UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

CLASSIFICAÇÃO DE NÓDULOS MAMÁRIOS UTILIZANDO SEGMENTAÇÃO E ANÁLISE DE DENSIDADE DE IMAGENS MAMOGRÁFICAS

STEFAN MULLER FERREIRA GONÇALVES

ORIENTADOR: PROF. DR. HOMERO SCHIABEL

SÃO CARLOS
2015

STEFAN MULLER FERREIRA GONÇALVES

CLASSIFICAÇÃO DE NÓDULOS MAMÁRIOS UTILIZANDO SEGMENTAÇÃO E ANÁLISE DE DENSIDADE DE IMAGENS MAMOGRÁFICAS

Trabalho de Conclusão de Curso apresentado
à Escola de Engenharia de São Carlos, da
Universidade de São Paulo

Curso de Engenharia Elétrica com Ênfase em
Eletrônica

Orientador: Prof. Dr. Homero Schiabel

São Carlos
2015

Muller Ferreira Gonçalves, Stefan
M958c Classificação de nódulos mamários utilizando segmentação e análise de densidade de imagens mamográficas / Stefan Muller Ferreira Gonçalves; orientador Homero Schiabel; coorientador Bruno Roberto Nepomuceno Matheus. São Carlos, 2015.

Monografia (Graduação em Engenharia Elétrica com ênfase em Eletrônica) -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2015.

1. CADx. 2. Mamografia. 3. Segmentação de imagem. 4. Densidade dos níveis de cinza. 5. Método de Otsu. I. Título.

FOLHA DE APROVAÇÃO

Nome: Stefan Muller Ferreira Gonçalves

Título: "Classificação de nódulos mamários utilizando segmentação e análise de densidade de imagens mamográficas"

Trabalho de Conclusão de Curso defendido e aprovado em 30 / 11 / 2015,

com NOTA 7,5 (sete , cinco), **pela Comissão Julgadora:**

Prof. Associado Homero Schiabel - (Orientador - SEL/EESC/USP)

Dr. Bruno Roberto Nepomuceno Matheus - (SEL/EESC/USP)

Dr. Mauro Masili - (Pós-Doutorando - SEL/EESC/USP)

Coordenador da CoC-Engenharia Elétrica - EESC/USP:
Prof. Dr. José Carlos de Melo Vieira Júnior

RESUMO

No Brasil, exceto na região Norte, o câncer de mama é o tipo mais comum entre as mulheres (INCA, 2015). Como inda não é possível prevenir esse tipo de câncer, o diagnóstico precoce aumenta consideravelmente as chances de cura da doença e o exame mamográfico é uma das maneiras eficazes de se realizar esse diagnóstico. Estudos mostram que, de todas as mamografias avaliadas, de 10 a 15% tem seu diagnóstico errado (PEIXOTO, CANELLA e AZEVEDO, 2007)e isso pode levar a biópsias desnecessárias, ou até mesmo, fazer com que a paciente não receba o tratamento necessário. Uma possível solução para esse problema é usar um sistema CADx (Computer-aided Diagnosis) que analisa e classifica o nódulo presente em uma mamografia baseado em diferentes atributos.

Seguindo uma das linhas de pesquisa do Laboratório de Análise e Processamento de Imagens Médicas e Odontológicas (LAPIMO) da Escola de Engenharia de São Carlos da USP, este trabalho tem como objetivo implementar um programa independente de plataforma que ajude a estabelecer a correlação entre a densidade de níveis de cinza do nódulo presente em uma mamografia e sua classificação (benigno ou maligno), ou seja, apenas um dos atributos que são levados em conta no sistema CADx.

Durante o desenvolvimento do programa foram implementados alguns filtros para melhorar as segmentação do nódulo e da mama que é feita através do método de Otsu (OTSU, 1979). A versão final do programa permite, através de sua interface gráfica, que o usuário ajuste diversos parâmetros relacionados aos filtros implementados e também a utilizar nódulos já segmentados por outros métodos.

Com o programa desenvolvido foi possível processar um grande número de mamografias, extraindo a densidade dos níveis de cinza do nódulo e da mama e com isso testar a eficácia de dois sistemas de classificação, a razão das médias do nódulo e da mama e o *K-Nearest Neighbours* (KNN), a fim de verificar a correlação entre a classificação do nódulo e sua densidade de níveis de cinza.

Com os resultados foi possível verificar que a acurácia na classificação do nódulo depende fortemente do sistema de classificação e do escâner usado na digitalização da mamografia. O sistema de classificação KNN é superior à razão das médias do nódulo e da mama, sendo possível com ele atingir uma acurácia de 69,81% para uma amostra de 100 imagens e 66,17% com uma amostra de 352 imagens, usando apenas a densidade de níveis de cinza como atributo.

Palavras-chave: CADx, Mamografia, Segmentação de imagem, Densidade dos níveis de cinza, Método de Otsu.

ABSTRACT

In Brazil, except in the North, breast cancer is the most common type among women (INCA, 2015). As is not yet possible to prevent this type of cancer, early diagnosis greatly increases the chances of curing the disease and mammography is one of the most effective ways to accomplish this diagnosis. Studies show that, from all the mammograms, 10 to 15% has a wrong diagnosis(PEIXOTO, CANELLA e AZEVEDO, 2007) and this can lead to unnecessary biopsies, or even cause te patient to not receive the treatment needed. A possible solution is to use a CADx (Computer-aided Diagnosis) scheme that is able to analyse and classify, based on different attributes, the nodule present on a mammography.

Following one of the research lines of the Medical and Dental image Processing and Analysis Laboratory (LAPIMO) from the Engineering School from USP São Carlos, this term paper has the objective to describe the implementation of a plataform-independent software that helps to establish the correlation between the grayscale density of a nodule present in a mammography and its classification (benign or malign), that is, only one of the attributes that are taken into account in a CADx scheme. During the software development, some filters that improve nodule and breast segmentation, that is done by Otsu method (OTSU, 1979), were created and implemented. On the final version, the user is able to adjust many parameters associated to the filters implemented and is also able to use nodules already segmented by another method.

With the software created was possible to process a large number of mammograms, extracting the nodule and breast grayscale density to test the effectiveness from two classification systems, the ratio between nodule and breast means and also the K-Nearest Neighbours (KNN), in order to verify the correlation between the nodule classification and its grayscale density.

With the results was possible to ascertain that the classification accuracy heavily depends on the classification system and the scanner, to digitalize the mammography, used. The KNN method is superior to the ration between nodule and breast means, the first one being able to achieve accuracy of 69,81% for a sample with 100 images and 66,17% for a sample with 352 images, using only the greyscale density as attribute.

Keywords: CADx, Mammography, Image segmentation, Grayscale density, Otsu Method.

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

API	*Application Programming Interface*
CADx	*Computer-aided Diagnosis*
DDSM	*Digital Database for Screening Mammography*
ECM	Exame Clínico das Mamas
JAI	Java *Advanced Imaging*
JDK	Java *Development Kit*
JSE	Java *Standard Edition*
JVM	Java *Virtual Machine*
LAPIMO	Laboratório de Análise e Processamento de Imagens Médicas e Odontológicas
RGB	*"Red, Green, Blue"*
ROC	*Receiver Operating Characteristic*
ROI	*Region of Interest*
TIFF	*Tagged Image File Format*
KNN	K-Nearest Neighbour
SVM	Support Vector Machines

Sumário

1 Introdução

1.1 Câncer de mama

O nome câncer é dado a uma série de doenças que estão relacionadas e em todas elas, alguma célula do corpo começa a se dividir descontroladamente e se espalha para tecidos vizinhos. Esse tipo de doença tem diversas manifestações clínicas, derivadas de variações genéticas e morfológicas (BUZARD, MALUF e LIMA, 2015).

No Brasil, especificamente o câncer de mama, é o tipo mais comum entre as mulheres em todas as regiões exceto na região Norte, onde fica em segundo lugar (INCA, 2015). Além disso, de acordo com a *World Cancer Research Fund International*, no mundo, o câncer de mama é o tipo mais comum entre as mulheres.

Ainda não é possível prevenir o câncer de mama em função da multiplicidade de fatores relacionados ao surgimento da doença, desta forma, o diagnóstico precoce é a melhor forma de combater a doença e dentre os métodos para detecção do câncer de mama destacam-se o Exame Clínico das Mamas (ECM) e a mamografia, essa última sendo o objeto de estudo deste trabalho (BUZARD, MALUF e LIMA, 2015).

1.2 Mamografia

A mamografia consiste na radiografia da mama e ajuda no diagnóstico de doenças relacionadas a ela. A detecção prematura do câncer de mama depende primariamente da mamografia, porém a análise de uma eventual anormalidade presente no raio-X da mama pode ser complicada e imagens com alto contraste, nitidez e resolução são necessárias para um diagnóstico correto e preciso. Devido a incertezas na análise das radiografias, existe uma taxa de falsos negativos e falsos positivos no diagnóstico por imagem (BOYLE, 2002).

Para a obtenção das mamografias é utilizado um aparelho projetado para obter radiografias das mamas chamado mamógrafo. Este aparelho permite identificar estruturas de até 0.3 mm de diâmetro e foi especialmente desenvolvido para trabalhar com tecidos de densidades próximas às encontradas nas mamas. O mamógrafo restringe e direciona o feixe de raio-X a fim de diminuir a dose de radiação a que a paciente é exposta e ao mesmo tempo melhorar a qualidade da imagem gerada (BICK e DIEKMANN, 2010).

Se a anormalidade encontrada na mamografia for suspeita, a paciente é chamada para complementar a avaliação, o que consiste em mais radiografias, ultrassonografias e, se necessário, até uma biópsia (CASSIDY, BISSETT, 2010).

1.3 Nódulos

Nódulos mamários podem ser causados por infecções, lesões, tumores não cancerosos e câncer (BUZARD, MALUF e LIMA, 2015). Sendo assim, podemos dividir os nódulos mamários em:

- Mastite: Causado por infecção e consequente inflamação do tecido da mama.
- Hematoma: Causado por lesões no tecido da mama que pode ocasionar a formação de um nódulo.
- Tumores não cancerosos:
  - Fibroadenoma: Muito comuns, são tumores sólidos, firmes e indolores.
  - Cisto mamário: Pequenas bolsas preenchidas com fluido que se formam dentro da mama.
  - Mama fibrocística: São caracterizadas por mamas irregulares e com sensação de pequenos grânulos dentro.
- Câncer: Tumor sólido e maligno que se desenvolve no tecido mamário.

O câncer de mama é mais comum na mama esquerda e em cerca de 50% dos casos ocorre no quadrante superior externo (CASSIDY, BISSETT, 2010).

1.4 Sistemas CADx

A utilização de esquema *Computer-aided Diagnosis* (CADx) é considerada uma das abordagens que aumentam a eficácia de exames mamográficos e um número considerável de pesquisas têm sido dedicadas no desenvolvimento e aprimoramento de técnicas usadas nesses sistemas (CHAN, 1999).

Estudos recentes como o de (LIU e TANG, 2014), que usam esquema CADx com sistema de classificação baseados em aprendizagem supervisionada e método *k-fold* de validação cruzada, resultam em programas com acurácia de até 94% na classificação de nódulos benignos e malignos.

É importante notar que trabalhos como o de (LIU e TANG, 2014), além de utilizar sistemas de classificação complexos, fazem uso de mais de 30 atributos relacionados a textura e geometria que devem ser extraídos do nódulo e por isso é de interesse melhorar a acurácia de sistemas mais simples e que levam em conta menos parâmetros, como o trabalho de (MATHEUS, 2015), o qual implementa um esquema CADx cujo classificador de nódulos apresenta 72% de acurácia usando apenas 4 atributos.

Seguindo uma das linhas de pesquisa do Laboratório de Análise e Processamento de Imagens Médicas e Odontológicas (LAPIMO) da Escola de Engenharia de São Carlos da USP, em especial relacionada ao trabalho de (MATHEUS, 2015), busca-se estabelecer a relação entre a classificação de um nódulo e sua densidade de níveis de cinza. Deseja-se verificar a acurácia do diagnóstico levando-se em conta apenas 1 dos 4 atributos (a densidade relativa dos níveis de cinza da mamografia) para determinar se este é ou não um parâmetro pertinente e que deveria fazer parte do sistema CADx completo.

1.5 Objetivos

O objetivo é desenvolver um programa, independente de plataforma, capaz de:

- Receber mamografias de casos conhecidos como entrada.
- Segmentar a mama e o nódulo tornando possível variar os parâmetros da técnica de segmentação.
- Calcular a densidade dos níveis de cinza da mama e do nódulo e retornar um arquivo com esses valores.

A proposta desse projeto é, operando dentro do esquema CADx do LAPIMO, na etapa de classificação de nódulos mamários, melhorar o diagnóstico por imagem através do desenvolvimento de um método pata estabelecer a relação entre a densidade dos níveis de cinza de um nódulo presente na mamografia e sa classificação deste nódulo (maligno ou benigno).

Para investigar essa possível relação é preciso construir uma estatística de teste confiável por meio de treinamento de técnica heurística que por sua vez requer uma grande amostra de imagens de casos conhecidos. Analisar e processar um grande número de imagens pode tomar um alto tempo de processamento e, buscando minimizar o tempo, o programa será implementado de forma a permitir o processamento paralelo das imagens. Java é uma linguagem *multithreading* nativa e isso será explorado no programa principal.

1.6 Disposição do trabalho

Este trabalho se dispõe da seguinte maneira:

- Capítulo 2: Descrição dos materiais e métodos utilizados no trabalho.
- Capítulo 3: Estudo preliminar e simplificado no MATLAB com uma amostra pequena de imagens.
- Capítulo 4: Descrição da engenharia de software utilizada no programa final bem como o detalhamento dos métodos implementados no programa.
- Capítulo 5: Apresentação e discussão dos resultados.
- Capítulo 6: Conclusões e propostas para desenvolvimentos futuros.

2 Materiais e Métodos

2.1 Base de dados utilizada, DDSM (HEATH e BOWYER, 2001)

O conjunto de imagens utilizada neste trabalho foram obtidas da base de dados denominada *Digital Database for Screening Mammography* (DDSM), que são mamografias tela/filme digitalizadas. Essa base de dados é um projeto conjunto envolvendo pesquisadores do Massachusetts General Hospital (D. Kopans, R. Moore), da Universidade do Sul Da Florida (K. Bowyer) e do Sandia National Laboratories - EUA (P. Kegelmeyer). Por ser de utilização gratuita e possuir grande acervo de imagens com informações técnicas e clínicas correspondentes, essa base de dados é bastante utilizada em artigos da área de diagnóstico mamográfico (NASCIMENTO e RAMOS, 2008). Dentre as informações disponíveis para cada mamografia, as que são importantes para esse trabalho são:

- Equipamento (escâner) utilizado na digitalização da mamografia.
- Quantidade de regiões de interesse ou *Region of Interest* (ROI) contidas na mamografia. Cada ROI corresponde a um nódulo suspeito.
- Coordenadas da ROI, em forma de *code-chain* (código em cadeia), contendo o nódulo a ser classificado.
- Classificação do nódulo.

2.2 Sistemas de classificação

Com o *toolbox* de aprendizado supervisionado do MATLAB foram explorados alguns sistemas de classificação variando-se o tamanho da amostra de mamografias, dentre eles estão:

- Árvores de decisão
- K-Nearest Neighbour (KNN)
- Support Vector Machines (SVM)
- Razão das densidades do nódulo e da mama (μ_n/μ_f)

Dos classificadores citados acima, por simplicidade, foram escolhidos o KNN e a razão das médias.

2.2.1 K-Nearest Neighbor (KNN)

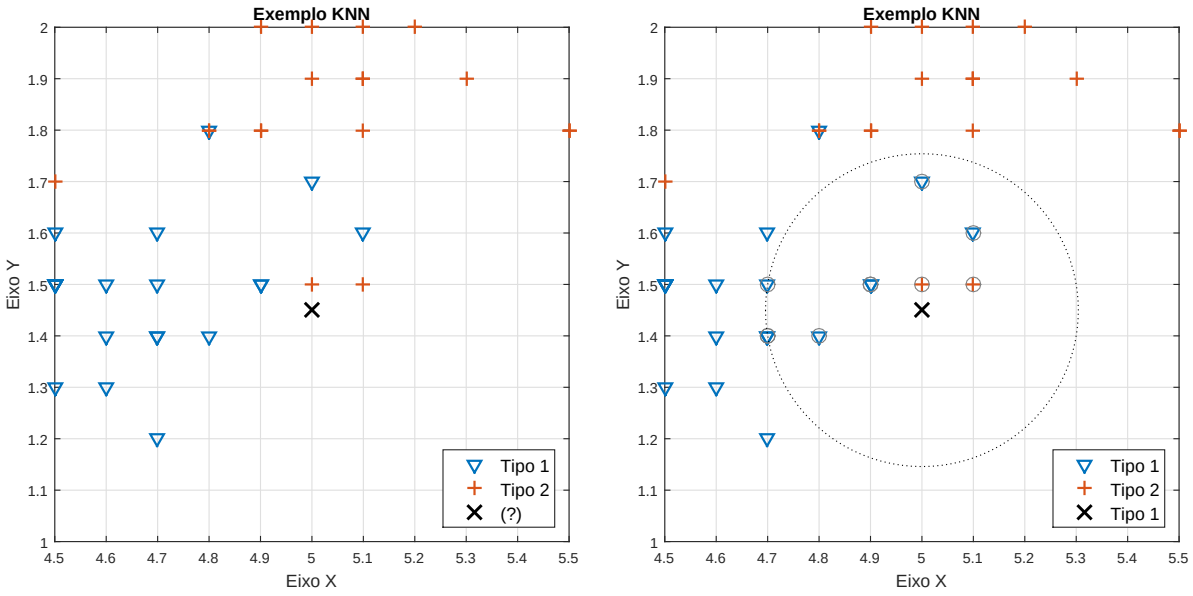
É um dos algoritmos mais simples de classificação, usado para classificar objetos com base nos exemplos de treinamento que estão mais próximos (MURTY, DEVI, 2011). Para utilizar o KNN é necessário:

- Conjunto de exemplos de treinamento (casos conhecidos)
- Definir uma métrica para calcular a distância do objeto até os exemplos de treinamento (geralmente utiliza-se a métrica euclidiana)
- Definir o valor de k (número de vizinhos mais próximos que serão considerados pelo algoritmo). Essa definição é feita através de tentativa criteriosa, onde busca-se um valor de k para se obter boa taxa de acerto.

Portanto, classificar um objeto desconhecido utilizando o algoritmo KNN consiste em:

- Calcular a distância entre o objeto a ser classificado e os exemplos de treinamento
- Identificar os k vizinhos mais próximos
- Utilizar o rótulo desses vizinhos para determinar a classificação do objeto desconhecido, com base em votação majoritária. Se houver empate no número de rótulos a escolha é arbitrária e portanto, para contornar esse problema, é melhor escolher k ímpar.

A Figura 1 a seguir ilustra o funcionamento do algoritmo para classificar um objeto representado pelo ponto **X**.



(a) KNN Antes (b) KNN Depois

Figura 1: Funcionamento do algoritmo K-Nearest Neighbor

No exemplo acima, a distância considerada é a distância euclidiana e o número de vizinhos mais próximos considerado foi arbitrariamente escolhido como $k = 8$. Desses 8 objetos mais próximos de **X**, 6 são do Tipo 1 e portanto, por votação majoritária, **X** é classificado como do Tipo 1.

O MATLAB implementa esse algoritmo e também permite realizar validação cruzada dos dados pelo método *k-fold* para treinar o conjunto de dados e obter um k que otimiza a acurácia do teste. Essa técnica consiste em particionar o conjunto de dados em k subconjuntos mutuamente exclusivos para utilizar alguns deles como exemplos de treinamento e o restante para validar o classificador. Isso traz o erro de validação mais próximo da realidade (KOHAVI, 1995).

2.2.2 Razão das médias do nódulo e da mama (μ_n/μ_f)

A razão das médias do nódulo e da mama de uma amostra grande de mamografias de casos conhecidos permite determinar um limiar ótimo tal que um caso desconhecido seja classificado como suspeito se sua razão μ_n/μ_f estiver acima desse limiar e não suspeito de estiver abaixo.

2.3 *Receiver Operating Characteristic Curve* (Curva ROC)

Em um teste binário (positivo ou negativo) de diagnóstico, os seguintes resultados são possíveis :

- Verdadeiro positivo (VP)
- Falso negativo (FN)
- Falso positivo (FP)
- Verdadeiro negativo (VN)

Sendo assim, podemos definir a sensibilidade S e a especificidade E como

$$S \triangleq \frac{\sum VP}{\sum VP + \sum FN}, \quad E \triangleq \frac{\sum VN}{\sum FP + \sum VN} \tag{1}$$

onde a Sensibilidade S representa a taxa de casos malignos que foram corretamente identificados e, analogamente, a Especificidade E representa a taxa de casos benignos que foram corretamente identificados. Na prática, não adianta S alto e E baixo, ou seja, não queremos $S = 1 - E$, mas sim $S > 1 - E$.

Para obter a curva *Receiver Operating Characteristic* (ROC), traça-se $S \times E$ variando-se o limiar de discriminação. A área embaixo dessa curva pode ser usada como indicador da eficácia de um sistema de classificação uma vez que quanto maior a área, maior a taxa de acerto do sistema. Essa área varia no intervalo $[0,1]$, sendo o máximo ideal 1, que representa 100% de sensibilidade e especificidade.

A acurácia, que também é um indicador da eficácia do sistema, pode ser calculada como:

$$A_{cc} = \frac{\sum VP + \sum VN}{\text{nº de imagens na amostra}} \tag{2}$$

ou ainda, em termos da prevalência P_v (probabilidade da existência do sinal na população), temos:

$$A_{cc} = S \cdot P_v + E \cdot (1 - P_v) \tag{3}$$

A acurácia representa a taxa de acerto do sistema e foi usada para testar a eficácia de ambos os sistemas de classificação (KNN e Razão das médias). É importante notar que a acurácia deve ser interpretada com cuidado uma vez que, para o diagnóstico de uma condição rara em determinada população, ambas sensitividade e especificidade podem ser altas mas a acurácia baixa. Analogamente, uma condição comum pode resultar em acurácia alta e sensibilidade e especificidade baixas (ZHU, 2010).

Para traçar a curva ROC referente ao sistema de classificação pela razão das médias, basta variar o limiar a partir do qual mamografias com densidade relativa superior serão classificadas como caso maligno e mamografias com densidades relativa inferior serão classificadas como benigno.

Para o sistema de classificação pelo método KNN, a obtenção da curva ROC não é tão simples, é preciso calcular, para cada ponto X, a probabilidade a posteriori que é dada por:

$$p(j|X) = \frac{\sum\limits_{i \in knn} W(i)\beta(i)}{\sum\limits_{i \in knn} W(i)} \tag{4}$$

onde

- X são os pontos de coordenada (μ_n, μ_f)
- j é a classe verdadeira de cada ponto (maligno=1 ou benigno=0)
- knn são os k vizinhos mais próximos
- $\beta(i) = 1$ se o ponto $X(i)$ é maligno e 0 caso contrário
- $W(i)$ é a função peso, $W(i) = \dfrac{1}{\sum\limits_{i \in knn}(1)}$ quando não especificada

De posse de (4) o resultado da classificação (maligno ou benigno) é dado por:

$$\hat{y} = \underset{y=1,2}{\operatorname{argmin}} \sum_{k=1}^{2} \hat{P}(k|x)C(y|k) \tag{5}$$

onde

- $\hat{y}$ é o resultado da classificação ($\hat{y} = 1$ para casos malignos e $\hat{y} = 0$ caso contrário)
- $\hat{P}(k|x)$ é a probabilidade posteriori dada por (4)
- $C(y|k)$ é o custo de classificar uma observação como y quando sua verdadeira classe é k. De modo geral, $C(y|k) = 0$ se $y(i) = k$ e $C(y|k) = 1$ caso contrário.

O MATLAB possui funções que simplificam esse processo.

2.4 Métodos de segmentação de imagem

2.4.1 Método de Otsu (OTSU, 1979)

O método de Otsu (OTSU, 1979) aproxima o histograma de uma imagem por duas funções Gaussianas e escolhe o limiar de forma a minimizar a variância intraclasses. Cada classe possui suas próprias características, ou seja, sua média e desvio-padrão.

Considere uma imagem digital I, de dimensões $M \times N$ e quantizada em L níveis de cinza. O primeiro passo é calcular o histograma p da imagem, dado por

$$p_i = \frac{n_i}{M \cdot N} \tag{6}$$

em que n_i é a quantidade de *pixels* da imagem I que possuem a intensidade de cinza i, para $i = 0, 1, 2, \ldots, L-1$. Assim, $M \cdot N = n_0 + n_1 + n_2 + \cdots + n_{L-1}$ e

$$\sum_{i=0}^{L-1} p_i = 1 \tag{7}$$

Seja k o nível de cinza que particiona o histograma da imagem em duas classes C_1 e C_2, em que a primeira e a segunda classe compreendem os *pixels* cujos níveis de cinza pertencem ao intervalo $[0,k]$ e $[k+1, L-1]$. Assim, podemos definir as probabilidades

- $P_1(k)$ é a probabilidade do nível de cinza k ser da classe C_1
- $P_2(k)$ é a probabilidade do nível de cinza k ser da classe C_2

$$P_1(k) = \sum_{i=0}^{k} p_i, \quad P_2(k) = \sum_{i=k+1}^{L-1} p_i \tag{8}$$

Como o histograma é aproximado por duas funções Gaussianas

$$m_1(k) = \sum_{i=0}^{k} iP(i|C_1) \tag{9}$$

e utilizando a regra de Bayes, temos

$$m_1(k) = \sum_{i=0}^{k} i\frac{P(i|C_1)P(i)}{P(C_1)} \tag{10}$$

$P(C_1) = P_1(k)$, $P(i)$ é o próprio p_i e $P(C_1|i)$ é sempre 1 pois i está no intervalo de cinza da própria classe C_1. Assim

$$m_1(k) = \frac{1}{P_1(k)} \sum_{i=0}^{k} ip_i \tag{11}$$

Similarmente,

$$m_2(k) = \frac{1}{P_2(k)} \sum_{i=k+1}^{L-1} ip_i \tag{12}$$

Por sua vez, a variância para cada distribuição de probabilidade pode ser determinada por

$$\sigma_1^2(k) = \frac{1}{P_1(k)} \sum_{i=0}^{k} (m_1(k) - p_i)^2 \tag{13}$$

$$\sigma_2^2(k) = \frac{1}{P_2(k)} \sum_{i=k+1}^{L-1} (m_2(k) - p_i)^2 \tag{14}$$

Após calcular σ_C^2 para todos os valores de k, determina-se o limiar ótimo k^* de acordo com a Eq.15

$$k^* = \min_{0 \leqslant k \leqslant L-1} \sigma_C^2(k) \tag{15}$$

A implementação do método de Otsu em Java encontra-se no **Anexo A**.

2.4.2 *Level Set*

Como explicado anteriormente, o programa permite usar arquivos externos para os nódulos, de forma a comparar resultados com técnicas de segmentação implementadas em outros programas. O LAPIMO possui uma coleção de nódulos já segmentados por diversos métodos, dentre eles o método *Level Set*. Essa técnica é moderna e possui a vantagem de garantir suavidade nos contornos da região segmentada (CHUNMING, HUANG E DING, 2011).

O método de *Level Set* foi proposto inicialmente para o estudo de frentes cuja velocidade depende da curvatura local como o crescimento de cristais e propagação de chamas, tal método introduziu algoritmos numéricos para o estudo destes problemas e mais tarde foi usado no processamento de imagens (OSHER, SETHIAN, 1988).

Para a segmentação de imagem, uma versão simples do método consiste em definir o contorno de segmentação como parte de uma superfície onde o nível de contorno é 0, ou seja, a *level set* zero (CHEN, 2008). Seja $\phi(x,y,t)$ uma superfície implícita tal que

$$\phi(x,y,t) = \pm d \tag{16}$$

onde (x,y) são pontos no domínio Ω da imagem I e d é a distância entre (x,y) e a *level set* zero. O sinal de d é positivo se (x,y) está fora do contorno e negativo se está dentro. Nesse caso, a curva de interesse é representada pelos pontos (x,y) para os quais $\phi(x,y,t) = 0$.

Para evoluir (16) no tempo usamos a regra da cadeia:

$$\begin{aligned} \phi_t + \phi_x x_t + \phi_y y_t &= 0 \\ \phi_t + (x_t, y_t) \cdot \nabla(\phi) &= 0 \end{aligned} \tag{17}$$

Fazendo $(x_t, y_t) = \vec{n} + \vec{s}$ onde $\vec{n}$ é o vetor normal a frente no ponto (x,y) e $\vec{s}$ é um vetor arbitrário qualquer. podemos reescrever o sistema (17) como:

$$\begin{aligned} \phi_t + (\vec{n} + \vec{s}) \cdot \nabla\phi &= 0 \\ \phi_t + \vec{n}\nabla\phi + \vec{s} \cdot \nabla\phi &= 0 \\ \phi_t + V_n|\nabla\phi| + \vec{s} \cdot \nabla\phi &= 0 \end{aligned} \tag{18}$$

Onde V_n é um escalar que controla a velocidade com que a superfície $\phi(x,y,t)$ se move na direção normal e $\vec{s}$ representa uma força que dita direção e velocidade de evolução da superfície. Quando $\phi(x,y,t) = 0$ se aproxima de detalhes na imagem, V_n cai para zero e $\vec{s}$ age de forma a prender o contorno nas bordas. O problema se resume a encontrar $(V_n)_{ij}$ e s_{ij} de tal forma a garantir que a superfície inicial convirja para o contorno desejado com curvatura suave.

Para resolver (18) em todo $(x,y) \in \Omega$, adotando $\phi(x,y,n\Delta t) = \phi^n_{xy}$ onde Δt representa o passo, temos:

$$\begin{aligned} \frac{\phi^{n+1}_{xy} - \phi^n_{xy}}{\Delta t} + (V_n)_{xy}|\nabla\phi^n_{xy}| + \vec{s}_{xy} \cdot \nabla\phi^n_{xy} &= 0 \\ \phi^{n+1}_{xy} = \phi^n_{xy} - \Delta t \left[(V_n)_{xy}|\nabla\phi^n_{xy}| + \vec{s}_{xy} \cdot \nabla\phi^n_{xy}\right] & \end{aligned} \tag{19}$$

Resolver numericamente a equação diferencial parcial dada por (19) é um problema de valor inicial e, dependendo de como for definido $\vec{s}$ e V_n, o contorno pode ou não convergir rapidamente para a região de interesse (CHEN, 2008). De forma geral, o esforço computacional para resolver (19) é grande e a escolha de seus parâmetros para trabalhar com as imagens que são foco deste trabalho provavelmente requer uma tese dedicada a isso. Contudo, no LAPIMO, o método foi implementado no MATLAB e os nódulos resultantes foram separados em uma pasta. Os resultados do programa utilizando esses nódulos são apresentados no Capítulo 5.

2.5 Cálculo das densidades de nível de cinza

Depois de segmentadas as imagens, por *Level set* ou utilizando o limiar dado pelo método de Otsu, as densidades do nódulo e da mama são calculadas.

Para calcular as densidades, seja Ω o domínio da imagem de dimensão $m \times n$ e $I_M : \Omega \mapsto \mathbb{R}$ uma imagem mamográfica em escala de cinza, temos segmentada I_M em duas regiões de domínio Ω_f para o fundo e Ω_n para o nódulo. Para cada ponto $s = (i, j) \in I_M$ temos:

$$B_{ij} = \begin{cases} 1 & \text{região de interesse} \\ 0 & \text{caso contrário} \end{cases} \quad i = 1,2,...m; \;\; j = 1,2,...n$$

e portanto as densidade dos níveis de cinza do nódulo e do fundo são, respectivamente,

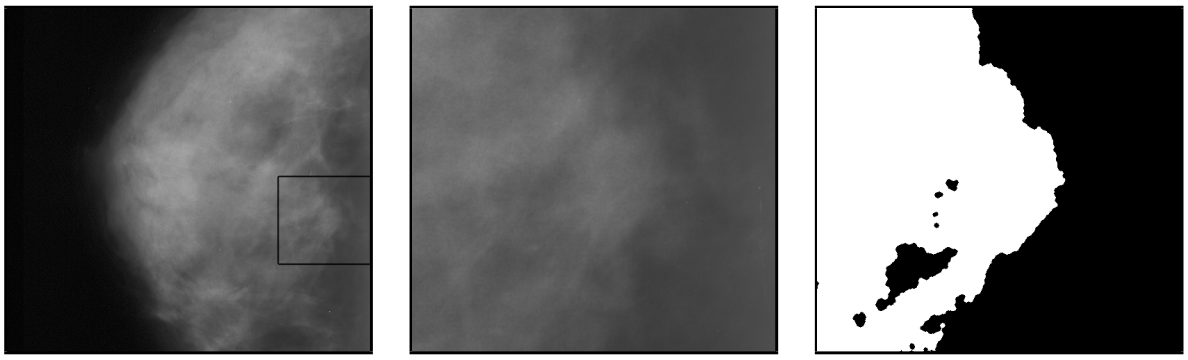
$$\mu_n = \frac{\int\limits_{\Omega_n} I_M(s)B_n(s)ds}{\int\limits_{\Omega_n} 1ds} \;\; , \;\; \mu_f = \frac{\int\limits_{\Omega} I_M(s)B_f(s)ds}{\int\limits_{\Omega} 1ds} \tag{20}$$

3 Estudo Preliminar

3.1 Introdução

Para se investigar a relação entre o nível de cinza e a classificação do nódulo foi implementado um programa simplificado no MATLAB, que foi escolhido pois possui *toolbox* de processamento de imagens e também de aprendizado supervisionado (*Classification Learner*), este último utilizado para auxiliar na escolha de um sistema de classificação adequado. Além disso, o MATLAB torna relativamente simples estudar a performance de um sistema classificador binário pois possui funções prontas para traçar curvas ROC e efetuar predições desses sistemas (SLABY, 2007).

Para o cálculo das densidades do nódulo e da mama, o programa implementado no MATLAB precisa de três entradas: a mamografia completa, a imagem do nódulo e a imagem binarizada do nódulo. Um exemplo das entradas é mostrado na Figura 2 abaixo.



(a) Mama esquerda (LEFT_CC) (b) Imagem contendo apenas nódulo (c) Imagem binarizada do nódulo

Figura 2: Exemplo de entrada para o programa do MATLAB

Nesta etapa foi utilizada uma amostra com número relativamente pequeno de imagens. No total foram 230 imagens das quais 34 são de casos benignos e 196 de casos malignos.

3.2 Resultados

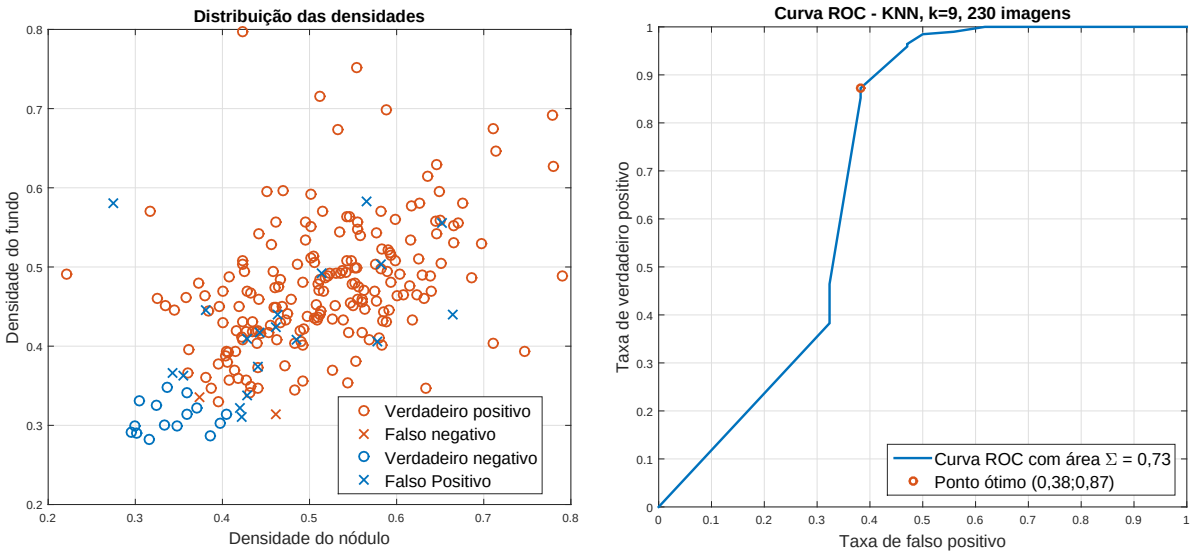
Nesta seção são apresentados os resultados para ambos os sistemas de classificação, o KNN e a razão das densidades do nódulo e da mama. A amostra foi escolhida de tal forma que as mamografias possuíam um par de imagens correspondentes ao nódulo original e ao nódulo binário. Nesta etapa do desenvolvimento do programa principal, as mamografias que se encaixavam nesse critério eram em sua maioria do caso maligno.

3.2.1 KNN

Os resultados referentes ao sistema de classificação KNN podem ser vistos nas Figuras 3a e 3b. Eles foram obtidos para o algoritmo KNN com $k = 9$ (obtido após algumas tentativas, verificando qual valor obtinha a maior taxa de acerto) e métrica euclidiana, ou seja, a distância entre dois pontos é dada por:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{21}$$

A Tabela 1 detalha os resultados obtidos nessas condições.



(a) Distribuição dos pontos (μ_n, μ_f). (b) Curva ROC

Figura 3: Resultados obtidos com o algoritmo KNN no MATLAB

Resultados Figura 3b				
Tamanho da amosta	Sensibilidade (S)	Especificidade (E)	Área sob a curva ROC (Σ)	Acurácia
230	0,8724	0,6176	0,7090	0,8347

Tabela 1: Tabela com resultados da Figura 3b

Os resultados acima mostram que o KNN obteve acurácia de 83,47% porém, como já mencionado, é preciso tomar cuidado ao interpretar a acurácia. Nesse estudo preliminar, como é possível notar na Figura 3a, a amostra possui muitos casos com a existência de sinal (maligno) e portanto o resultado da acurácia pode estar distorcido.

3.2.2 Limiar pela densidade relativa μ_n/μ_f

Calculados μ_n e μ_f por (20), é possível traçar a curva ROC da Figura 4.

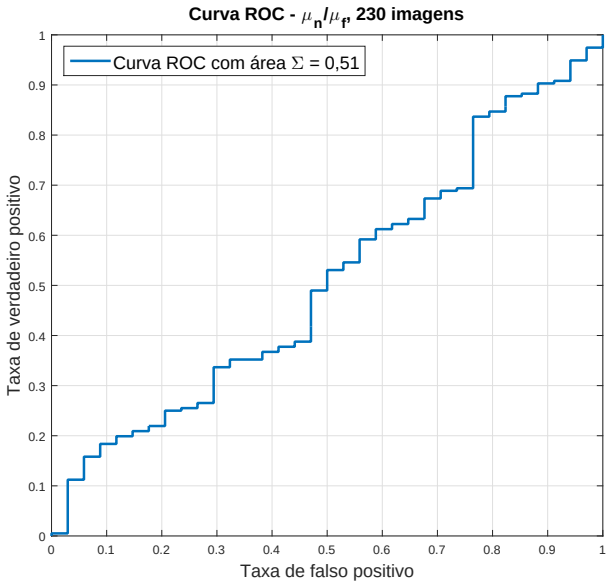


Figura 4: Curva ROC variando a razão μ_n e μ_f

A Curva ROC obtida mostra que a razão das densidades pode não ser um bom sistema de classificação porém, ainda é cedo para descarta-lo pois a amostra de imagens usada possui muitos casos malignos e poucos benignos, o que pode gerar resultados inconsistentes.

3.2.3 Conclusão

Levando em conta apenas os resultados obtidos com o classificador KNN, ou seja, analisando as densidades dos nódulos e dos fundos separadamente, é plausível assumir que exista uma relação entre a densidade dos níveis de cinza e a classificação dos nódulos uma vez que a sensibilidade, especificidade e acurácia tiveram bons resultados.

O classificador pela razão μ_n/μ_f não apresentou resultados satisfatórios, talvez pela presença de ruído e não-homogeneidade da luminosidade dos *pixels* da radiografia ou pelo fato de haver muitos mais casos malignos do que benignos. Ainda assim, os dois classificadores serão usados nos resultados do programa final a fim de serem comparados novamente já que a amostra disponível nesta etapa não é suficiente para analisar corretamente os dois sistemas de classificação.

4 Programa Principal

4.1 Introdução

O fato de Java ser uma linguagem multiplataforma foi determinante em sua escolha. Além disso, a linguagem torna a tarefa de criar interfaces gráficas com o usuário relativamente simples.

Em resumo, o programa deve ser capaz de construir uma estrutura de dados contendo, para cada mamografia, as seguintes informações:

- Classificação do nódulo, benigno ou maligno. (Informação lida de arquivo de *overlay*, processo detalhada na seção 4.6.6 deste capítulo)
- Densidade do nódulo. (Calculado pelo programa)
- Densidade do fundo. (Calculado pelo programa)

Depois, esse conjunto de dados é usado para um treinamento heurístico dos dois sistemas de classificação explorados no Capítulo 3, o KNN e a razão das médias do nódulo e da mama. Por último, calcula-se a taxa de acerto para classificação dos nódulos pela densidade de nível de cinza.

Um dos problemas encontrados foi o tempo de processamento para cada imagem. Por se tratar de imagens de 16 bits e de alta resolução (12 Mega *Pixels*) o custo computacional é grande, tanto de memória quanto de tempo de processamento (SCHILDT, 2013). A solução encontrada foi a implementação de *multithreading*, tornando o programa capaz de segmentar e processar várias imagens ao mesmo tempo.

4.2 Diagrama de Arquitetura

Antes de se desenvolver um *software*, deve-se projetar a estrutura do sistema. Esse processo é denominado modelagem e um dos métodos para realiza-lo é através do Diagrama de Arquitetura.

Para isso, considera-se a interface com o usuário, processamento de entrada, funções de processo e processamento de saída.

Esse diagrama pode ser visto na Figura 5 ao lado e permite visualizar como o programa se relaciona com o ambiente, ou seja, ele mostra o relacionamento entre o programa básico e o usuário, entradas e saídas (PRESSMAN, 2011).

Um maior detalhamento do **Sistema de segmentação e cálculo da densidade de níveis de cinza** pode ser visto no Diagrama de Fluxo de Dados, na seção 4.3.

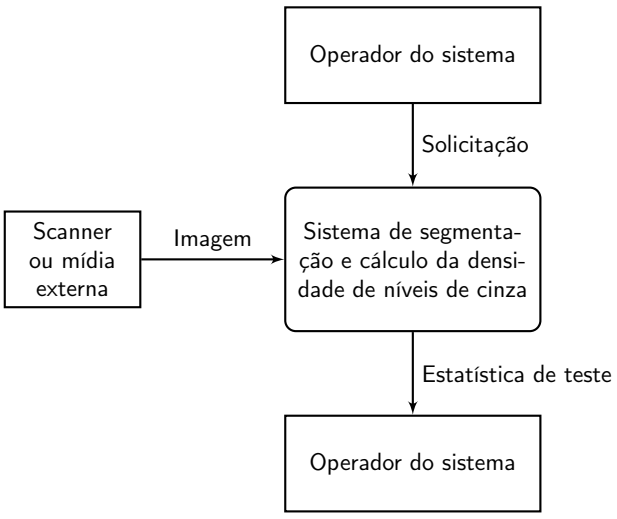


Figura 5: Diagrama de Arquiteturas

4.3 Diagrama de Fluxo de Dados

O diagrama de fluxo de dados é um modelo que procura representar o programa como uma rede de processos e permite visualizar o fluxo de informação (PRESSMAN, 2011). O diagrama da Figura 6 é

referente ao programa desenvolvido neste trabalho e detalha os processos para construção da estatística de teste, que se resume a:

- Cálculo do limiar
  O *threshold* é calculado via método de Otsu (OTSU, 1979).
- Binarização da imagem
  Utiliza-se o *thresdhold* calculado na etapa anterior para a binarização da imagem. No caso do nódulo, quando a área ocupada por ele for maior que um certo valor (o programa permite o controle deste parâmetro) o *threshold* é alterado, levando em conta um valor de *offset* (também pode ser alterado pelo usuário) de forma que o nódulo seja bem segmentado. Esse processo é detalhado no item 4.6.2 deste capítulo.
- Remoção de bordas
  Se aplica apenas para o fundo da mamografia. Nesta etapa as bordas brancas são removidas e o usuário pode ativar, desativar ou ajustar essa função.
- Segmentação
  Um algoritmo conhecido como *Flood Filling* (SCHILDT, 2013) é utilizado tanto para segmentar a mama e o nódulo. O detalhamento desta etapa se encontra no item 4.6.4 deste capítulo.
- Definição da ROI
  O arquivo de *overlay* é carregado e as coordenadas da ROI são definidas na imagem.
- Cálculo das densidades
  Com a ROI definida, as densidades do nódulo e do fundo são calculadas. Um par ordenado (μ_n, μ_f) com essas densidades é guardado em um arquivo para mais tarde ser usado na definição do sistema de classificação.

O Diagrama de Fluxo de Dados a seguir ilustra de forma simples e concisa as diversas etapas até o cálculo das médias de nível de cinza de uma mamografia, que corresponde a saída do programa.

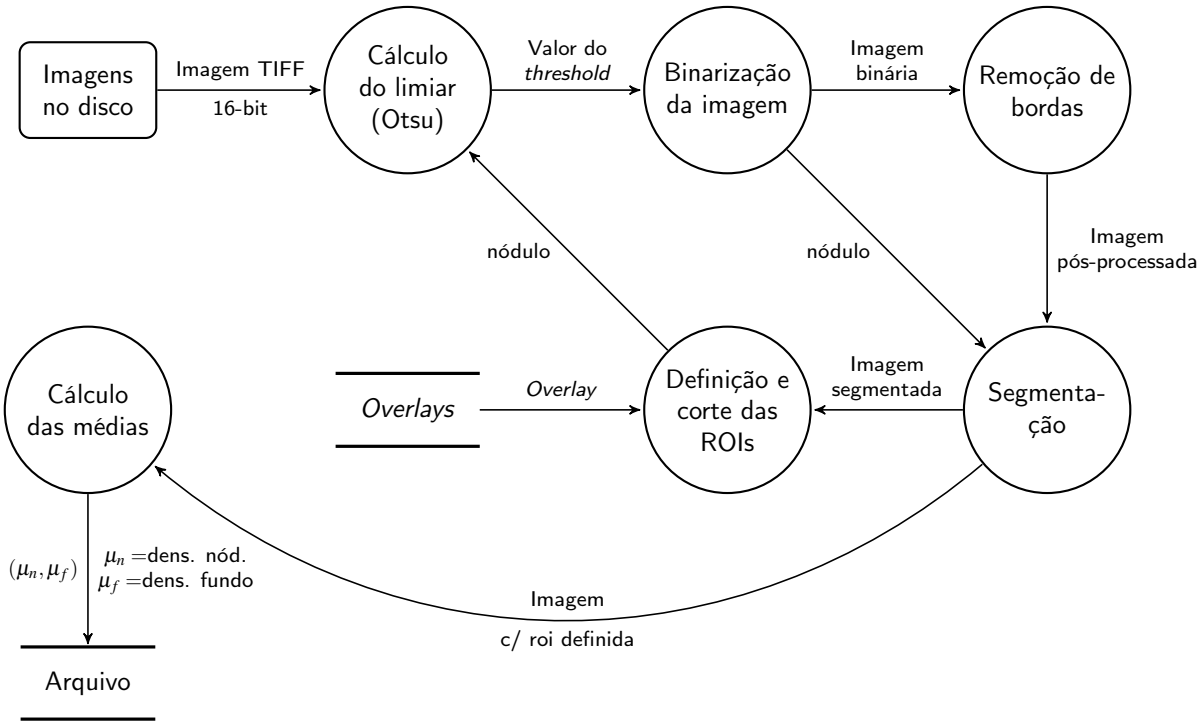


Figura 6: Diagrama de Fluxo de Dados

4.4 Interface gráfica

Para o facilitar a interação do usuário com o programa, foi criada uma interface gráfica simples para o programa. Dentro dela é possível ajustar diversos parâmetros referentes a segmentação das imagens e às saída do programa. As Figuras 7 e 8 mostram a interface com o usuário do programa e menus de configurações.

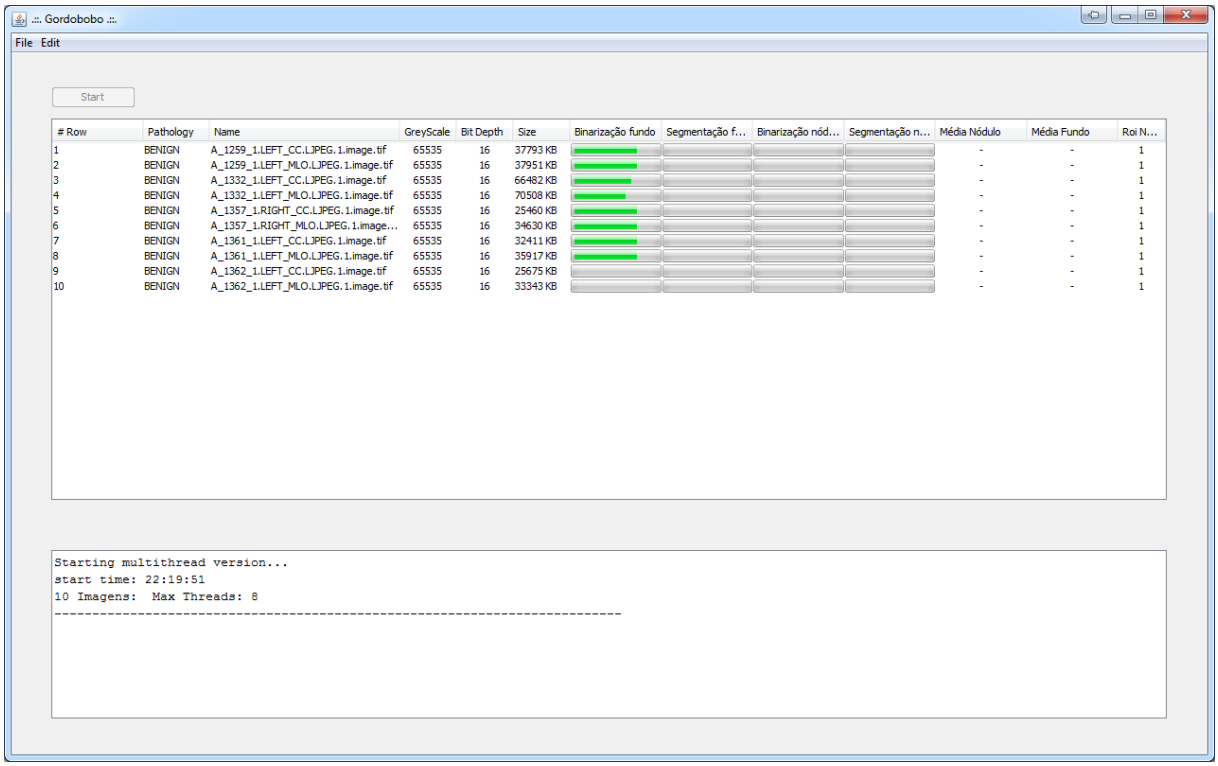


Figura 7: Interface gráfica principal

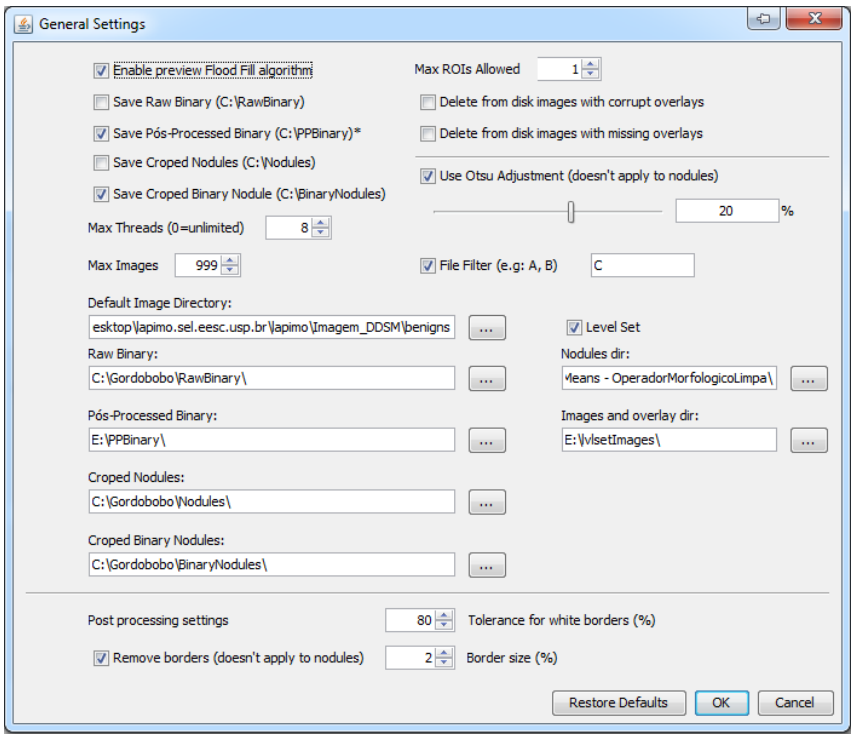


Figura 8: Edit -> Settings

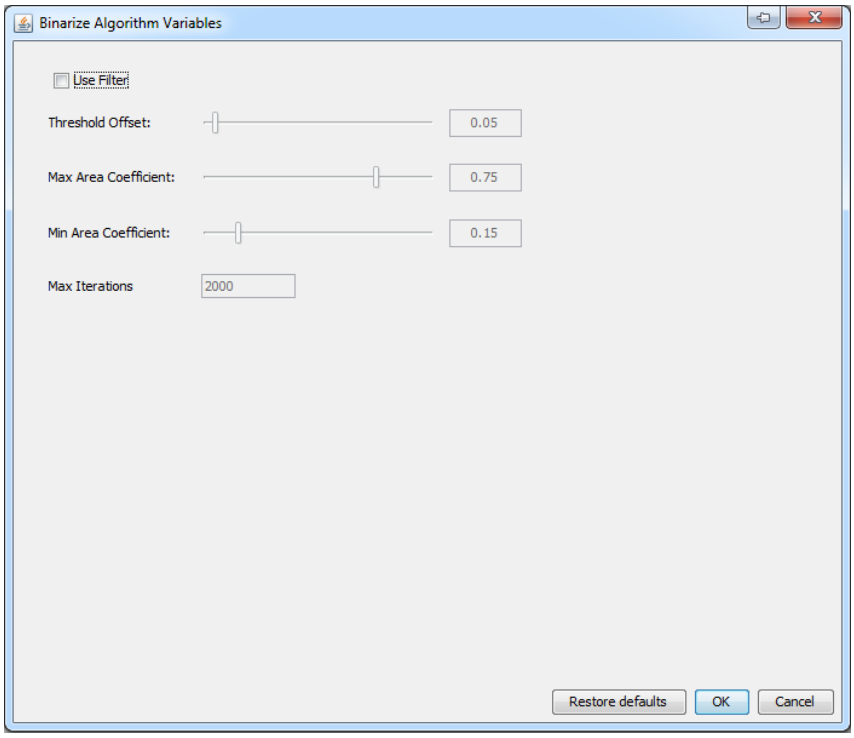


Figura 9: Edit -> Binarize filter

4.5 *Multithreading*

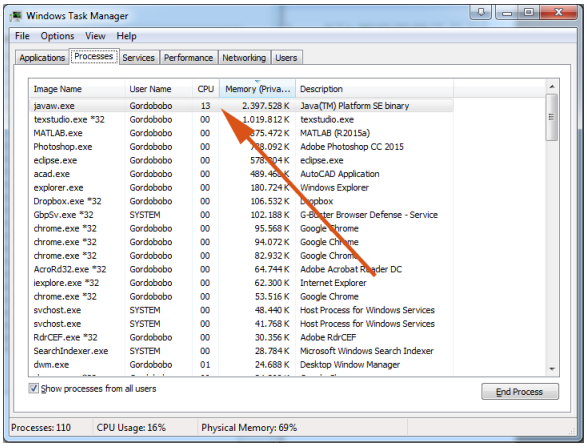
Na primeira versão do programa não havia suporte a *multithreading* e tudo era feito em um único processo, uma imagem de cada vez. Contudo, o processamento de cada imagem é independente e além disso Java tem suporte nativo a *multithreading*, facilitando assim a sua implementação. O resultado foi um programa que permitiu a utilização de todos os processadores lógicos da máquina, o que consequentemente resultou em uma economia substancial de tempo. Na prática, como pode ser visto na Figura 12, o programa demora aproximadamente cinco vezes menos tempo em relação a versão *singlethread* do programa.

No Menu *Settings* (Figura 8), o usuário pode limitar o número de *threads* que o programa pode alocar.

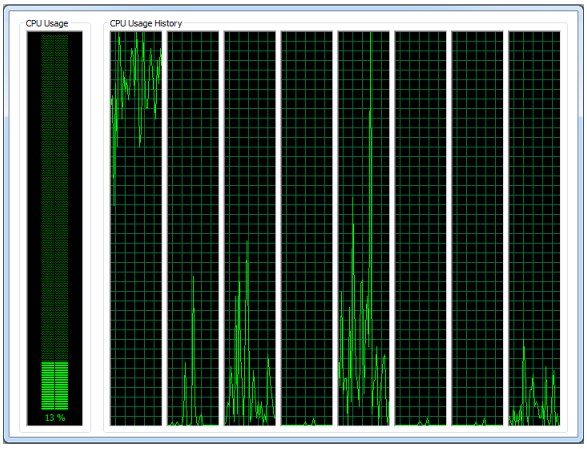
A configuração da máquina na qual foi executado o programa é:

- CPU: Mobile QuadCore Intel Core i7-3630QM, 2400 Hz (24x100)
- MB: LG A560-T.BG77P1 (Versão da BIOS QLGALQ122)
- RAM: 16 GB (2x 8 GB) 2133 MHz DDR3 Corsair Vengeance (CMSX16GX3M2B2133C1)
- Disk: Samsung SSD 250 GB 840 EVO
- Chipset North Bridge: Intel Ivy Bridge-MB IMC
- Chipset South Bridge: Intel Panther Point HM77
- GPU: nVIDIA Geforce GT 640M (GK107M)
- OS: Microsoft Windows 7 Ultimate (Versão 6.1.7601.18869)
- Java: Version 8 Update 60 (build 1.8.0-60-b27)

As Figuras 10 e 11 a seguir mostram o uso da CPU pelas versões *singlethread* e *multithread*, respectivamente, do programa.

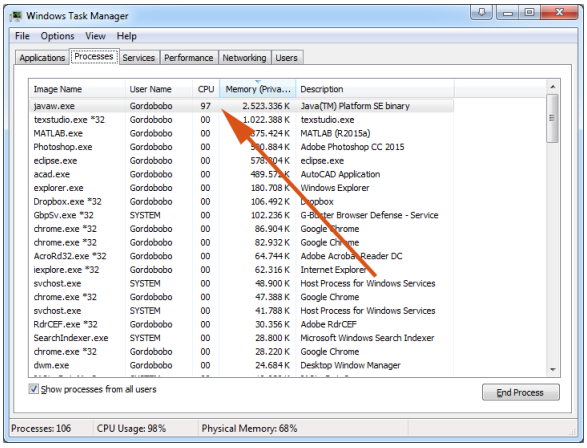


(a) Gerenciador de tarefas: Processos

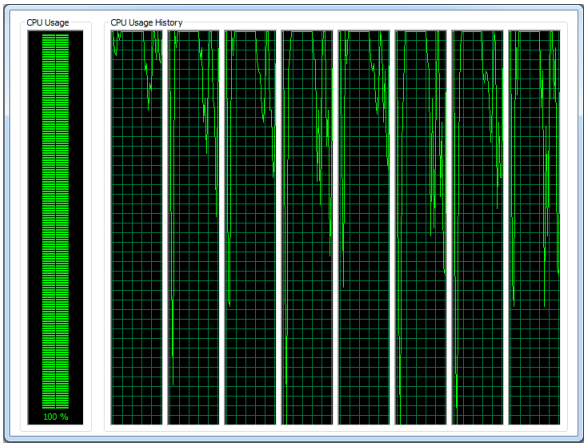


(b) Gerenciador de tarefas: Performance

Figura 10: Versão *singlethread* do programa



(a) Gerenciador de tarefas: Processos



(b) Gerenciador de tarefas: Performance

Figura 11: Versão *multithreading* do programa

O ganho por paralelismo de um código é dado pela Lei de *Amdahl* (SCHILDT, 2013). Seja f a porcentagem do programa que não pode rodar em paralelo, ou seja, é estritamente serial e n o número de processos, o ganho máximo de performance é:

$$S(n) = \frac{1}{f + \left(\frac{1-f}{n}\right)} \tag{22}$$

O gráfico da Figura 12 ilustra o ganho máximo de performance de um programa $(1-f)\%$ paralelo utilizando a lei de *Amdahl* e tanto os asteriscos como a curva pontilhada representam o ganho de performance do programa desenvolvido neste trabalho.

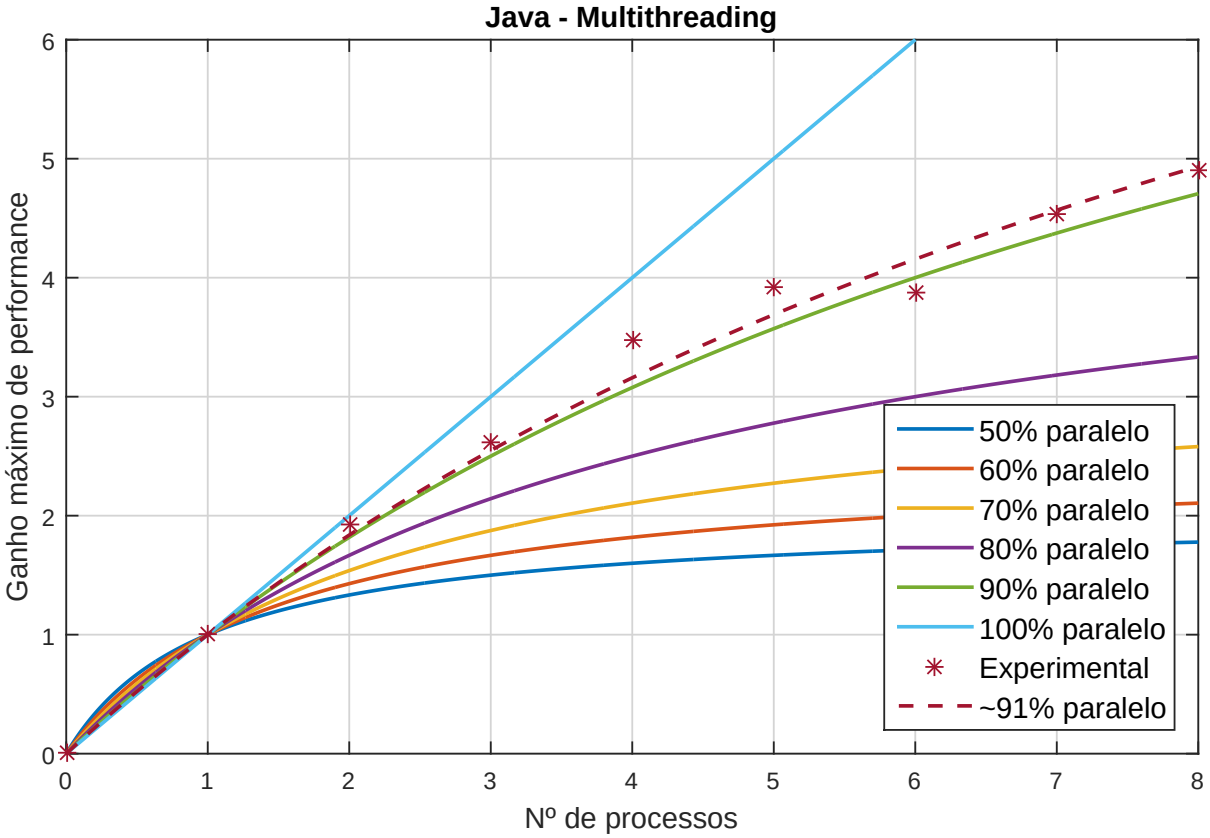


Figura 12: Gráfico de $S(n)$ para $f = 0.5$; $f = 0.4$; $f = 0.3$; $f = 0.2$; $f = 0.1$; $f = 0$ e f experimental

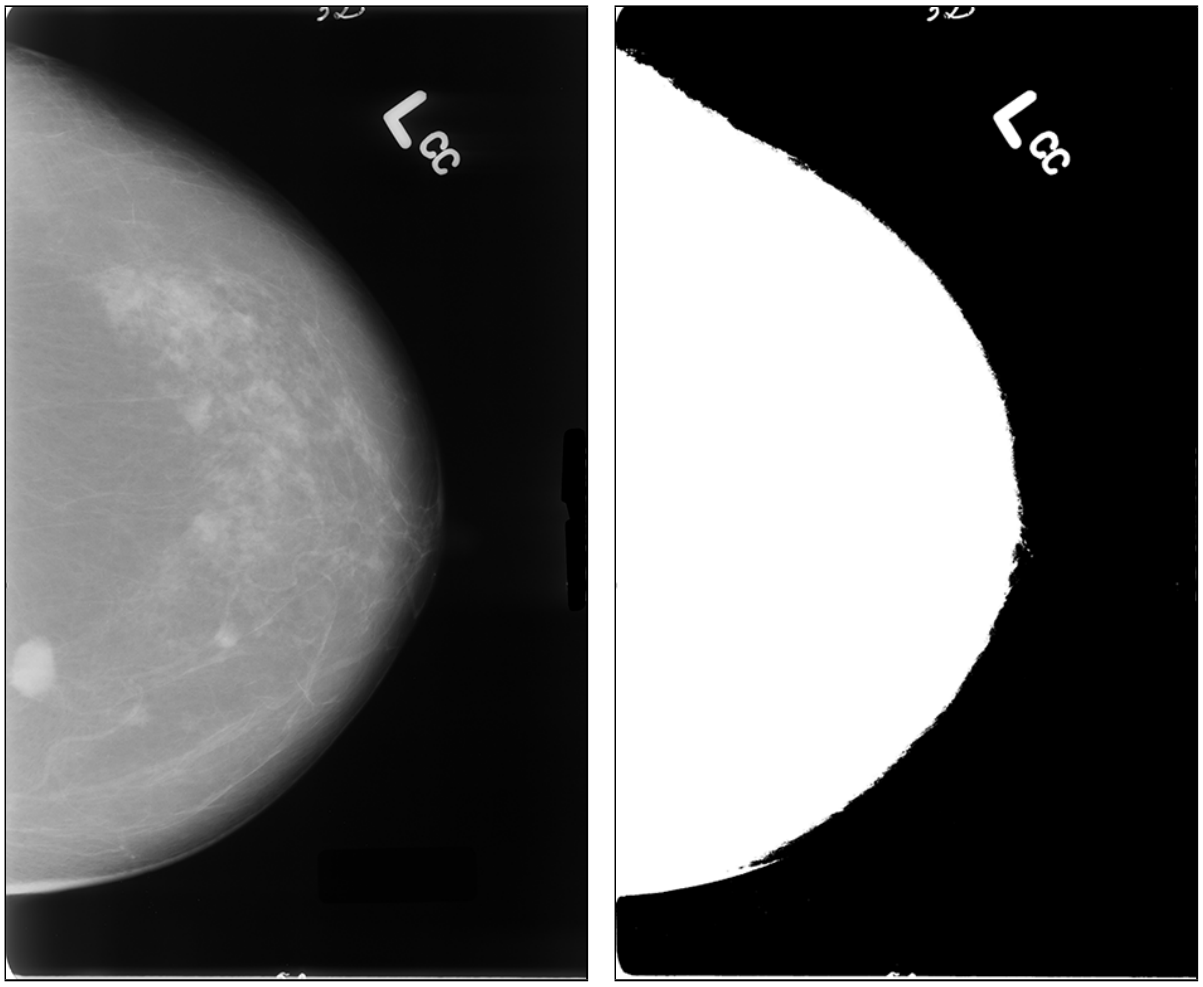
Na figura acima, os pontos experimentais foram obtido em uma máquina com 8 processadores lógicos e é possível notar claramente no gráfico que o ganho máximo ocorre justamente para 8 processos.

A curva pontilhada é resultado da interpolação dos pontos obtidos experimentalmente medindo-se o tempo de execução do programa quando este rodava limitando-se o número de *threads* que podiam ser criados entre 1 até 8. Essa curva foi interpolada pelo método de mínimos quadrados - algoritmo de Levenberg-Marquardt (MORÉ, 1978).

4.6 Processos implementados no programa

4.6.1 Binarização da mama

Com o *threshold* calculado pelo método de Otsu (OTSU, 1979) descrito na seção 2.4.1 do Capítulo 2, a imagem passa pela binarização e em seguida é convertida de 16 bits (0 e 65535) para 8 bits (0 e 255). Esse procedimento pode ser visto na Figura 13.



(a) Mamografia original, TIFF de 16 bits (b) Mamografia depois da binarização, TIFF de 8 bits

Figura 13: Binarização utilizando o limiar de Otsu

Existe, no menu *Settings* do programa, campo onde o usuário pode ajustar o *threshold* calculado por Otsu de modo que a mama seja melhor segmentada em casos onde esse limiar provoca muita perda do fundo da mama. Nesta etapa, a imagem binária ainda possui os *tags*, imperfeições nas bordas, ruído e objetos no fundo preto que não pertencem a mama. Todas essas imperfeições são removidas antes do cálculo da densidade dos níveis de cinza do fundo.

4.6.2 Binarização do nódulo

A binarização do nódulo é análoga à do fundo e um novo *threshold* é calculado levando em consideração apenas a região do nódulo. Porém, em alguns casos, foi verificado que o nódulo binário resultante estava totalmente branco ou totalmente preto. Isso se deve ao fato de que, nesses casos, ou o nódulo não está bem definido dentro da ROI ou o nódulo toma praticamente toda a ROI. Para resolver esse problema o programa calcula a área branca do nódulo após a binarização e se for maior ou menor que um certo valor (que pode ser ajustado pelo usuário dentro do programa), o *threshold* é ajustado. O *offset* de ajuste do *threshold* também pode ser ajustado pelo usuário (A Figura **??** mostra a janela do programa onde é possível ajustar esses valores). Por exemplo, na Figura 14, o programa é ajustado para:

- Aumentar o *threshold* em 20% enquanto a razão de *pixels* brancos for 50% ou mais.
- Diminuir o *threshold* em 20% enquanto a razão de *pixels* brancos for 10% ou menos.

(a) Nódulo original (b) Binarização, *threshold* original (c) Binarização, *threshold* + 20%

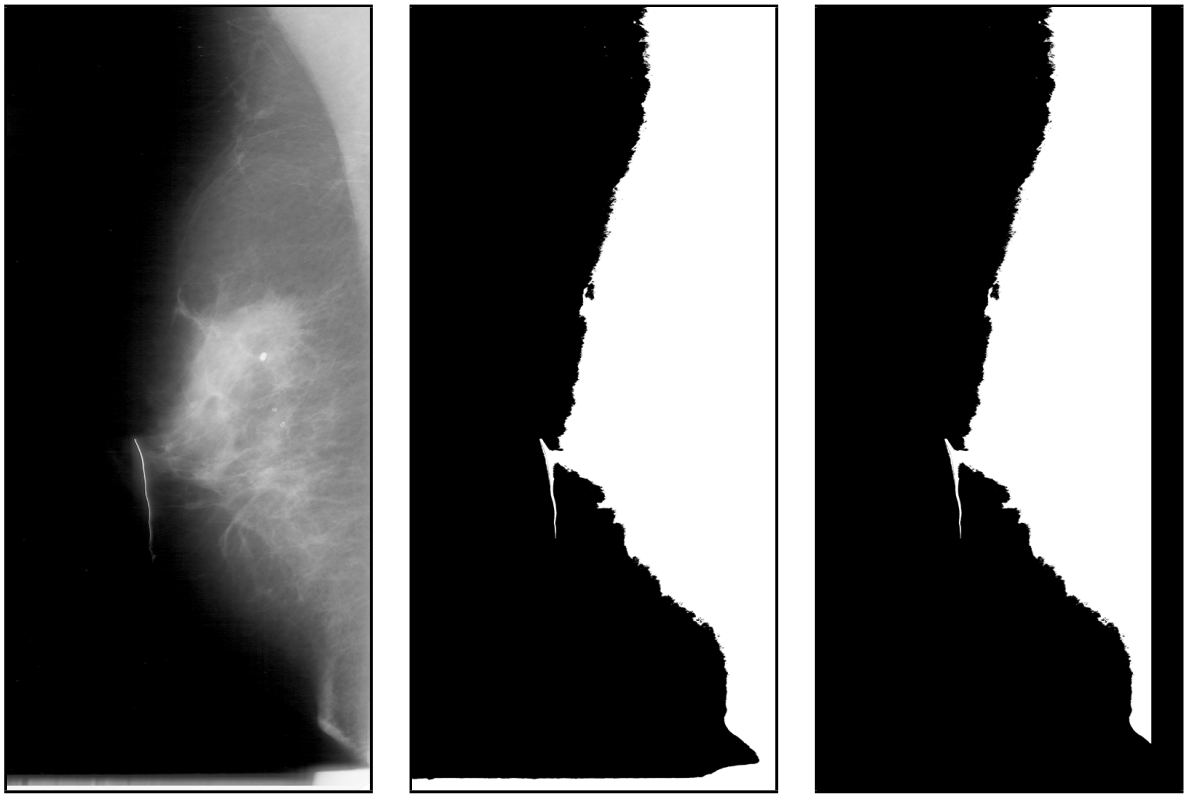
Figura 14: Resultado do ajuste automático de *threshold*

Os parâmetros para o controle desse filtro são:

- *Threshold offset*: Ajuste do *threshold* de Otsu
- *Max area Coefficient*: Área máxima que o nódulo pode assumir
- *Min area Coefficient*: Área mínima que o nódulo pode assumir

4.6.3 Pós-processamento (remoção das bordas)

O pós-processamento ocorre apenas para o fundo e se resume a eliminação de bordas brancas. A Figura 15 compara uma mamografia antes e depois de passar por esta etapa.



(a) Mamografia original (b) Antes do pós-processamento (c) Depois do pós-processamento

Figura 15: Remoção de bordas

Essa etapa foi implementada porque os extremos das mamografias podem estar esbranquiçados e essa região não corresponde a mama. O programa varre as linhas e colunas correspondentes a esses

extremos e calcula a razão de *pixels* brancos, se essa razão for maior que um certo valor essas linhas ou colunas são removidas (tornam-se pretas). Os parâmetros para esse filtro são:

- Tamanho das bordas: Porcentagem da imagem que é considerada borda
- Tolerância: Porcentagem da linha ou coluna que pode ser branca

Ambos esses parâmetros podem ser ajustados pelo usuário dentro do programa, como pode ser visto na Figura **??**. Depois dessa etapa a imagem binária ainda pode conter *tags* e objetos que não pertençam à mama mas somente na etapa seguinte que a mama é isolada e segmentada.

Cada *thread* responsável pela binarização dá início a outro que por sua vez é encarregado pelo pós-processamento.

4.6.4 Segmentação da mama

Para segmentar a mama, utiliza-se um método chamado *Flood Fill*. Esse algoritmo funciona a partir de um ponto inicial dentro do objeto a ser segmentado, procurando por *pixels* brancos para preencher.

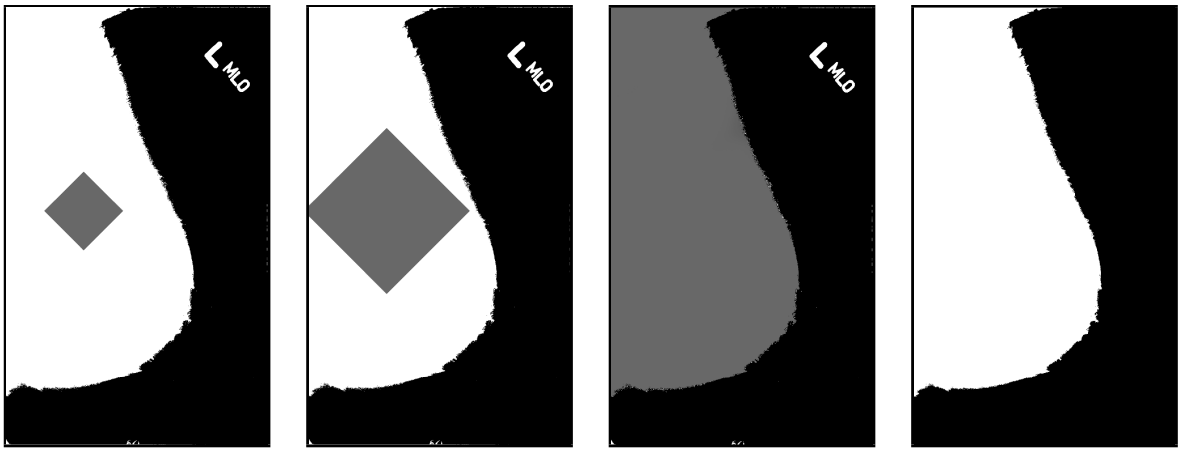
Esse algoritmo pode ser implementado recursivamente ou iterativamente, entretanto um **Ambiente de Tempo de Execusão Java** (JRE) suporta, no máximo, aproximadamente dez mil chamadas recursivas do *Flood Fill* antes que a memória alocada para a pilha acabe. Isso só permite trabalhar com imagens pequenas, com resoluções próximas de 200×200 *pixels*. Isso ocorre, porque em Java e linguagens como C e C++, variáveis locais são automaticamente armazenadas na *call stack* (pilha de chamadas) a cada chamada e recuperadas da pilha quando a chamada termina (BURGER e BURGE, 2009).

Por esse motivo foi implementada a versão iterativa, que embora mais complexa, permite alocar toda a memória disponível e portanto trabalhar com imagens muito maiores.

O método iterativo por sua vez possui duas variantes:

- *depth-first*: Implementa sua própria *stack* (pilha) com muito mais memória do que a pilha de chamadas padrão do método recursivo (BURGER e BURGE, 2009).
- *breadth-first*: Usa estrutura de dados do tipo *queue* (fila) ao invés de *stack* (pilha) para armazenar as coordenadas dos *pixels* ainda não visitados (BURGER e BURGE, 2009).

A variante implementada foi a segunda, *breadth-first* que utiliza o conceito de *LinkedList* e, devido ao método de exploração de *pixels* vizinhos, utiliza muito menos memória que a versão *depth-first*. A Figura 16 abaixo ilustra a segmentação da mama pelo algoritmo *Flood Fill - breadth-first*.



(a) n iterações (b) $m > n$ iterações (c) completo (d) mama segmentada

Figura 16: Segmentação pelo algoritmo *Flood Fill - breadth-first*

O pseudocódigo para esse algoritmo e sua implementação em Java podem ser vistos no **Anexo B**.

Esse algoritmo depende de um ponto inicial que deve necessariamente estar dentro da região a ser segmentada. Como a mama é uma região conexa e é o maior objeto de uma mamografia, o centro de

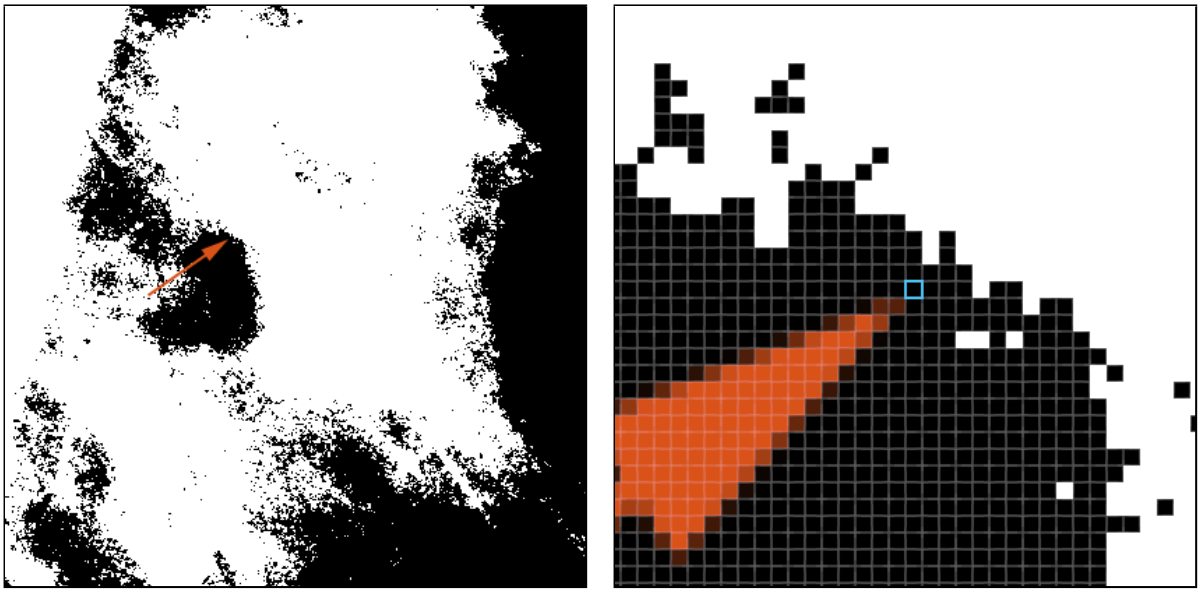
massa está sempre dentro dela. Para o cálculo do centro de massa temos:

Seja B uma imagem binária onde $B(i,j) = 1$ corresponde a região de interesse e 0 caso contrário, então:

$$(\bar{x},\bar{y}) \quad | \quad \bar{x} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{M} jB(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{M} B(i,j)}; \quad \bar{y} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{M} iB(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{M} B(i,j)} \tag{23}$$

4.6.5 Segmentação do nódulo

A segmentação do nódulo é feita de forma análoga à da mama inteira porém, como o nódulo na imagem binária nem sempre é uma região conexa, o centro de massa dado por (23) pode não estar dentro do nódulo e portanto, antes do cálculo do centro de massa começar, é feita uma operação morfológica de abertura na imagem (GONZALES, 2007) com uma cruz 3×3 como elemento estruturante. Se mesmo depois do processo de abertura, o centro de massa ainda não estiver dentro da região branca (correspondente ao nódulo), a imagem passa por processos de dilatação (GONZALES, 2007) com o mesmo elemento estruturante anterior, até que seu centro de massa pertença ao nódulo. As Figuras 17 e 18 mostram o mesmo nódulo, antes de depois dos processos morfológicos, até o centro de massa pertencer a região branca. A seta laranja aponta para o centro de massa.

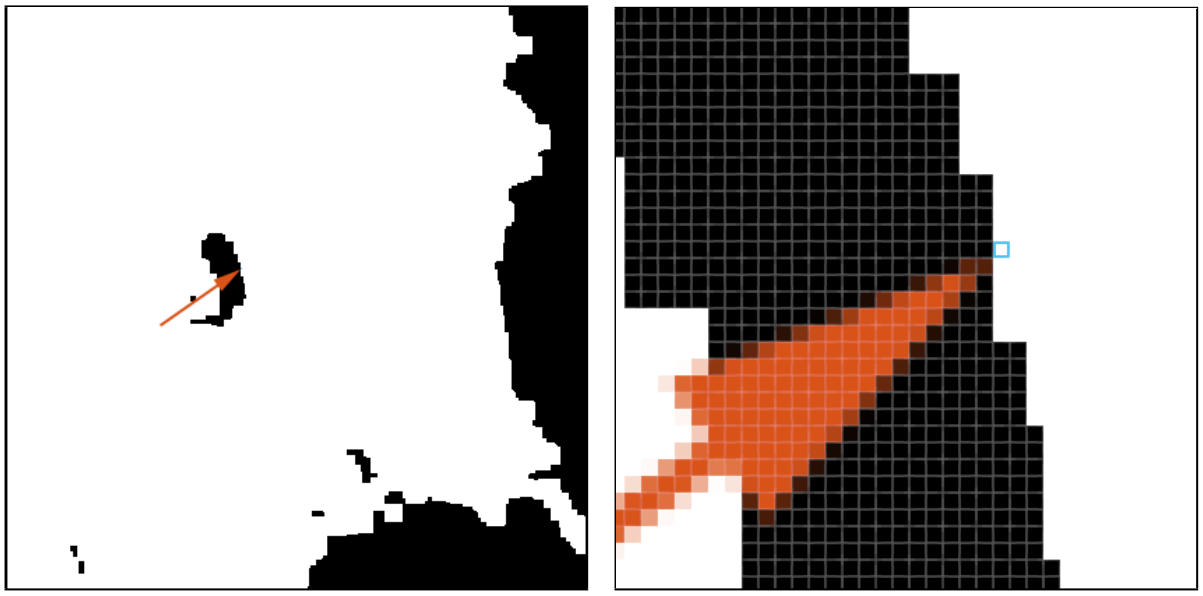


(a) Nódulo binário original, tamanho original

(b) Nódulo binário original, *zoom* 12x

Figura 17: Nódulo antes dos processos morfológicos, centro de massa fora da região branca

Na Figura 17 existem várias regiões desconexas (ilhas de *pixels* brancos) e isso faz com que, em alguns casos, o centro de massa não pertença à região de interesse e portanto o algoritmo *FloodFill* não tem um ponto inicial válido para segmentar o nódulo. Depois de cada processo de dilatação, o centro de massa é recalculado e de novo é checado se o centro de massa pertence à região branca. A Figura 18 mostra o nódulo depois de passar pelos processos morfológicos. É importante enfatizar que isso ocorre apenas para um número pequeno de imagens e, quando ocorre, o número de erosões consecutivas necessárias para ajuste do centro de massa é pequeno (cinco operações no máximo).



(a) Nódulo depois da morfologia, tamanho original (b) Nódulo depois da morfologia, *zoom* 12x

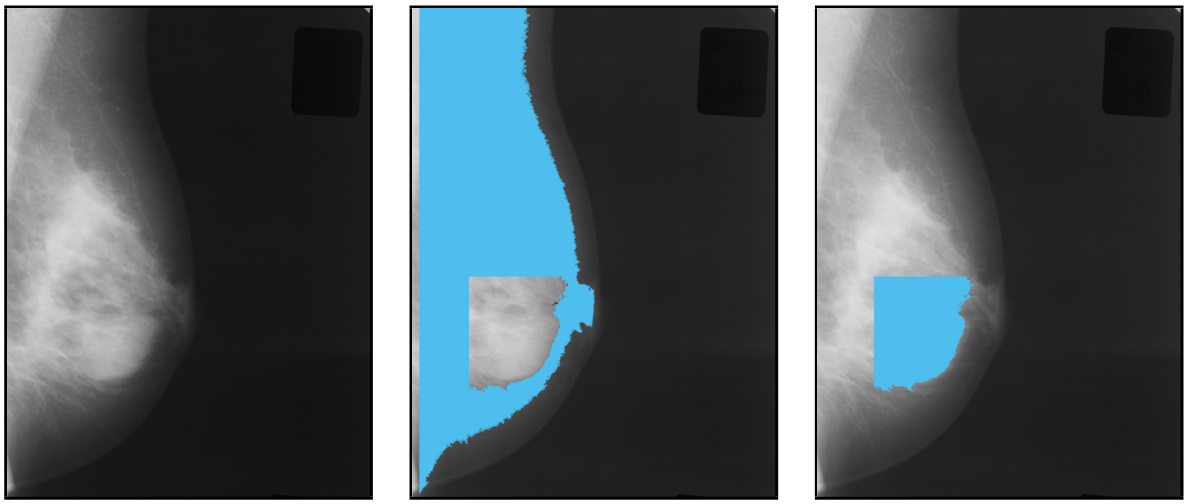
Figura 18: Nódulo depois dos processos morfológicos, centro de massa dentro da região branca

As Figuras 17 e 18 mostram um caso único de um nódulo que precisou ser dilatado 30 vezes, porém a diferença na médias dos níveis de cinza do nódulo original e do nódulo erodido foi insignificante. Assim como a mama, antes de ser segmentado o nódulo passa por processo de abertura para eliminar ruídos e depois de ser segmentado, passa por processo de fechamento para suavizar contornos.

4.6.6 Definição da ROI

Para a definição da região de interesse ROI é carregado um arquivo de *overlay* que contém suas coordenadas. O programa permite ao usuário filtrar mamografias pela quantidade de ROIs que elas possuem e trabalhar, por exemplo, apenas com mamografias que possuam duas regiões de interesse. Por simplicidade, os resultados apresentados no Capítulo 5 foram obtidos utilizando mamografias que possuem apenas uma única ROI.

4.6.7 Cálculo da densidade de níveis de cinza



(a) Mamografia original (b) Pixels usados na média do fundo (c) Pixels usados na média do nódulo

Figura 19: Cálculo da densidade de níveis de cinza

Por simplicidade, no estudo preliminar no Capítulo 3, o cálculo da densidade de níveis de cinza do fundo não excluía a região do nódulo, inserindo assim, um erro na média final do fundo. Um dos refinamentos da versão em Java foi justamente excluir da média do fundo, os *pixels* correspondentes ao nódulo, ou seja, no programa principal $\Omega_f \neq \Omega$. Na Figura 19 acima é possível ver, em azul, os *pixels* usados no cálculo das médias do fundo e do nódulo.

4.7 Outras técnicas de segmentação

Foi implementado no programa principal um recurso que permite ao usuário definir uma pasta com nódulos binários, já segmentados. Esse recurso permite testar técnicas de segmentação que foram implementadas no MATLAB ou em outro programa qualquer. Uma vez definida a pasta com os nódulos, o programa procura automaticamente na pasta de imagens (que também é definida pelo usuário) aquelas que correspondem aos nódulos segmentados (baseando-se nos nomes dos arquivos). Desse modo, é possível segmentar apenas o fundo pelo programa e utilizar o nódulo já segmentado no cálculo de sua média. Com a introdução dessa ferramenta foi possível comparar diversas técnicas de segmentação. Esses resultados encontram-se no Capítulo 5, item 5.5.

5 Resultados e Discussões

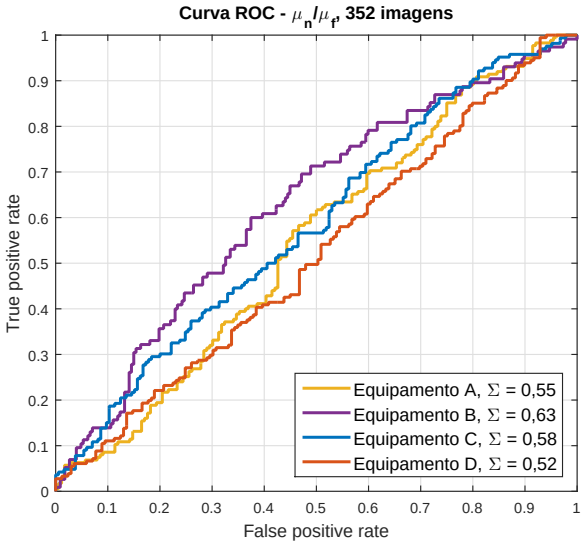
O programa deve, resumidamente, receber uma mamografia como entrada e retornar as densidades dos níveis de cinza do nódulo e da mama, permitindo o ajuste de parâmetros relacionados a segmentação da mama e do nódulo, dentre eles:

- Ajuste fixo do limiar Otsu para a mama.
- Ajuste automático do limiar Otsu para o nódulo.
- Remoção de borda.

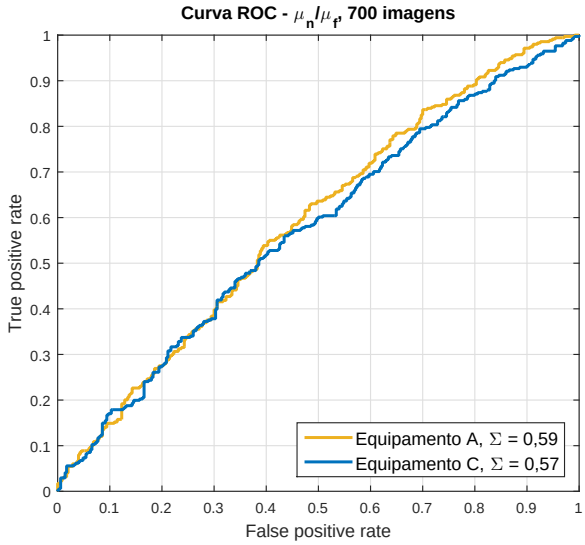
Depois de calculadas as densidades, estas são exportadas para o MATLAB onde é feita a análise da eficácia dos dois sistemas de classificação detalhados no Capítulo 2 (KNN e Razão das densidades μ_n/μ_f). A eficácia é determinada por curvas ROC com classe positiva referindo-se a casos malignos e classe negativa a casos benignos. Os resultados são expostos em tabelas e os melhores são destacados em cinza e foram escolhidos com base na acurácia, dada pela equação (2).

5.1 Diferença entre equipamentos

A Figura 20 mostra resultados do sistema de classificação pela razão das médias, para quatro tipos de escâneres diferentes, cada um com uma amostra de 352 mamografias divididas igualmente em casos benignos e malignos. O tamanho da amostra de 352 imagens foi escolhido pois corresponde ao número máximo de imagens disponíveis para o equipamento **D**. Os esquipamentos **A** e **C** possuem um número grande de imagens (1574 e 730 respectivamente) foram comparadas as performance desses dois escâneres usando amostras maiores, com 700 imagens.



(a) Limiar μ_n/μ_f, 352 imagens



(b) Limiar μ_n/μ_f, 700 imagens

Figura 20: Curvas ROC dos classificadores KNN e Limiar μ_n/μ_f, para diferentes equipamentos

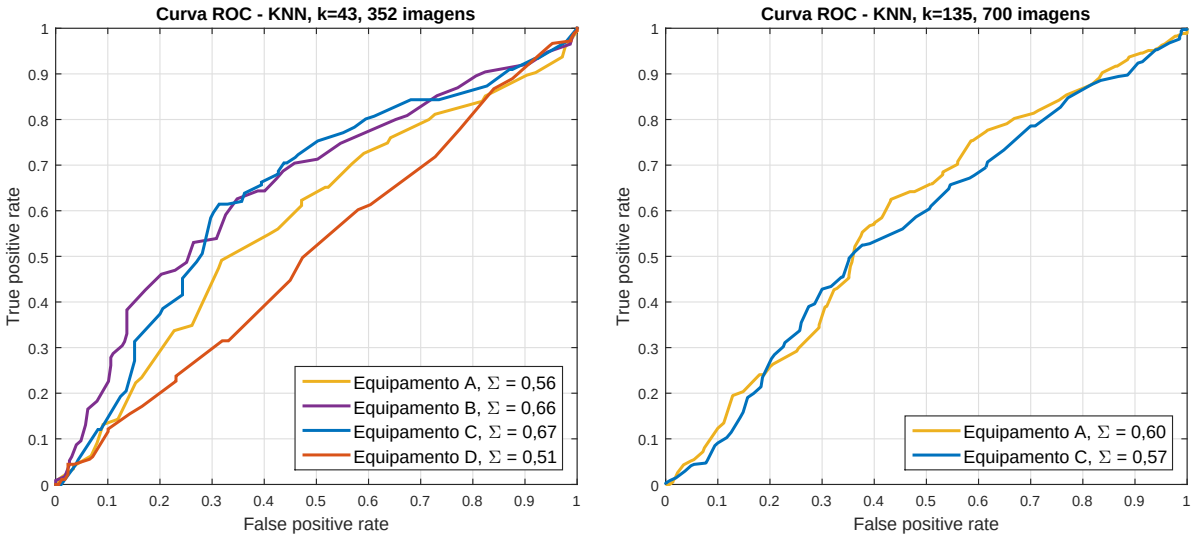
Resultados Figuras 20a e 20b

Equipamento	Tamanho da amosta	Sensibilidade (S)	Especificidade (E)	Área (Σ)	Acurácia
A	352	0,6057	0,5114	0,5455	0,5586
	700	0,5501	0,5943	0,5902	0,5722
B	352	0,7130	0,5110	0,6268	0,6120
C	352	0,5663	0,5351	0,5822	0,5507
	700	0,5601	0,5657	0,5738	0,5629
D	352	0,5414	0,4911	0,5177	0,5163

Tabela 2: Tabela com resultados das Figuras 20a e 20b

Os resultados da Tabela 2 mostram que a acurácia pode vir a variar muito com o escâner utilizado. O escâner **D** possui acurácia de 51,63% enquanto o **B** atinge 61,20%, o que representa uma melhora de 18,53% apenas trocando o equipamento.

A Figura 21 mostra os resultados para o sistema de classificação com o KNN.



(a) KNN $k = 43$, 352 imagens

(b) KNN $k = 135$, 700 imagens

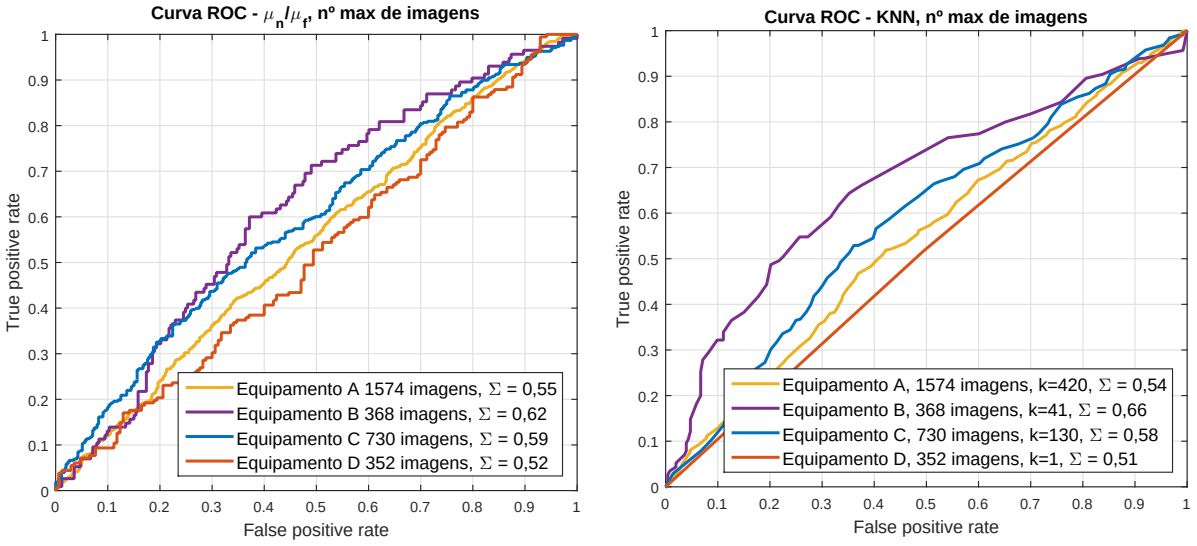
Figura 21: Curvas ROC dos classificadores KNN e Limiar μ_n/μ_f, para diferentes equipamentos

Resultados Figuras 21a e 21b					
Equipamento	Tamanho da amosta	Sensibilidade (S)	Especificidade (E)	Área (Σ)	Acurácia
A	352	0,6114	0,6364	0,5341	0,6239
	700	0,6562	0,5286	0,5990	0,5924
B	352	0,7391	0,5463	0,6543	0,6427
C	352	0,7169	0,5676	0,6694	0,6423
	700	0,5865	0,5429	0,5686	0,5647
D	352	0,5138	0,5621	0,5067	0,5380

Tabela 3: Tabela com resultados das Figuras 21a e 21b

Comparando os resultados da Tabela 3 com os da Tabela 2 é possível ver que o KNN possui performance superior à razão das médias em todos os casos, atingindo o máximo de 64,27% para o equipamento **B**. Além disso, o escâner D possui a pior performance e os equipamento C, mesmo dobrando o número de imagens na amostra, manteve praticamente a mesma acurácia.

A Figura 22 mostra a performance para uma amostra com o máximo número de imagens por equipamento.



(a) Limiar μ_n/μ_f

(b) KNN

Figura 22: Curvas ROC com número máximo disponível de imagens para cada equipamento

Resultados Figura 22a					
Equipamento	Tamanho da amosta	Sensibilidade (S)	Especificidade (E)	Área (Σ)	Acurácia
A	1574	0,5293	0,5392	0,5465	0,5343
B	368	0,6000	0,6285	0,6178	0,6196
C	730	0,5661	0,5597	0,5918	0,5630
D	352	0,5001	0,5500	0,5116	0,5242

Tabela 4: Tabela com resultados da Figura 22a

Resultados Figura 22b					
Equipamento	Tamanho da amosta	Sensibilidade (S)	Especificidade (E)	Área (Σ)	Acurácia
A	1574	0,5089	0,5873	0,5436	0,5482
B	368	0,6957	0,5692	0,6643	0,6087
C	730	0,6508	0,5341	0,5812	0,5945
D	352	0,5030	0,5000	0,5116	0,5016

Tabela 5: Tabela com resultados da Figura 22b

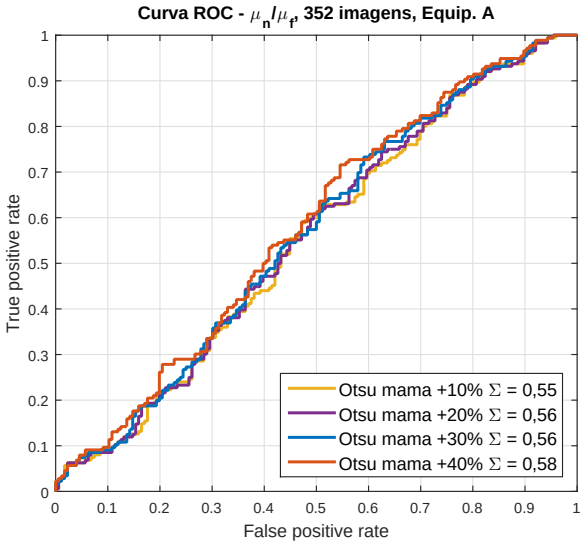
Nas Tabelas 4 e 5, o KNN possui performance similar à razão das médias.

Nos resultados apresentados até agora, o equipamento **B** possui a melhor performance e o sistema de classificação KNN ainda é superior a razão das médias, chegando a 64,27% de acurácia na Tabela 3. O equipamento **C** possui performance similar ao **B**, podendo ser até superior se levarmos em conta as Tabelas 4 e 5, onde o tamanho da amostra dele é bem maior que a do **B** mas com a acurácia se mantendo muito próxima.

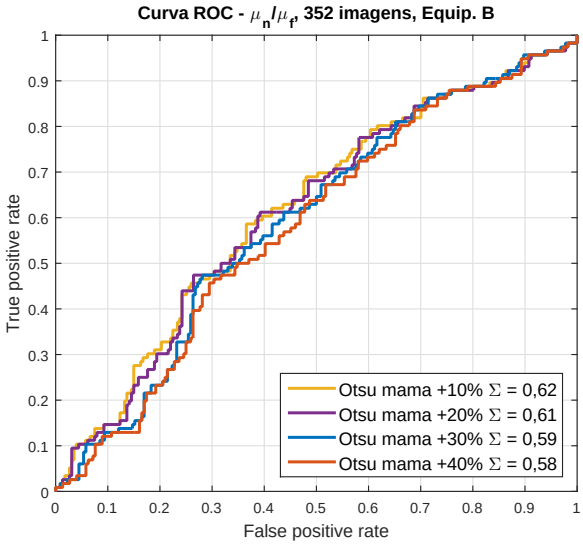
O equipamento **D** é o de pior performance, fato evidenciado tanto pelas curvas ROC quanto pela acurácia.

5.2 Efeito do ajuste do limiar Otsu para a mama

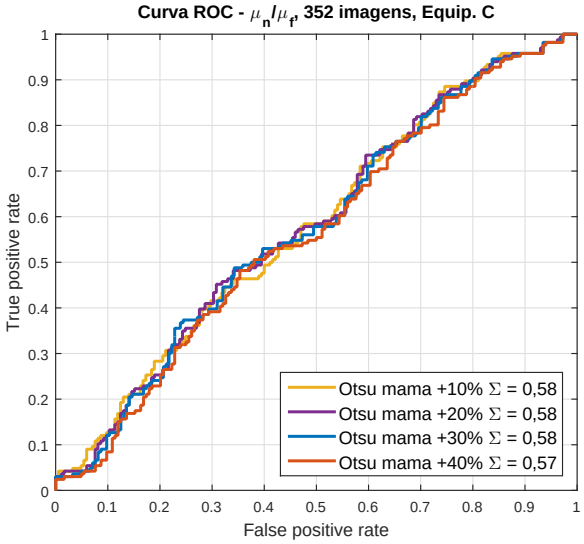
A Figura 23 compara resultados do sistema de classificação pela razão, para diferentes equipamentos, obtidos variando-se apenas o limiar Otsu da mama.



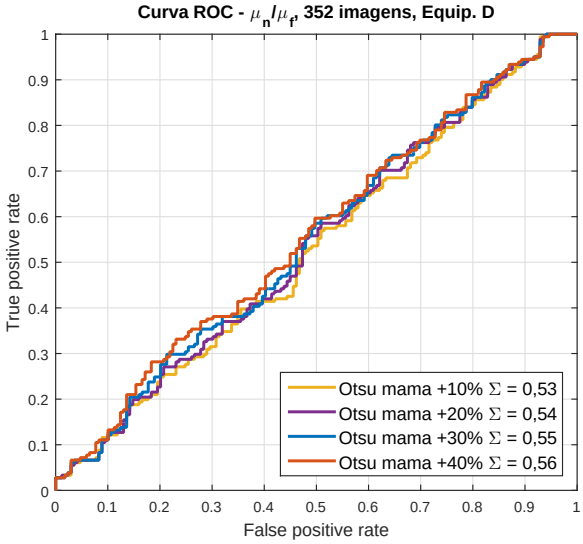
(a) Equipamento **A**



(b) Equipamento **B**



(c) Equipamento **C**



(d) Equipamento **D**

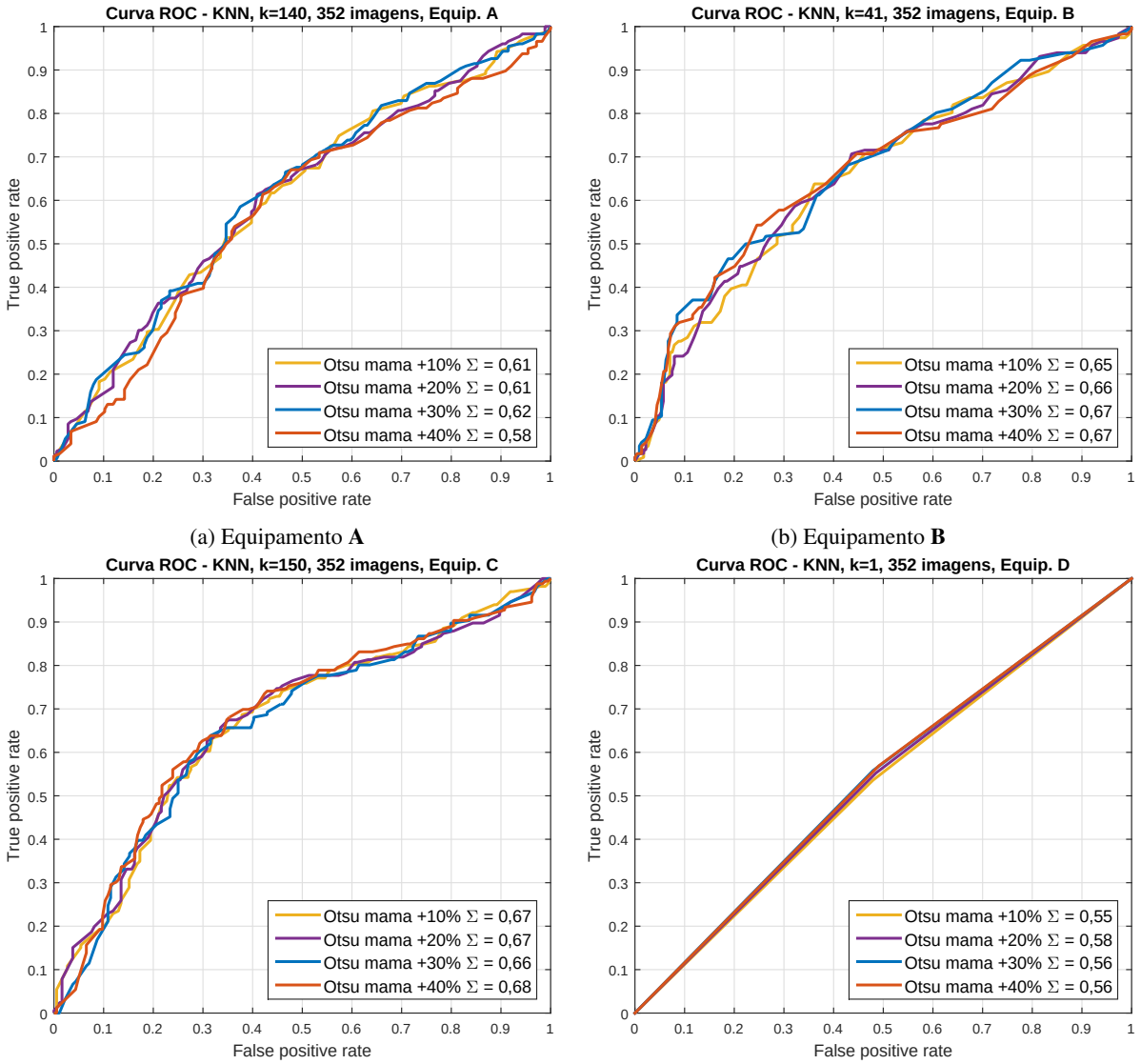
Figura 23: Classificador pela razão μ_n/μ_f: Variação do limiar Otsu da mama em todos os equipamentos

Resultados Figura 23

Equip.	Otsu da mama	Tam. da amosta	Sensibilidade (S)	Especificidade (E)	Área (Σ)	Acurácia
A	+10%	352	0,5543	0,5511	0,5506	0,5527
	+20%	352	0,5455	0,5511	0,5548	0,5483
	+30%	352	0,5341	0,5682	0,5626	0,5512
B	+10%	352	0,5862	0,6344	0,6177	0,6103
	+20%	352	0,6121	0,6079	0,6120	0,6100
	+30%	352	0,5862	0,5848	0,5937	0,5855
C	+10%	352	0,5783	0,5297	0,5798	0,5540
	+20%	352	0,5783	0,5203	0,5800	0,5493
	+30%	352	0,5301	0,6033	0,5752	0,5667

Tabela 6: Tabela com resultados do classificador pela razão μ_n/μ_f da Figura 23

A Figura 24 compara resultados do KNN, para diferentes equipamentos, obtidos variando-se o limiar Otsu da mama.



(a) Equipamento **A**

(b) Equipamento **B**

(c) Equipamento **C**

(d) Equipamento **D**

Figura 24: classificador KNN: Variação do limiar Otsu da mama nos equipamentos **A**, **B** **C** e **D**

Resultados Figura 24

Equip.	Otsu da mama	Tam. da amosta	Sensibilidade (S)	Especificidade (E)	Área (Σ)	Acurácia
A	+10%	352	0,6171	0,5625	0,6060	0,5740
	+20%	352	0,6136	0,5909	0,6083	0,6023
	+30%	352	0,5852	0,6260	0,6150	0,6056
B	+10%	352	0,7069	0,5674	0,6527	0,6222
	+20%	352	0,7069	0,5639	0,6575	0,6354
	+30%	352	0,6810	0,5714	0,6709	0,6262
C	+10%	352	0,6988	0,6000	0,6691	0,6494
	+20%	352	0,6747	0,6486	0,6703	0,6617
	+30%	352	0,6566	0,6522	0,6626	0,6544

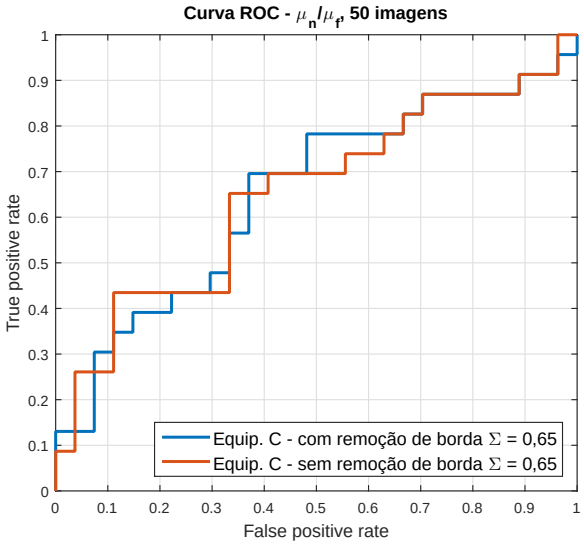
Tabela 7: Tabela com resultados do classificador KNN da Figura 24

Nas Tabelas 6 e 7 os resultados para o ajuste de Otsu da mama com +40% foram omitidos pois, com esse valor de ajuste, a mama não é segmentada corretamente. Isso ocorre porque, com um limiar muito alto, parte do fundo é segmentado junto com a mama.

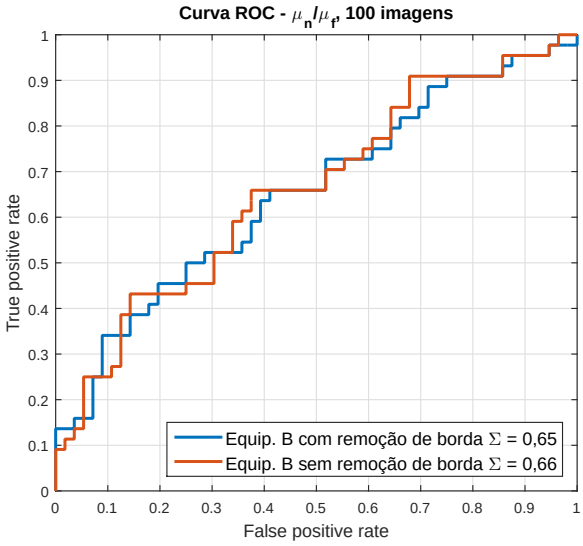
Comparando a acurácia de ambos os sistemas de classificação, o KNN obteve resultados melhores, chegando a 66,17% de acurácia, contra 61,03% com a razão das médias. O equipamento **C** tem a melhor performance nesse teste, isso fica claro com a Figura 24c. Com base nesses resultados, escolhemos o valor de ajuste para o limiar Otsu da mama de +20% como o padrão para o programa principal.

5.3 Remoção de borda

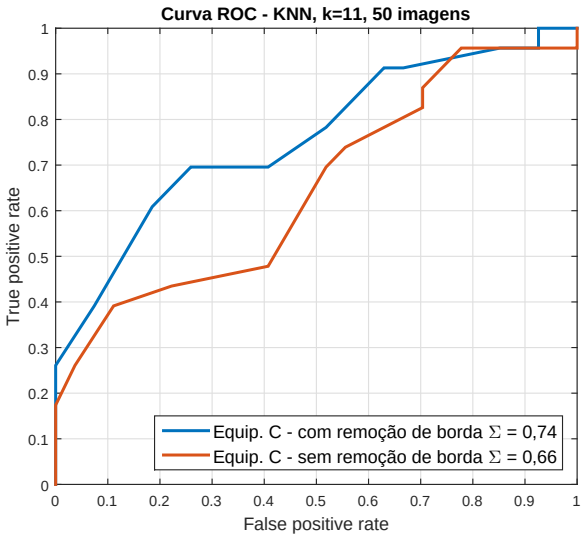
A Figura 25 expões os resultados obtidos para ambos os sistemas de classificação com e sem a opção de remoção de borda



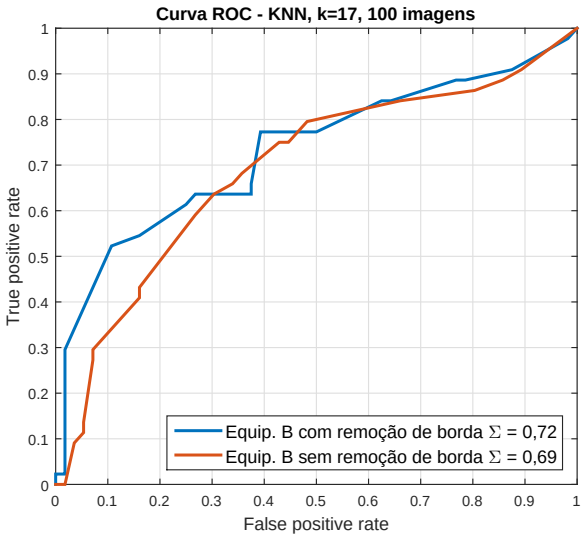
(a) Classificação pela razão μ_n/μ_f, 50 imagens



(b) Classificação pela razão μ_n/μ_f, 100 imagens



(c) Classificação por KNN, 50 imagens



(d) Classificação por KNN, 100 imagens

Figura 25: Efeito da remoção de bordas

Resultados das Figuras 25a e 25b					
Remoção de Borda	Tamanho da amosta	Sensibilidade (S)	Especificidade (E)	Área (Σ)	Acurácia
Sim	50	0,6957	0,6296	0,6506	0,6627
Sim	100	0,6591	0,6500	0,6526	0,6546
Não	50	0,6957	0,6296	0,6506	0,6627
Não	100	0,6591	0,5893	0,6591	0,6242

Tabela 8: Tabela com resultados do classificador pela razão μ_n/μ_f das Figuras 25a e 25b

Resultados das Figuras 25c e 25d					
Remoção de Borda	Tamanho da amosta	Sensibilidade (S)	Especificidade (E)	Área (Σ)	Acurácia
Sim	50	0,7391	0,6296	0,7367	0,6844
Sim	100	0,7500	0,6429	0,7200	0,6965
Não	50	0,5652	0,7407	0,6610	0,6530
Não	100	0,7273	0,6607	0,6922	0,6940

Tabela 9: Tabela com resultados do classificador KNN das Figuras 25c e 25d

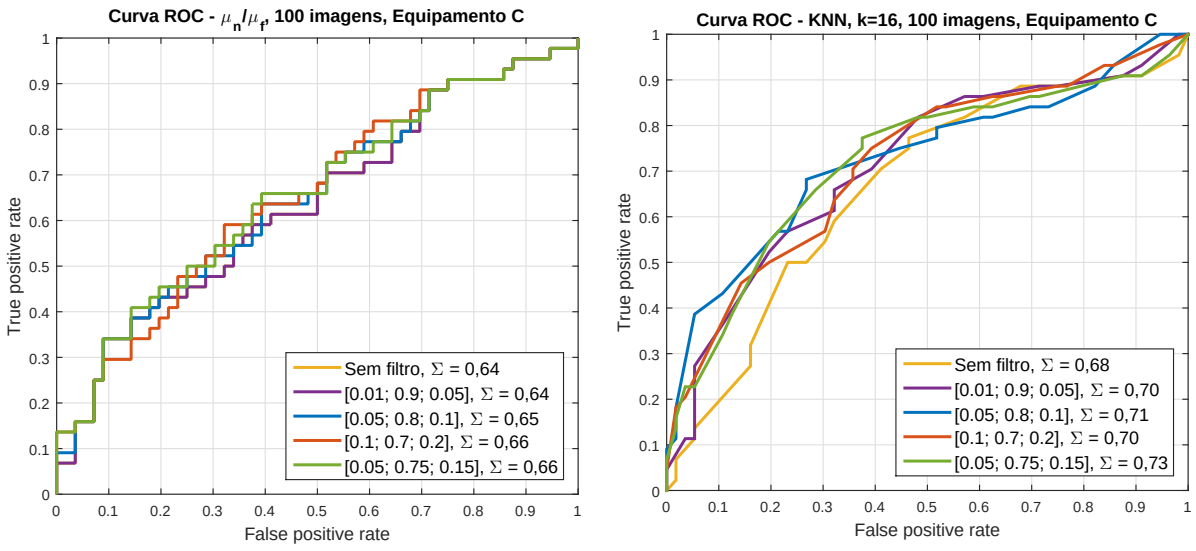
Resultados com a opção de remoção de borda ativada apresentam, embora não muito, performance superior para ambos os tamanhos de amostra e o KNN mais uma vez obteve a maior acurácia.

5.4 Efeito da variação dos parâmetros do filtro Otsu do nódulo

No caso do nódulo, como explicado no Capítulo 4, item 4.6.2, o programa permite variar parâmetros de um filtro que mantém o nódulo dentro de um limite de tamanho. Esses parâmetros são:

- *Threshold offset*: Ajuste do *threshold* de Otsu
- *Max area Coefficient*: Área máxima que o nódulo pode assumir
- *Min area Coefficient*: Área mínima que o nódulo pode assumir

A Figura 26 mostra resultados para uma amostra de 100 imagens do equipamento **C** com Otsu da mama ajustado para $+20\%$ onde varia-se os parâmetros do filtro Otsu para o nódulo. Esses parâmetros estão no formato $[offset, A_{max}, A_{min}]$ na legenda do gráfico e nas Tabelas 10 e 11



(a) Limiar μ_n/μ_f (b) KNN

Figura 26: Performance de ambos os classificadores variando-se o filtro do nódulo

Resultados Figura 26a					
Filtro	Tamanho da amosta	Sensibilidade (S)	Especificidade (E)	Área (Σ)	Acurácia
–	100	0,6136	0,5893	0,6380	0,6015
$[0,01;\ 0,9;\ 0,05]$	100	0,6136	0,5893	0,6380	0,6015
$[0,05;\ 0.8;\ 0,1]$	100	0,6364	0,6071	0,6489	0,6218
$[0,1;\ 0,7;\ 0,2]$	100	0,6364	0,6071	0,6554	0,6218
$[0,05;\ 0,75;\ 0,15]$	100	0,6591	0,6071	0,6595	0,6331

Tabela 10: Tabela com resultados do classificador pela razão μ_n/μ_f da Figura 26a

Resultados Figura 26b					
Filtro	Tamanho da amosta	Sensibilidade (S)	Especificidade (E)	Área (Σ)	Acurácia
–	100	0,6364	0,7143	0,6841	0,6754
$[0,01;\ 0,9;\ 0,05]$	100	0,6364	0,7500	0,6958	0,6932
$[0,05;\ 0,8;\ 0,1]$	100	0,7500	0,6429	0,7123	0,6965
$[0,1;\ 0,7;\ 0,2]$	100	0,7045	0,6786	0,6983	0,6916
$[0,05;\ 0,75;\ 0,15]$	100	0,6818	0,7143	0,7271	0,6981

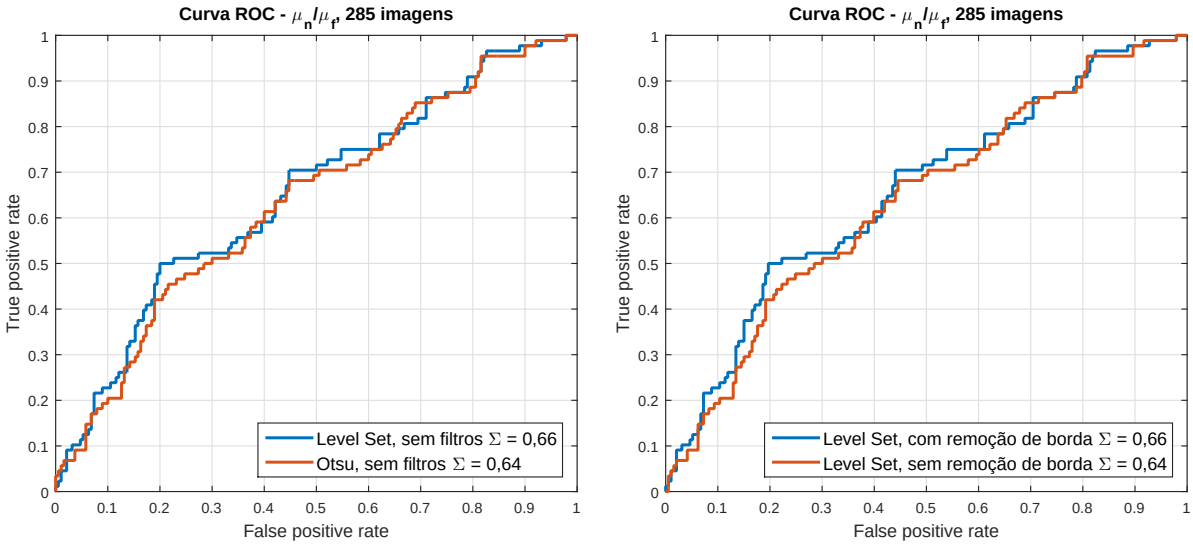
Tabela 11: Tabela com resultados do classificador KNN da Figura 26b

Por se tratar de uma amostra relativamente pequena, não são muitos os nódulos que caem fora do limite estabelecido pelos parâmetros do filtro, e consequentemente o número de imagens afetada por esse filtro é pequeno. Portanto, já era de se esperar que o resultado, embora positivo, não seja muito expressivo. Os melhores parâmetros ajustados foram [0,05; 0,75; 0,15] e por isso esses valores foram escolhidos como os padrões do programa.

O KNN obteve, de novo, melhores resultados que a razão das médias.

5.4.1 Resultados *Level Set*

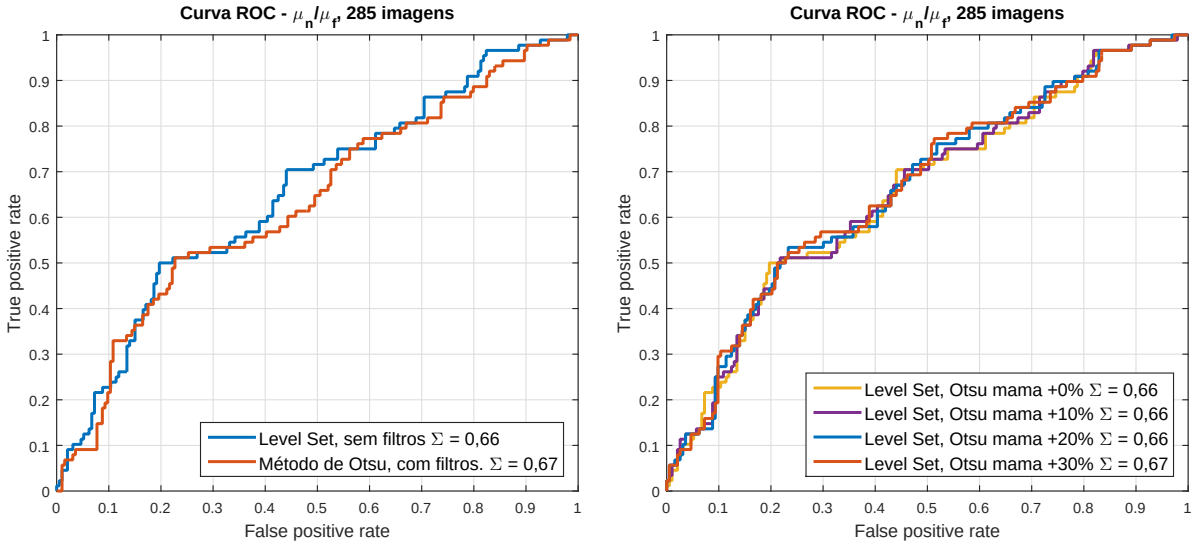
As imagens usadas para comparar os métodos de Otsu e *Level Set* nas Figuras 27 à 30 são as mesmas para ambos os métodos e correspondem a mamografias provenientes do equipamento **C**. As Figuras 27 e 28 refere-se ao classificador pela razão μ_n/μ_f e essas figuras mostram resultados obtidos variando-se diversos parâmetros do programa.



(a) *Level Set* × Otsu padrão

(b) Efeito da remoção de borda

Figura 27: Método de segmentação *Level Set*, classificador pela razão μ_n/μ_f



(a) *Level Set* × Otsu otimizado (b) Efeito do ajuste do limiar Otsu da mama

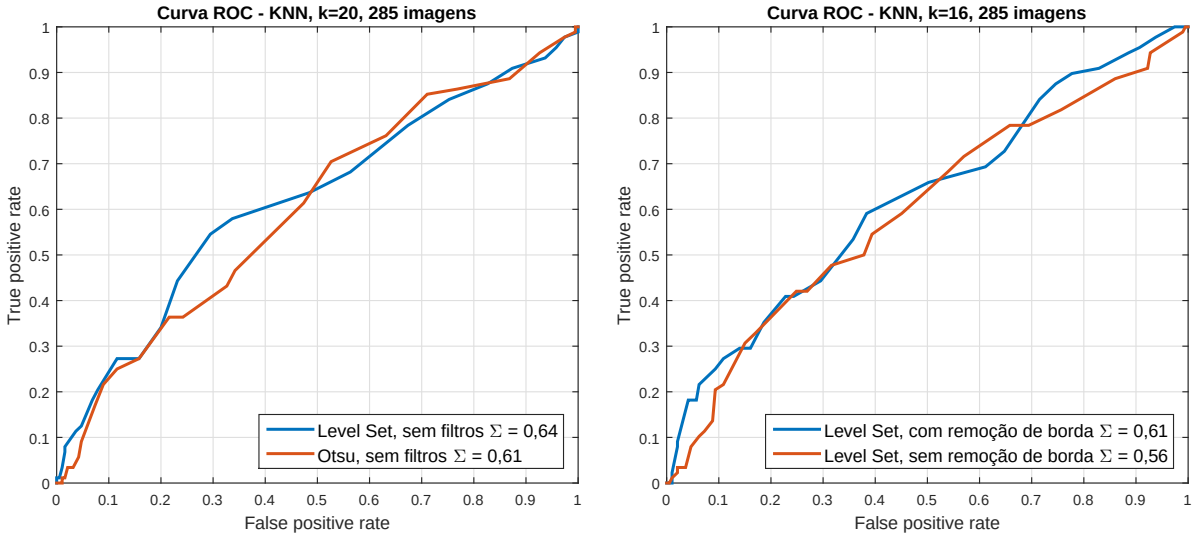
Figura 28: Método de segmentação *Level Set*, classificador pela razão μ_n/μ_f

Resultados Figuras 27 e 28								
Método	Amostra	Filt. mama	Filt. Nod.	Rem. borda	(S)	(E)	(Σ)	Acurácia
Level Set	285	Não	-	Sim	0,7045	0,5526	0,6525	0,6002
Otsu	285	Não	Não	Sim	0,6818	0,5526	0,6379	0,5931
Otsu	285	+20%	[0,05; 0,75; 0,15]	Sim	0,7019	0,5445	0,6655	0,5938

Tabela 12: Tabela com resultados do classificador pela razão μ_n/μ_f das Figuras 27 e 28

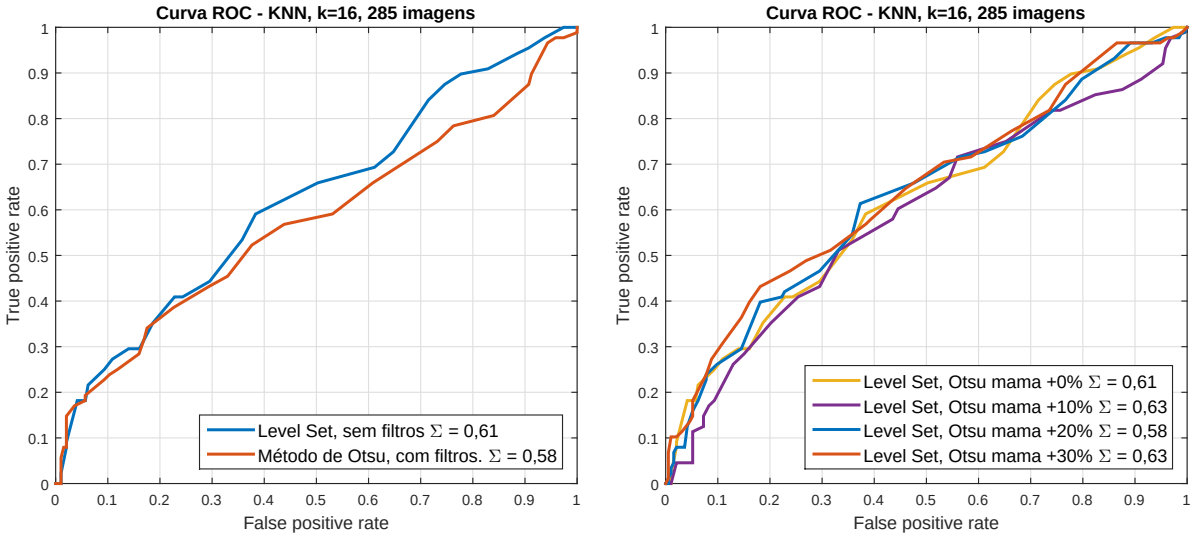
Os resultados com o *Level Set* apresentam acurácia superior mas a diferença não é expressiva.

As Figuras 29 e 30 mostram a performance do classificador KNN.



(a) *Level Set* × Otsu padrão (b) Efeito da remoção de borda

Figura 29: Método de segmentação *Level Set*, classificador KNN



(a) *Level Set* × Otsu otimizado (b) Efeito do ajuste do limiar Otsu da mama

Figura 30: Método de segmentação *Level Set*, classificador KNN

Resultados Figuras 29 e 30

Método	Amosta	Filt. mama	Filt. Nod.	Rem. borda	(S)	(E)	(Σ)	Acurácia
Level Set	285	Não	-	Sim	0,6250	0,6263	0,6444	0,6259
Otsu	285	Não	Não	Sim	0,5682	0,6477	0,6123	0,6228
Otsu	285	+20%	[0,05; 0,75; 0,15]	Sim	0,6023	0,5361	0,5818	0,5368

Tabela 13: Tabela com resultados do classificador KNN das Figuras 29 e 30

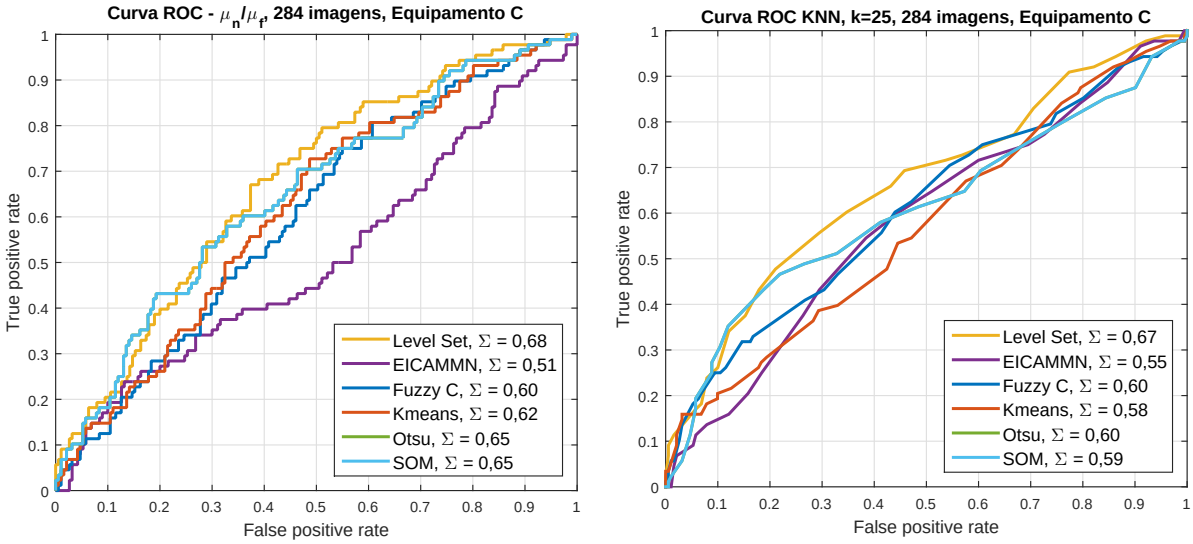
Embora o método de *Level Set* tenha resultados melhores em todos os casos (Tabelas 12 e 13), essa melhora não é expressiva e não justificaria, pelo menos para a análise da densidade dos níveis de cinza, a implementação dessa técnica.

5.5 Outros métodos de segmentação

Com o recurso de usar nódulos externos já segmentados, exploram-se alguns métodos desenvolvidos previamente por outros pesquisadores e os quais têm nódulos já segmentados no LAPIMO. Os métodos comparados aqui são:

- *Level Set* (OSHER e SETHIAN, 1988)
- EICAMM (RIBEIRO, 2013)
- Fuzzy C (KELLER, 2012)
- Kmeans (DALMIYA, DASGUPTA e DATTA, 2012)
- Otsu (OTSU, 1979)
- SOM (AHIRWAR e JADON, 2011)

A Figura 31 compara esses diversos métodos, para classificação baseada na densidade de níveis de cinza, em um só gráfico.



(a) Limiar μ_n/μ_f (b) KNN

Figura 31: Performance de ambos os classificadores para outros métodos de segmentação

Com esses resultados fica visível que, para a classificação pela densidade de níveis de cinza, os métodos de Otsu e Level Set apresentam os melhores resultados.

6 Conclusão

O objetivo inicial do trabalho era implementar um programa capaz de retornar as densidades dos níveis de cinza tanto do nódulo quanto da mama, de uma grande amostra de imagens, para que, com um sistema de classificação adequado, seja possível estabelecer qual a relação entre essas densidades e a classificação do nódulo. Se a relação existisse, seria viável usar a densidade de níveis de cinza da mamografia como um dos atributos analisados dentro de um sistema CADx mais completo.

No desenvolvimento do programa foi utilizada o método de Otsu (OTSU, 1979) para a segmentação das imagens, já implementado anteriormente no LAPIMO, e sua aplicação exigiu o cuidado para que ele fosse aplicado apenas na região de interesse (mama ou nódulo), dependendo da etapa de processamento da mamografia. Durante a programação de um meio para conseguir isso, foram desenvolvidos dois filtros, uma para a mama e outro para o nódulo, que melhoram suas respectivas segmentações ajustando o *threshold* calculado pelo método de Otsu. Esses filtros são descritos na seção 4.6, Capítulo 4. Além disso, foi desenvolvido um algoritmo para remoção de bordas brancas que é descrito na seção 4.6.3 do mesmo capítulo.

Como o LAPIMO trabalho com outros métodos de segmentação de imagem além de Otsu, implementou-se um meio de se utilizar nódulos já segmentados em imagens separadas. Essa implementação permitiu comparar a performance de diferentes métodos (seção 5.5, Capítulo 5).

O programa foi implementado de forma a fazer o uso de *multithreading*, uma vez que o processamento de cada imagem é independente e uma versão *singlethread* do programa deixava o restante dos processadores lógicos da máquina desocupados. A versão final do programa possui interface gráfica intuitiva, onde é possível ajustar diversos parâmetros referentes aos filtros desenvolvidos neste trabalho e também diversas facilidades como limitar o número de *threads* e gravar a última pasta de arquivos usada para carregar as imagens do disco.

Uma vez que o programa foi implementado e o objetivo foi alcançado, as densidades do nódulo e da mama foram carregadas no MATLAB a fim de testar a eficácia de dois sistemas de classificação, a simples razão entre as densidades e o método KNN, comparando seus resultados.

Os resultados mostram que a performance de ambos os classificadores são bastante sensíveis ao tamanho da amostra de imagens, ao tipo de escâner usado na digitalização das mamografias e ao método de segmentação implementado. No estudo preliminar (Capítulo 3) havia uma amostra pequena de casos benignos comparada a de casos malignos e portanto a performance de ambos os classificadores podem gerar resultados distorcidos. No programa principal, com amostras de tamanhos similares para ambos os casos, o classificador KNN obteve acurácia de 69,81% para uma amostra de 100 imagens e 66,17% com uma amostra de 352 imagens (Tabela 7 e 11), já a classificação pela razão das médias mostrou performance inferior ao KNN em todos os casos.

Comparando os resultados para diferentes equipamentos, os escâneres **B** e **C** obtiveram resultados bem melhores que os **A** e **D**. Esse fato mostra que, para usar um sistema CADx no auxílio do diagnóstico por imagem, é necessário que o equipamento utilizado na digitação das mamografias seja bem calibrado e obtenha resultados satisfatórios no treinamento heurístico.

Variando o método de segmentação das mamografias, observou-se que o *Level Set* obteve a maior acurácia porém o programa não tem suporte nativo a este método e deve receber nódulos já segmentados para utilizá-lo, o que deixa espaço para desenvolvimentos futuros do programa. Mesmo otimizando os parâmetros para o método de Otsu, o *Level Set* ainda obteve acurácia superior para ambos os sistemas de classificação.

Com isso, mostrou-se que, não só existe uma relação entre a densidade relativa de níveis de cinza de um nódulo com sua classificação, como essa relação depende muito do equipamento, tamanho da amostra e do sistema de classificação adotados.

Por último, a segmentação de um grande número de imagens mostrou-se um processo demorado que pode levar horas dependendo do tamanho da amostra. O programa processa em média 25 imagens por minuto utilizando 8 *threads* mas cai para 5 imagens por minuto na versão *singlethread*. Essa diferença de performance justificou a implementação de *multithreading* e máquinas com um número maior de processadores lógicos teriam performance ainda maior no treinamento heurístico.

6.1 Desenvolvimentos futuros

O programa desenvolvido neste trabalho possibilita muitas melhorias, por exemplo:

- Otimização do código, resultando em menor tempo de execução e melhor gerenciamento de memória.
- Introdução de novas técnicas de segmentação (*Level Set*), que apresentem melhores resultados.
- Introdução de análise de outros aspectos do nódulo ou integração com um sistema CADx mais completo, procurando aumentar a sensibilidade e acurácia dos resultados.
- Ajuste mais fino dos parâmetros dos filtros presentes no programa.

Além disso, como os resultados são sensíveis ao tipo de equipamento de aquisição da imagem de raio-X, pode-se implementar um algoritmo para suavizar a não-homogeneidade da luminosidades dos *pixels* das mamografias. Esse tipo de abordagem é usada em técnicas avançadas de *Level Set* e mostram melhorias nos resultados (CHUNMING, HUANG E DING, 2011).

7 Referências Bibliográficas

AHIRWAR, A.; JADON, R. S. **Characterization of tumor region using SOM and Neuro Fuzzy techniques in Digital Mammography.** International Journal of Computer Science and Information Technology, v. 3, n. 1, p. 199-211, 2011.

ARORA, S.; BARAK, B. **Computational complexity:** a modern approach. Cambridge University Press, 2009.

BICK, U.; DIEKMANN, F. **Digital Mammography**. Berlin: Springer Science & Business Media, 2010. 219p.

BOYLE, P. **Current situation of screening for cancer**. Annals of Oncology-English Edition, v. 13, n. 4, p. 189-198, 2002.

BURGER, W., BURGE, M. J. **Principles of Digital Image Processing:** Core Algorithms Springer, 2009. 329p.

BUZARD, A. C.; MALUF, F. C.; LIMA, C. M. R. **MOC 2015 - Manual de Oncologia Clínica do Brasil - Tumores Sólidos.** 13th ed. São Paulo: Dendrix, 2015. 764p.

CASSIDY, J.; BISSET, D.; SPENCER, R.; PAYNE, M. **Oxford Handbook of Oncology** 3th ed. Oxford: OUP Oxford, 2010. 896p.

CHAN, H. et al. **Improvement of Radiologists' Characterization of Mammographic Masses by Using Computer-aided Diagnosis: An ROC Study** 1. Radiology, v. 212, n. 3, p. 817-827, 1999.

CHEN, T. F. **Medical image segmentation using level sets**. Technical Report. Canada, University of Waterloo, 2008.

CHUNMING, L.; HUANG, R.; DING, Z.; GATENBY, C.; METAXAS D. N.; GORE, J. C. **A Level Set Method for Image Segmentation in the Presence of Intensity Inhomogeneities With Application to MRI**, Transactions on Image Processing, v. 20, n. 7, jul. 2011.

DALMIYA, S.; DASGUPTA, A.; DATTA, S. **Application of Wavelet based K-means Algorithm in Mammogram Segmentation.** International Journal of Computer Applications, v. 52, n. 15, p. 15-19, 2012.

EISBERG, R.; RESNICK, R. **Física Quântica –átomos, moléculas. Sólidos, núcleos e partículas**. Rio de Janeiro: Campus, 1994.

GONZALES, R. C.; WOODS, R. E. **Digital image processing.** 3th ed. New Jersey: Prentice Hall, 2007. 976p.

HANLEY, J A.; MCNEIL, B. J. **The meaning and use of the area under a receiver operating characteristic (ROC) curve**. Radiology, v. 143, n. 1, p. 29-36, 1982.

INSTITUTO NACIONAL DO CÂNCER. INCA, 2015. Disponível em: <http://www.inca.gov.br/conteudo_view.asp?id=1932>. Acesso em: 2015.

KELLER, B. M. et al. **Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation.** Medical physics, v. 39, n. 8, p. 4903-4917, 2012.

KOHAVI, R. et al. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. In: Ijcai. 1995. p. 1137-1145.

LIU, X.; TANG, J. **Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method.** Systems Journal, IEEE, v. 8, n. 3, p. 910-920, 2014.

MARKEY, M. K. **Physics of Mammographic Imaging.**, 1st ed. CRC Press, 2012. 317p.

MATHEUS, B. R. N. **Sistema JAVA para gerenciamento de esquema CADx em mamografia.** 2015. Tese (Doutorado em Processamento de Sinais de Instrumentação) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2015. Disponível em: <http://www.teses.usp.br/teses/disponiveis/18/18152/tde-22102015-093201/>. Acesso em: 2015-11-13.

MORÉ, J. J. **The Levenberg-Marquardt algorithm: implementation and theory.** In: Numerical analysis. Springer Berlin Heidelberg, 1978. p. 105-116.

MURTY, M. N.; DEVI, V. S. **Pattern Recognition**. London: Springer London, 2011. 263p.

NASCIMENTO, M. Z.; RAMOS, R. P. **Combinando duas visões mamográficas em extração de características com Ridgelet.** In: XI Congresso Brasileiro de Informática em Saúde, Campos do Jordão. XI Congresso Brasileiro de Informática em Saúde. 2008.

OSHER, S.; SETHIAN, A. **Fronts propagating with curvature dependent speed**: Algorithms based on Hamilton-Jacobi formulations. Journal of Computational Physics, v. 79, p. 12-49, 1988.

OTSU, N. **A threshold selection method from gray-level histograms**. Automatica, v. 11, n. 285-296, p. 23-27, 1979.

PEIXOTO, J. E.; CANELLA, E.; AZEVEDO, A. C. **Mamografia: da prática ao controle.** Rio de Janeiro: Gráfica Esdeva, 2007.

PRESSMAN, R. **Engenharia de Software** McGraw-Hill 7th ed, 2011.

RIBEIRO, P. B. **Esquema CADx para classificação de nódulos em imagens mamográficas digitais baseado na segmentação pelo modelo EICAMM.** 2013. Tese (Doutorado em Processamento de Sinais de Instrumentação) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2013. Disponível em:
<http://www.teses.usp.br/teses/disponiveis/18/18152/tde-01072013-101058/>.
Acesso em: 2015-12-03.

SCHILDT, H. **Java**: A Beginner's Guide 6th ed. McGraw-Hill Education, 2013 728p.

SLABY, A. **ROC analysis with Matlab**. In: Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on. IEEE, 2007. p. 191-196.

ZHU, W. et al. **Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations.** NESUG proceedings: health care and life sciences, Baltimore, Maryland, p. 1-9, 2010.

Anexo A - Implementação do Método de Otsu em Java

```
ImagePlus img =new ImagePlus(listOfImages
[nlin].getPath());
ImageProcessor ip =  img.getProcessor();
int bit=ip.getBitDepth();
int niveisCinza=(int) Math.pow(2, bit)-1;
int limiar;
int threshold=0;
double w0=0 , uT=0 , ut1=0, SigB2=0,
SigT2=0, nf=0, nmin=0;
double n=0, x=0, y=0;
double [] ni = null;
double [] pI = null;
ni = new double[niveisCinza+1];
pI = new double[niveisCinza+1];
int linha = ip.getHeight();
int coluna = ip.getWidth();
n = linha*coluna;
nmin = -1.0;
for (int i = 0; i <= niveisCinza; i++)
{
    ni[i]=0;
}
for(int i = 0; i < coluna; i++)
{
    for(int j = 0; j < linha; j++)
    {
        int temp = ip.getPixel(i, j);
        if ((temp>=0)&&(temp<=niveisCinza))
            ni[temp] = ni[temp]+1;
    }
}
for (int i = 0; i <= niveisCinza; i++)
{
    pI[i] = ni[i]/n;
}
uT = 0.0;
for (int i = 0; i <= niveisCinza; i++)
{
    uT = uT + i * pI[i];
}
SigT2 = 0.0;
for (int i = 0; i <= niveisCinza; i++)
{
    SigT2=SigT2+(i-uT)*(i-uT)*pI[i];
}
int j = -1;
int k = -1;
for (int i = 0; i <= niveisCinza; i++)
{
    if ((j<0) && (pI[i] > 0.0))
        j=i;
    if (pI[i] > 0.0)
        k=i;
}
for (int t=j; t<=k; t++)
{
    ut1 = 0.0;
    w0 = 0.0;
for (int i = 0; i <= t; i++)
{
    ut1 = (ut1 + (i * pI[i]));
    w0 = w0 + pI[i];
}
x = uT*w0-ut1;
x = x*x;
y = w0*(1.0 - w0);
if (y>0.0)
    x = x/y;
else
    x = 0.0;
SigB2 = x;
nf = SigB2 / SigT2;
if (nf >= nmin)
{
    nmin = nf;
    threshold = t - 1;
}
}
limiar = threshold;
return limiar;
```

Anexo B - Pseudocódigo e código em Java do algoritmo *Flood Fill*

```
Crie uma fila vazia Q
Insira o ponto inicial (u,v) na fila:
ENQUEUE(Q,(u,v))
Enquanto Q não estiver vazio faça
Pegue a próxima coordenada da frente da fila:
(x,y) ← DEQUEUE(Q)
Se (x,y) estiver dentro da imagem && I(x,y) = 1 Então
I(x,y) ← label
ENQUEUE(Q,(x+1,y))
ENQUEUE(Q,(x,y+1))
ENQUEUE(Q,(x,y−1))
ENQUEUE(Q,(x−1,y))
Fim_Se
Fim_Enquanto
Retorna
```

Pseudocódigo (BURGER e BURGE, 2009)

```
int ymassi=(int) (long) ymass;
int xmassi=(int) (long) xmass;
LinkedList<Point> q = new LinkedList<Point>();
q.addFirst(new Point(xmassi, ymassi));
while (!q.isEmpty())
{Point n = q.removeLast();
int u = n.x; int v = n.y;
if((u>=xini)&&(u<xmax)&&(v>=yini)&&(v<ymax)&&ipb.getPixel(u,v)==255)
{ipb.putPixel(u, v, 50);
q.addFirst(new Point(u+1, v));
q.addFirst(new Point(u, v+1));
q.addFirst(new Point(u, v-1));
q.addFirst(new Point(u-1, v));}}
```

Implementação em Java