	3	1
4	63	1	1	3	0
4	61	1	1	3	0
4	52	1	1	3	0
4	36	1	1	3	0
4	41	?	?	3	0
4	59	1	1	3	0
5	51	4	4	2	1
4	36	1	1	3	0
5	40	4	3	3	1
4	49	1	1	3	0
4	37	2	3	3	0
4	46	1	1	3	0
4	63	1	1	3	0
4	28	2	1	3	0
4	47	2	1	3	0
4	42	2	1	3	1
5	44	4	5	3	1
4	49	4	4	3	0
5	47	4	5	3	1
5	52	4	5	3	1
4	53	1	1	3	1
5	83	3	3	3	1
4	50	4	4	?	1
5	63	4	4	3	1
4	82	?	5	3	1
4	54	1	1	3	0
4	50	4	4	3	0
5	80	4	5	3	1
5	45	2	4	3	0
5	59	4	4	?	1
4	28	2	1	3	0
4	31	1	1	3	0
4	41	2	1	3	0
4	21	3	1	3	0
5	44	3	4	3	1
5	49	4	4	3	1
5	71	4	5	3	1
5	75	4	5	3	1
4	38	2	1	3	0

4	60	1	3	3	0
5	87	4	5	3	1
4	70	4	4	3	1
5	55	4	5	3	1
3	21	1	1	3	0
4	50	1	1	3	0
5	76	4	5	3	1
4	23	1	1	3	0
3	68	?	?	3	0
4	62	4	?	3	1
5	65	1	?	3	1
5	73	4	5	3	1
4	38	2	3	3	0
2	57	1	1	3	0
5	65	4	5	3	1
5	67	2	4	3	1
5	61	2	4	3	1
5	56	4	4	3	0
5	71	2	4	3	1
4	49	2	2	3	0
4	55	?	?	3	0
4	44	2	1	3	0
0	58	4	4	3	0
4	27	2	1	3	0
5	73	4	5	3	1
4	34	2	1	3	0
5	63	?	4	3	1
4	50	2	1	3	1
4	62	2	1	3	0
3	21	3	1	3	0
4	49	2	?	3	0
4	36	3	1	3	0
4	45	2	1	3	1
5	67	4	5	3	1
4	21	1	1	3	0
4	57	2	1	3	0
5	66	4	5	3	1
4	71	4	4	3	1
5	69	3	4	3	1
6	80	4	5	3	1
3	27	2	1	3	0
4	38	2	1	3	0
4	23	2	1	3	0
5	70	?	5	3	1



4	46	4	3	3	0
4	61	2	3	3	0
5	65	4	5	3	1
4	60	4	3	3	0
5	83	4	5	3	1
5	40	4	4	3	1
2	59	?	4	3	0
4	53	3	4	3	0
4	76	4	4	3	0
5	79	1	4	3	1
5	38	2	4	3	1
4	61	3	4	3	0
4	56	2	1	3	0
4	44	2	1	3	0
4	64	3	4	?	1
4	66	3	3	3	0
4	50	3	3	3	0
4	46	1	1	3	0
4	39	1	1	3	0
4	60	3	?	?	0
5	55	4	5	3	1
4	40	2	1	3	0
4	26	1	1	3	0
5	84	3	2	3	1
4	41	2	2	3	0
4	63	1	1	3	0
2	65	?	1	2	0
4	49	1	1	3	0
4	56	2	2	3	1
5	65	4	4	3	0
4	54	1	1	3	0
4	36	1	1	3	0
5	49	4	4	3	0
4	59	4	4	3	1
5	75	4	4	3	1
5	59	4	2	3	0
5	59	4	4	3	1
4	28	4	4	3	1
5	53	4	5	3	0
5	57	4	4	3	0
5	77	4	3	4	0
5	85	4	3	3	1
4	59	4	4	3	0
5	59	1	5	3	1

4	65	3	3	3	1
4	54	2	1	3	0
5	46	4	5	3	1
4	63	4	4	3	1
4	53	1	1	3	1
4	56	1	1	3	0
5	66	4	4	3	1
5	66	4	5	3	1
4	55	1	1	3	0
4	44	1	1	3	0
5	86	3	4	3	1
5	47	4	5	3	1
5	59	4	5	3	1
5	66	4	5	3	0
5	61	4	3	3	1
3	46	?	5	?	1
4	69	1	1	3	0
5	93	1	5	3	1
4	39	1	3	3	0
5	44	4	5	3	1
4	45	2	2	3	0
4	51	3	4	3	0
4	56	2	4	3	0
4	66	4	4	3	0
5	61	4	5	3	1
4	64	3	3	3	1
5	57	2	4	3	0
5	79	4	4	3	1
4	57	2	1	?	0
4	44	4	1	1	0
4	31	2	1	3	0
4	63	4	4	3	0
4	64	1	1	3	0
5	47	4	5	3	0
5	68	4	5	3	1
4	30	1	1	3	0
5	43	4	5	3	1
4	56	1	1	3	0
4	46	2	1	3	0
4	67	2	1	3	0
5	52	4	5	3	1
4	67	4	4	3	1
4	47	2	1	3	0
5	58	4	5	3	1

4	28	2	1	3	0
4	43	1	1	3	0
4	57	2	4	3	0
5	68	4	5	3	1
4	64	2	4	3	0
4	64	2	4	3	0
5	62	4	4	3	1
4	38	4	1	3	0
5	68	4	4	3	1
4	41	2	1	3	0
4	35	2	1	3	1
4	68	2	1	3	0
5	55	4	4	3	1
5	67	4	4	3	1
4	51	4	3	3	0
2	40	1	1	3	0
5	73	4	4	3	1
4	58	?	4	3	1
4	51	?	4	3	0
3	50	?	?	3	1
5	59	4	3	3	1
6	60	3	5	3	1
4	27	2	1	?	0
5	54	4	3	3	0
4	56	1	1	3	0
5	53	4	5	3	1
4	54	2	4	3	0
5	79	1	4	3	1
5	67	4	3	3	1
5	64	3	3	3	1
4	70	1	2	3	1
5	55	4	3	3	1
5	65	3	3	3	1
5	45	4	2	3	1
4	57	4	4	?	1
5	49	1	1	3	1
4	24	2	1	3	0
4	52	1	1	3	0
4	50	2	1	3	0
4	35	1	1	3	0
5	?	3	3	3	1
5	64	4	3	3	1
5	40	4	1	1	1
5	66	4	4	3	1

4	64	4	4	3	1
5	52	4	3	3	1
5	43	1	4	3	1
4	56	4	4	3	0
4	72	3	?	3	0
6	51	4	4	3	1
4	79	4	4	3	1
4	22	2	1	3	0
4	73	2	1	3	0
4	53	3	4	3	0
4	59	2	1	3	1
4	46	4	4	2	0
5	66	4	4	3	1
4	50	4	3	3	1
4	58	1	1	3	1
4	55	1	1	3	0
4	62	2	4	3	1
4	60	1	1	3	0
5	57	4	3	3	1
4	57	1	1	3	0
6	41	2	1	3	0
4	71	2	1	3	1
4	32	2	1	3	0
4	57	2	1	3	0
4	19	1	1	3	0
4	62	2	4	3	1
5	67	4	5	3	1
4	50	4	5	3	0
4	65	2	3	2	0
4	40	2	4	2	0
6	71	4	4	3	1
6	68	4	3	3	1
4	68	1	1	3	0
4	29	1	1	3	0
4	53	2	1	3	0
5	66	4	4	3	1
4	60	3	?	4	0
5	76	4	4	3	1
4	58	2	1	2	0
5	96	3	4	3	1
5	70	4	4	3	1
4	34	2	1	3	0
4	59	2	1	3	0
4	45	3	1	3	1

5	65	4	4	3	1
4	59	1	1	3	0
4	21	2	1	3	0
3	43	2	1	3	0
4	53	1	1	3	0
4	65	2	1	3	0
4	64	2	4	3	1
4	53	4	4	3	0
4	51	1	1	3	0
4	59	2	4	3	0
4	56	2	1	3	0
4	60	2	1	3	0
4	22	1	1	3	0
4	25	2	1	3	0
6	76	3	?	3	0
5	69	4	4	3	1
4	58	2	1	3	0
5	62	4	3	3	1
4	56	4	4	3	0
4	64	1	1	3	0
4	32	2	1	3	0
5	48	?	4	?	1
5	59	4	4	2	1
4	52	1	1	3	0
4	63	4	4	3	0
5	67	4	4	3	1
5	61	4	4	3	1
5	59	4	5	3	1
5	52	4	3	3	1
4	35	4	4	3	0
5	77	3	3	3	1
5	71	4	3	3	1
5	63	4	3	3	1
4	38	2	1	2	0
5	72	4	3	3	1
4	76	4	3	3	1
4	53	3	3	3	0
4	67	4	5	3	0
5	69	2	4	3	1
4	54	1	1	3	0
2	35	2	1	2	0
5	68	4	3	3	1
4	68	4	4	3	0
4	67	2	4	3	1

3	39	1	1	3	0
4	44	2	1	3	0
4	33	1	1	3	0
4	60	?	4	3	0
4	58	1	1	3	0
4	31	1	1	3	0
3	23	1	1	3	0
5	56	4	5	3	1
4	69	2	1	3	1
6	63	1	1	3	0
4	65	1	1	3	1
4	44	2	1	2	0
4	62	3	3	3	1
4	67	4	4	3	1
4	56	2	1	3	0
4	52	3	4	3	0
4	43	1	1	3	1
4	41	4	3	2	1
4	42	3	4	2	0
3	46	1	1	3	0
5	55	4	4	3	1
5	58	4	4	2	1
5	87	4	4	3	1
4	66	2	1	3	0
0	72	4	3	3	1
5	60	4	3	3	1
5	83	4	4	2	1
4	31	2	1	3	0
4	53	2	1	3	0
4	64	2	3	3	0
5	31	4	4	2	1
5	62	4	4	2	1
4	56	2	1	3	0
5	58	4	4	3	1
4	67	1	4	3	0
5	75	4	5	3	1
5	65	3	4	3	1
5	74	3	2	3	1
4	59	2	1	3	0
4	57	4	4	4	1
4	76	3	2	3	0
4	63	1	4	3	0
4	44	1	1	3	0
4	42	3	1	2	0

4	35	3	?	2	0
5	65	4	3	3	1
4	70	2	1	3	0
4	48	1	1	3	0
4	74	1	1	1	1
6	40	?	3	4	1
4	63	1	1	3	0
5	60	4	4	3	1
5	86	4	3	3	1
4	27	1	1	3	0
4	71	4	5	2	1
5	85	4	4	3	1
4	51	3	3	3	0
6	72	4	3	3	1
5	52	4	4	3	1
4	66	2	1	3	0
5	71	4	5	3	1
4	42	2	1	3	0
4	64	4	4	2	1
4	41	2	2	3	0
4	50	2	1	3	0
4	30	1	1	3	0
4	67	1	1	3	0
5	62	4	4	3	1
4	46	2	1	2	0
4	35	1	1	3	0
4	53	1	1	2	0
4	59	2	1	3	0
4	19	3	1	3	0
5	86	2	1	3	1
4	72	2	1	3	0
4	37	2	1	2	0
4	46	3	1	3	1
4	45	1	1	3	0
4	48	4	5	3	0
4	58	4	4	3	1
4	42	1	1	3	0
4	56	2	4	3	1
4	47	2	1	3	0
4	49	4	4	3	1
5	76	2	5	3	1
5	62	4	5	3	1
5	64	4	4	3	1
5	53	4	3	3	1

4	70	4	2	2	1
5	55	4	4	3	1
4	34	4	4	3	0
5	76	4	4	3	1
4	39	1	1	3	0
2	23	1	1	3	0
4	19	1	1	3	0
5	65	4	5	3	1
4	57	2	1	3	0
5	41	4	4	3	1
4	36	4	5	3	1
4	62	3	3	3	0
4	69	2	1	3	0
4	41	3	1	3	0
3	51	2	4	3	0
5	50	3	2	3	1
4	47	4	4	3	0
4	54	4	5	3	1
5	52	4	4	3	1
4	30	1	1	3	0
3	48	4	4	3	1
5	?	4	4	3	1
4	65	2	4	3	1
4	50	1	1	3	0
5	65	4	5	3	1
5	66	4	3	3	1
6	41	3	3	2	1
5	72	3	2	3	1
4	42	1	1	1	1
4	80	4	4	3	1
0	45	2	4	3	0
4	41	1	1	3	0
4	72	3	3	3	1
4	60	4	5	3	0
5	67	4	3	3	1
4	55	2	1	3	0
4	61	3	4	3	1
4	55	3	4	3	1
4	52	4	4	3	1
4	42	1	1	3	0
5	63	4	4	3	1
4	62	4	5	3	1
4	46	1	1	3	0
4	65	2	1	3	0



4	57	3	3	3	1
4	66	4	5	3	1
4	45	1	1	3	0
4	77	4	5	3	1
4	35	1	1	3	0
4	50	4	5	3	1
4	57	4	4	3	0
4	74	3	1	3	1
4	59	4	5	3	0
4	51	1	1	3	0
4	42	3	4	3	1
4	35	2	4	3	0
4	42	1	1	3	0
4	43	2	1	3	0
4	62	4	4	3	1
4	27	2	1	3	0
5	?	4	3	3	1
4	57	4	4	3	1
4	59	2	1	3	0
5	40	3	2	3	1
4	20	1	1	3	0
5	74	4	3	3	1
4	22	1	1	3	0
4	57	4	3	3	0
4	57	4	3	3	1
4	55	2	1	2	0
4	62	2	1	3	0
4	54	1	1	3	0
4	71	1	1	3	1
4	65	3	3	3	0
4	68	4	4	3	0
4	64	1	1	3	0
4	54	2	4	3	0
4	48	4	4	3	1
4	58	4	3	3	0
5	58	3	4	3	1
4	70	1	1	1	0
5	70	1	4	3	1
4	59	2	1	3	0
4	57	2	4	3	0
4	53	4	5	3	0
4	54	4	4	3	1
4	53	2	1	3	0
0	71	4	4	3	1

5	67	4	5	3	1
4	68	4	4	3	1
4	56	2	4	3	0
4	35	2	1	3	0
4	52	4	4	3	1
4	47	2	1	3	0
4	56	4	5	3	1
4	64	4	5	3	0
5	66	4	5	3	1
4	62	3	3	3	0