DETECÇÃO DE LESÕES EM MAMOGRAFIAS ATRAVÉS DA ASSIMETRIA DAS MAMAS E EXTRAÇÃO DE CARACTERÍSTICAS COM ÍNDICE DE GETIS-ORD

LESION DETECTION IN MAMMOGRAMS THROUGH THE ASYMMETRY OF THE BREASTS AND FEATURE EXTRACTION WITH INDEX GETIS-ORD

DETECCIÓN DE LESIONES EN LAS MAMOGRAFÍAS A TRAVÉS DE LA ASIMETRÍA DE LAS MAMAS Y EXTRACCIÓN DE CARACTERÍSTICAS CON EL ÍNDICE GETIS-ORD

Antônio Marcos Vieira Sales
Aristófanes Corrêa Silva
Anselmo Cardoso de Paiva

Resumo: O câncer de mama é aquele que tem início nas células das mamas. A principal forma de prevenção e diagnóstico precoce é através de exames de mamografia. Este trabalho tem como objetivo principal apresentar uma metodologia de auxílio à detecção de lesões em mamografias a partir da determinação de regiões suspeitas por nível de simetria. Técnicas de Processamento de Imagem foram usadas para preparar as mamografias e, em seguida, o nível de simetria entre a mama esquerda e a direita foi medido com coeficiente de correlação cruzada e distância euclidiana. O índice de Getis-Ord na sua forma geral foi usado para extrair características das imagens para treinar uma Máquina de Vetores de Suporte que classificou regiões das mamografias em lesão e não lesão. A metodologia, de modo geral, apresentou 80,11% de sensibilidade, 84,41% de especificidade e 84,38% de acurácia.
Palavras-chave: Câncer de mama. Mamografia. Coeficiente de correlação cruzada. Distância euclidiana. Índice de Getis-Ord. Máquina de vetores de suporte.

Abstract: Breast cancer is one that starts in the cells of the breast. The main form of prevention and early diagnosis is through mammograms. This work has as main goal to present a methodology to aid in the detection of lesions on mammograms from the determination of suspicious regions by level of symmetry. Image processing techniques were used to prepare the mammograms and then the degree of symmetry between left and right breasts was measured using cross-correlation coefficient and Euclidean distance. The index Getis-Ord was used to extract features from images to train a Support Vector Machine which classified regions of mammograms in lesion and non-lesion. The methodology, in general, showed 80.11% sensitivity, 84.41% specificity and 84.38% accuracy.
Keywords: Breast cancer. Mammography. Cross-correlation coefficient. Euclidean distance. Index Getis-Ord. Support vector machine.

Resumen: El cáncer de mama comienza en las células de los senos. La principal forma de prevención y diagnóstico precoz es a través de mamografías. Este trabajo tiene como objetivo principal presentar una metodología para ayudar en la detección de lesiones en las mamografías a partir de la determinación de las regiones sospechosas por nivel de simetría. Técnicas de procesamiento de imágenes se utilizaron para preparar las mamografías y luego el nivel de simetría entre el pecho izquierdo y derecho se midió utilizando el coeficiente de correlación cruzada y la distancia euclidiana. El índice Getis-Ord se utilizó para extraer características de las imágenes para formar una máquina de vectores de soporte que las regiones clasificadas de mamografías en lesión y no la lesión. La metodología, en general, mostró 80,11% de sensibilidad, especificidad 84,41% y 84,38% de precisión.
Palabras clave: Cáncer de mama. Mamografía. Coeficiente de correlación cruzada. Distancia euclídea. Índice Getis-Ord. Máquina de vectores soporte.

1 INTRODUÇÃO

O câncer de mama, de acordo com a *American Cancer Society* (ACS), é o segundo tipo de câncer a causar mais mortes em mulheres (ACS, 2012). Segundo o Instituto Nacional do Câncer (INCA), também é o tipo de câncer que mais acomete mulheres em todo o mundo, respondendo por, aproximadamente, 22% dos casos novos a cada ano. A estimativa para o ano de 2012 (também válida para 2013) foi de 52.680 novos casos no Brasil, onde as taxas de mortalidade por essa doença continuam elevadas, muito provavelmente porque a doença ainda é diagnosticada em estágios avançados (INSTITUTO NACIONAL DO CÂNCER, 2011).

O fato de a chance de cura ser bem maior quando o câncer é detectado precocemente e as estimativas de que grande parte das lesões deixa de ser detectada no exame feito pelo

*Artigo recebido em julho de 2013
Aprovado em outubro de 2013

especialista, devido à dificuldade de identificar padrões na imagem, são um grande incentivo à pesquisa e ao desenvolvimento de métodos e sistemas que visam a auxiliar o especialista na detecção (sistemas CAD - *Computer-Aided Detection*) e diagnóstico (sistemas CADx - *Computer-Aided Diagnosis*) da doença, indicando áreas suspeitas, assim como anormalidades mascaradas (ROCHA et al., 2011).

O exame mais usado na prevenção e diagnóstico do câncer de mama é a mamografia ou radiografia da mama, que permite a descoberta de lesões em fase inicial, imperceptíveis ao exame do toque. É difícil determinar a efetividade da mamografia, que depende de fatores relacionados às características da própria mama, da lesão, dos recursos disponíveis e da interpretação do especialista (INSTITUTO NACIONAL DO CÂNCER, 2012). A análise feita em imagens mamográficas pode ser melhorada com o auxílio de processamento digital de imagens e visão computacional. Assim, a principal contribuição das pesquisas na área de detecção e diagnósticos baseados em processamento de imagens vêm da dificuldade humana em localizar e classificar áreas lesionadas.

Este artigo está organizado em 6 seções. Na seção 2 serão listados alguns trabalhos relacionados. A seção 3 contém os fundamentos teóricos para a compreensão da metodologia proposta. Esta é descrita na seção 4, seguida dos resultados apresentados na seção 5. As considerações finais são feitas na seção 6.

2 TRABALHOS RELACIONADOS

Estudos anteriores vêm analisando o nível de assimetria entre mamografias esquerda e direita da mesma paciente e o risco de desenvolver câncer. Um trabalho realizado em Scutt; Lancaster; Manning (2006) detectou uma assimetria de volume maior em pacientes que chegaram a desenvolver câncer até o fim da pesquisa do que em pacientes que permaneceram normais.

Em Lee (1997), o autor propôs uma metodologia de registro de mamografias com intuito de fazer um mapeamento entre duas imagens, por meio do cálculo do fluxo óptico, entre a imagem fonte e a imagem destino. Já em Rodrigues (2010), o autor realizou uma comparação entre diversas métricas de similaridade no contexto de corregistro não rígido de imagens médicas, sendo o coeficiente de correlação cruzada uma das métricas consideradas mais robustas.

A utilização de medidas chamadas de descritores espaciais em conjunto com técnicas de aprendizado de máquina apresenta constante crescimento no processamento de imagens médicas. Um estudo feito por Braz Júnior (2008) utilizou várias dessas medidas (Índice de Moran, Coeficiente de Geary, Índice de Getis-Ord e função K de Ripley), sendo as características extraídas submetidas para treinamento e classificação com Máquina de Vetores de Suporte.

Uma metodologia de detecção de massas baseada na análise de simetria foi proposta em Ericeira (2011). Na dissertação, o autor utilizou registro bilateral das mamografias da mesma paciente e a função conhecida como variograma cruzado para extração de características de pares de regiões correspondentes das imagens.

Este trabalho tem como objetivo geral apresentar uma metodologia para detecção de regiões lesionadas em mamografias, partindo da determinação de pares de regiões suspeitos correspondentes na mama esquerda e na direita, com posterior classificação individual de tais regiões. Para isso, pares de mamografias são processados de forma que se possa medir a similaridade entre regiões correspondentes do par. Assim, regiões suspeitas são determinadas pelo nível de assimetria encontrado entre regiões da mama esquerda e da mama direita da mesma paciente através do índice de similaridade conhecido como coeficiente de correlação cruzada (*cross-correlation*) e da distância euclidiana. Após esta determinação, o índice de Getis-Ord, em sua forma geral, é utilizado em cada região suspeita isoladamente para extração de características. Posteriormente, estas características são usadas na formação de vetores para treinamento e classificação em "lesão" e "não lesão" com uma Máquina de Vetores de Suporte.

3 FUNDAMENTAÇÃO TEÓRICA

A seguir é apresentada uma breve explanação dos assuntos abordados e das técnicas utilizadas na metodologia deste trabalho. Dentre as quais estão: técnicas de processamento de imagem, medidas estatísticas e de correlação utilizadas e classificação por Máquina de Vetores de Suporte.

3.1 Redução de ruído

Muitas imagens apresentam pequenas falhas conhecidas como ruído, que pode ser adquirido na obtenção da própria imagem ou por causas externas. Uma técnica bastante utilizada para redução de ruído e que se mostra muito eficaz em imagens de mamografia é a aplicação do filtro da mediana (LEE, 1997). A mediana é uma das medidas estatísticas que representam uma tendência central para um conjunto de dados. Com ela, aplica-se um filtro de vizinhança onde o valor de um *pixel* da imagem é substituído pela mediana dos valores dos pixels de sua vizinhança.

3.2 Segmentação por *Watershed*

Imagens de mamografia costumam apresentar elementos chamados de "artefatos", que ficam no fundo da imagem e geralmen-

te contêm alguma informação sobre o exame ou sobre o material utilizado no mesmo. Estes elementos podem prejudicar etapas futuras do processamento das imagens, visto que não representam o objeto de interesse, no caso a mama.

As técnicas utilizadas para isolar o objeto de interesse na imagem são conhecidas como técnicas de segmentação. Na segmentação por *Watershed*, a imagem é comparada a um relevo topográfico, onde os níveis de cinza mais altos são os picos e os níveis de cinza mais baixos são vales. Os vales vão sendo gradativamente inundados até que as águas divididas pelo relevo se encontrem nos picos, onde são formadas as linhas divisórias encontradas pelo algoritmo, descrito com maiores detalhes em Gonzalez; Woods (2002). A técnica faz uso de morfologia matemática para geração dos limites do objeto de interesse. Estes limites podem ser criados com elementos estruturantes de dilatação de 3x3 pixels como descrito em Ericeira (2011). Neste trabalho, aplica-se o algoritmo da mesma forma.

3.3 Registro de imagem

Vários fatores podem contribuir para a existência de diferenças espaciais entre as imagens de mamografia esquerda e direita da mesma paciente, tais como: a estrutura dos tecidos da mama, o posicionamento da mama durante o exame, a compressão aplicada sobre a mama no exame, entre outras. Para tentar diminuir essas diferenças, faz-se uso do que se chama de registro de imagem.

Normalmente em processamento de imagens médicas trabalha-se com registro rígido e registro deformável. O primeiro busca diminuir as diferenças globais entre as imagens, ou seja, aplica transformações de rotação, translação e escala. Já o segundo tem a função de melhorar o resultado obtido com o registro rígido, aplicando deformações na imagem alvo até deixá-la o mais próximo possível da imagem fonte. O registro deformável conhecido com "Demons" (THIRION, 1995) baseia-se na ideia de que uma rede regular de forças é aplicada sobre a imagem deformando os seus contornos.

3.4 Coeficiente de correlação cruzada

O coeficiente de correlação cruzada é uma medida para estimar o grau de correlação entre duas séries ou funções. Em processamento de imagens, estas funções são as próprias imagens (ou regiões das imagens). A medida é recomendada para imagens que tenham sido adquiridas pelo mesmo sensor, como é o caso de um par de imagens mamográficas da mesma paciente.

Sejam *A* e *B* duas imagens (ou duas regiões correspondentes das imagens) compostas pela mesma quantidade de *pixels*, onde a_k é o pixel de índice *k* na imagem *A* e b_k é o pixel de índice *k* na imagem , o coeficiente pode ser usado como uma medida de similaridade entre as imagens e é descrito pela seguinte equação (MITCHELL, 2010):

$$\rho = \frac{\sum_k a_k b_k}{\sqrt{\sum_k a_k^2 \sum_k b_k^2}} \quad (1)$$

Na equação 1, ρ varia de 0 a 1 e quanto mais próximo de 1 mais similares são as imagens. Neste trabalho este valor é utilizado para medir a similaridade entre regiões correspondentes da mama esquerda e da direita da mesma paciente.

3.5 Índice de Getis-Ord

O índice de Getis-Ord é uma medida de associação espacial que, em sua forma geral (GETIS;ORD, 1992), pode ser usado como indicador do grau de associação entre os pixels de uma região da imagem para uma dada distância. Assim, conjunto dos índices de uma região para várias distâncias pode ser usado como um descritor espacial da região. O índice é descrito a seguir:

$$G(d) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(d) x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, \; j \neq \mathrm{i} \quad (2)$$

Na equação 2, G(d) equivale ao quociente entre o somatório de todos os pares de *pixels* (o produto entre esses *pixels*), que estejam até a distancia *d* entre si, pelo somatório de todos os pares de *pixels* possíveis, com o *pixel i* diferente do *pixel j*. A matriz de vizinhança w_{ij} recebe o valor 1 quando o par de pixels (x_i, x_j) está a uma distância menor ou igual a *d* um do outro.

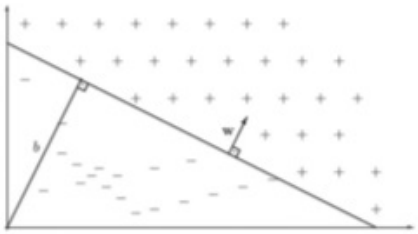
O conceito de distância é relativo e pode ser aplicado de acordo com a forma de utilização da medida. Neste trabalho a distância *d* igual a 1 indica uma janela de *3x3 pixels* ao redor do *pixel* em questão, assim como d igual a 2 determina uma janela de *5x5 pixels*, e assim sucessivamente.

3.6 Máquina de vetores de suporte

Uma Máquina de Vetores de Suporte (SVM), do inglês *Support Vector Machine*, é uma forma de aprendizado de máquina que busca classificar um conjunto de dados (conjunto de teste) a partir de um modelo baseado nas características obtidas de outro conjunto (conjunto de treinamento). Sendo esses dados

dispostos em um espaço dimensional e pertencendo a classes de dados distintas, tenta-se determinar um hiperplano (uma função) que separe ao máximo estas classes uma da outra, como na figura 1.

Figura 1 - Espaço vetorial contendo as duas classes linearmente separáveis



Fonte: Almeida et. al. (2012)

Um conjunto de dados pode ser linearmente separável, quando seus dados podem ser separados por um hiperplano, ou não. É possível o trabalho com conjuntos não linearmente separáveis pelo uso de funções chamadas *kernel*, capazes de realizar o mapeamento dos dados para um espaço de dimensão mais elevada. No presente trabalho, a função kernel utilizada foi a *Radial Basis Function* (RBF) da equação 3, que geralmente apresenta bons resultados, como analisado em Keerthi; Lin (2003).

$$K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right), \gamma > 0 \quad (3)$$

4 METODOLOGIA

A metodologia tem por objetivo detecção de regiões lesionadas em mamografias, partindo da determinação de pares de regiões suspeitos correspondentes na mama esquerda e na direita, com posterior classificação individual de tais regiões. Está dividida em sete etapas, a saber: aquisição das mamografias, pré-processamento, registro das mamografias, divisão das mamografias em regiões de 16x16 pixels, determinação de regiões suspeitas com coeficiente de correlação cruzada (CCC) e distância euclidiana (DE), extração de características com Índice de Getis-Ord e, finalmente, classificação com SVM. A figura 2 demonstra as etapas da metodologia e a saída de cada uma.

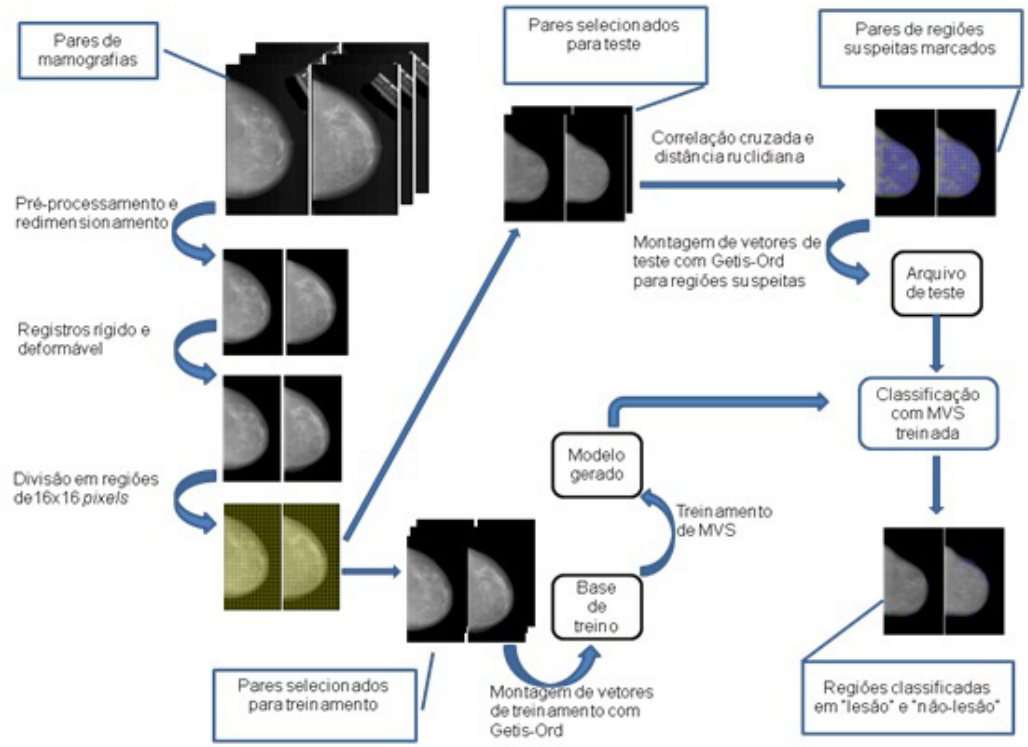
4.1 Aquisição das mamografias

Foram obtidas imagens de base mamografias disponibilizada na Internet para propósitos de pesquisa e bastante utilizada pela comunidade científica, o *Digital Database for Screening Mammography* (DDSM), que contém 2620 casos adquiridos a partir de hospitais e instituições (HEATH et al., 2001). Foram selecionados no total 499 pares de imagens do tipo crânio-caudal (CC).

4.2 Pré-processamento das mamografias

Todas as imagens passaram por processos de redução de ruído e segmentação e depois

Figura 2 – Representação gráfica da metodologia.



Fonte: Sales (2012).

foram redimensionadas, utilizando-se a biblioteca de programação para visão computacional OpenCV (BRADSKI; KAEHLER, 2008).

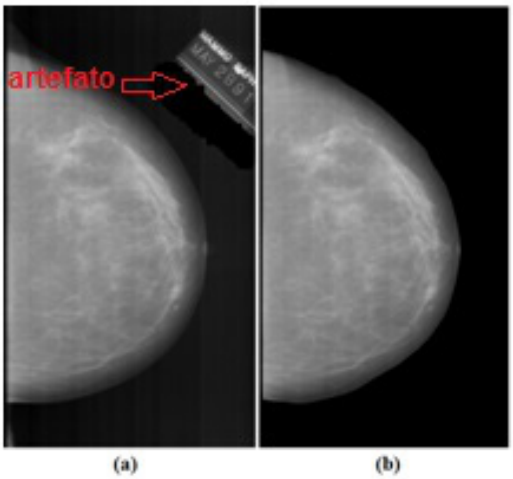
4.2.1 Redução de ruído

Para a redução de ruído foi aplicado um filtro de mediana de tamanho 5x5 pixels em toda a imagem. Este processo é realizado devido à presença de eventuais ruídos na imagem adquiridos, possivelmente, durante a digitalização.

4.2.2 Segmentação da mamas

O processo de segmentação das mamas utilizou uma implementação da técnica de *Watershed*, com aplicação de operador morfológico de dilatação, para estimar a borda da mama. Esta etapa visa a isolar o objeto de interesse, no caso a mama, além de aumentar a precisão e rapidez do registro. A figura 3 mostra o resultado da etapa de segmentação, destacando em (a) o artefato removido.

Figura 3 – Mamografia antes e após segmentação



(a) Mamografia antes da segmentação, destacando do artefato. (b) Mamografia após segmentação. Fonte: Elaborado pelos autores

4.2.3 Redimensionamento das mamografias

Etapa em que todas as imagens foram redimensionadas para o tamanho de 512 *pixels* de largura com uma altura proporcional à original (às imagens originais possuem em média dimensões da ordem de 4000 *pixels* de largura por 6000 *pixels* de altura aproximadamente). A perda de informação causada pela redução das mamografias não demonstrou ter grande impacto na metodologia, comparado ao tempo ganho com o processamento das imagens reduzidas.

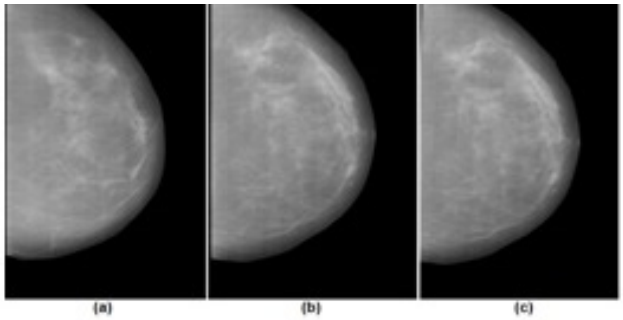
4.3 Registro das mamografias

Esta etapa é dividida em registro rígido e registro deformável, ambos tentam transformar a mama direita tendo a esquerda como base. Para aplicação de ambas as técnicas de registro, foi utilizado o conjunto de ferramentas *The Insight Segmentation and Registration Toolkit* (ITK) (ITK, 2012).

4.3.1 Registro Rígido

Foram utilizadas no registro rígido (Figura 4) transformações de escala, rotação e translação com intuito de diminuir as diferenças globais entre mama esquerda e mama direita, como posição das mamas, rotação e tamanho, que podem ter sido criadas durante realização do exame.

Figura 4 – Registro rígido da mama direita com base na mama esquerda



(a) Mama esquerda.(b) Mama direita. (c) Mama direita após registro rígido. Fonte: Elaborado pelos autores

4.3.2 Registro deformável

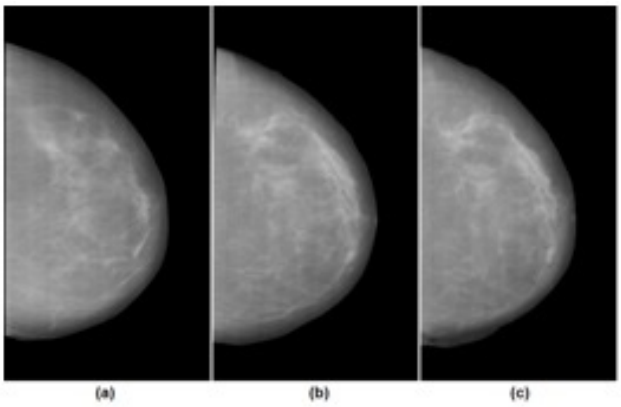
O registro deformável visa melhorar a saída do registro rígido (Figura 5), visto que as transformações lineares aplicadas por este modificam somente alguns aspectos (como posicionamento e escala) da mamografia, e foi feito aplicando-se uma implementação do algoritmo "Demons".

4.4 Divisão das mamografias em regiões de 16x16 pixels

As imagens foram divididas em regiões de 16x16 pixels , de forma que se pudesse fazer a correspondência entre regiões (somente regiões internas às mamas) nas mamas esquerda e direita da mesma paciente. Este tamanho de região foi utilizado por ter sido constatado empiricamente que com tamanhos maiores muitas lesões ficavam localizadas em fronteiras de regiões, o que dificultava a caracterização das regiões como área lesionada. Já tamanhos menores reduziam muito as ca-

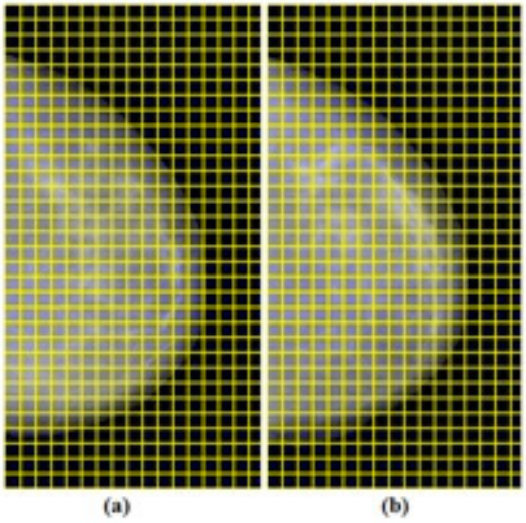
racterísticas de vizinhança de cada região. A divisão em regiões pode ser vista na figura 6.

Figura 5 – Registro rígido e registro deformável da mama direita com base na mama esquerda



(a) Mama esquerda. (b) Mama direita após registro rígido. (c) Mama direita após registro deformável.
Fonte: Elaborado pelos autores

Figura 6 – Divisão das mamografias em regiões de 16x16 pixels.



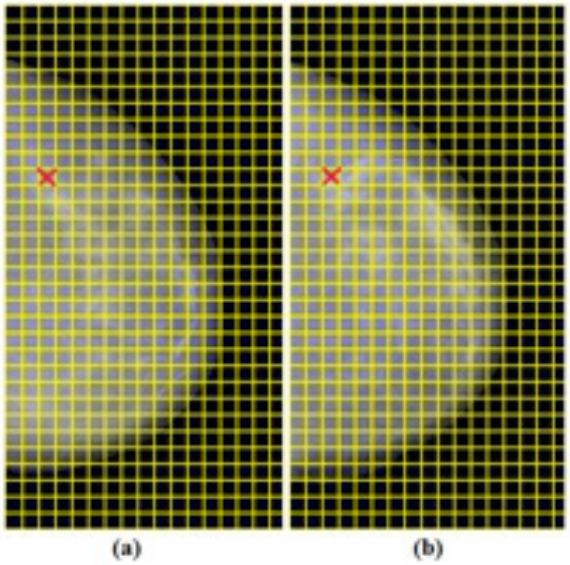
(a) Mama esquerda. (b) Mama direita.
Fonte: Elaborado pelos autores

4.5 Determinação de regiões suspeitas com CCC e DE

Esta etapa buscou classificar previamente os pares de regiões de cada par de mamografias em suspeitos (mais assimétricos) e não suspeitos (mais simétricos). Para cada par de regiões é calculado um valor de *CCC* e um valor de DE. Um exemplo de regiões correspondentes entre as mamas direita e esquerda é destacado na figura 7.

As variáveis *a* e *b* da equação 1, mostrada anteriormente, são os níveis de cinza dos pixels correspondentes na região esquerda e na direita. A figura 8 exemplifica a correspondência entre *pixels* de um par de regiões hipotéticas.

Figura 7 – Regiões correspondentes no par de mamografias.

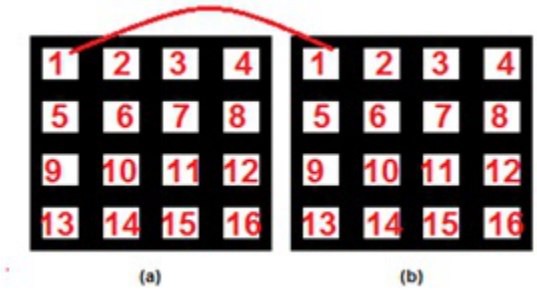


(a) Mama esquerda. (b) Mama direita. Regiões correspondentes marcadas com um X vermelho.
Fonte: Elaborado pelos autores

Uma vez calculados os valores de CCC e de DE para cada par de regiões, usam-se os valores medianos de cada medida como limiares de similaridade. O fato de a mediana ser menos influenciada por valores extremos do que, por exemplo, a média, foi determinante na escolha da mesma para uso como limiar de similaridade.

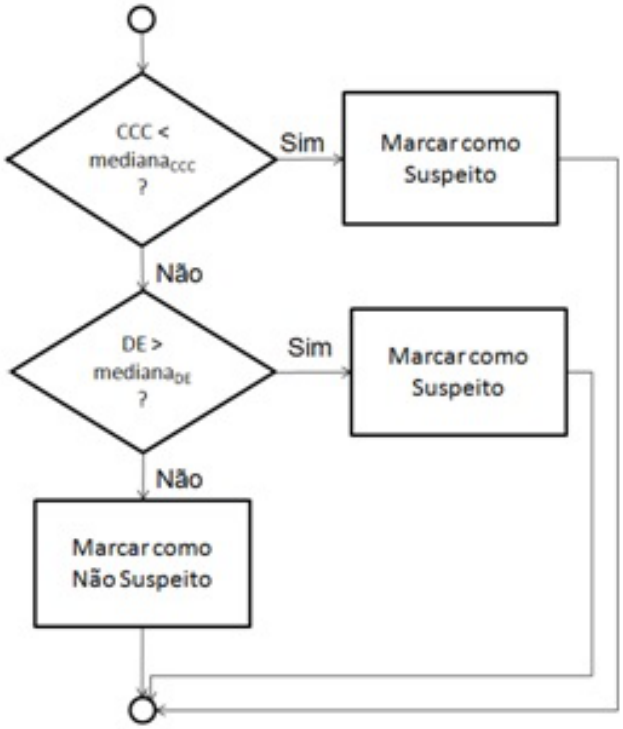
Quanto maior o grau de assimetria, menor o valor de *CCC* e maior a DE. Assim, os pares de regiões classificados como suspeitos de imediato foram aqueles com valores abaixo da mediana de *CCC*. Porém, há regiões visivelmente assimétricas que possuem altos valores de CCC mas com uma distribuição de níveis de cinza bem homogênea. Nestes casos usa-se, além do CCC, a mediana da DE como contraprova, sendo que os pares com DE acima da mediana da mesma medida para o par de mamas são classificados como suspeitos. O restante dos pares de regiões é considerado não suspeito. Este processo pode ser visualizado na figura 9.

Figura 8 – Correspondência entre pixels de regiões hipotéticas de 4x4 *pixels*



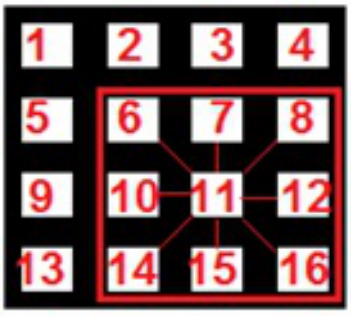
(a) Região esquerda. (b) Região direita.
Fonte: Sales (2012)

Figura 9 – Determinação de regiões suspeitas por assimetria usando CCC e DE



Fonte: Sales (2012).

Figura 10 – Distância 1 para cálculo de índice de Getis-Ord.



Obs: Pares possíveis com pixel 11 para distância 1 em destaque. Fonte: Sales (2012).

4.6 Extração de características com índice de Getis-Ord.

Para o cálculo do índice de Getis-Ord de uma região são considerados todos os pares de pixels possíveis dentro da mesma região para distâncias de 1 a 14 *pixels*. Aqui a distância é considerada como uma janela ao redor do *pixel* que cresce na vertical e na horizontal como mostrado na figura 10, uma região hipotética de 4x4 *pixels* (distância 1 ao redor do pixel 11).

De acordo com a Equação 2, se na iteração do cálculo o valor de x_i estiver no nível de cinza do pixel 11, os pares (x_i, x_j) incluídos na equação serão os níveis de cinza dos pixels (11,6), (11,7), (11,8), (11,10), (11,12), (11,14), (11,15) e (11,16). A distância 15, para regiões de 16x16 *pixels*, que englobaria todos os pixels da região, não foi considerada, visto que ela sempre geraria um índice de valor 1 para qualquer caso, portanto não serviria como medida de distinção. No seu lugar foi utilizada a média dos níveis de cinza da região, para ajudar a distinguir regiões homogêneas com uma diferença muito grande entre os seus níveis de cinza (pois, regiões homogêneas tendem a possuir índices de Getis-Ord parecidos). Os valores do índice de Getis-Ord variam de 0 a 1 e a média dos níveis de cinza foi colocada na mesma escala.

Desta forma, para cada região isolada (não mais aos pares), montou-se um vetor com 15 características, sendo as 14 primeiras os índices de Getis-Ord para distâncias de 1 a 14 e a última a média dos níveis de cinza da região. Com os vetores de características das regiões obtidas a partir dos casos selecionados para treino, foi formada uma base de treinamento para uma SVM. Já para os casos selecionados para teste foram gerados vetores de características somente para as regiões consideradas suspeitas.

4.7 Classificação final com SVM

Foi gerado um modelo a partir da base de treinamento para ser usado na classificação final das regiões suspeitas em lesão e não lesão. Nesta etapa os vetores de características das regiões suspeitas isoladas são submetidos à SVM treinada. Foi utilizado o software *LIBSVM* (CHANG; LIN, 2011) para treinamento e classificação com SVM.

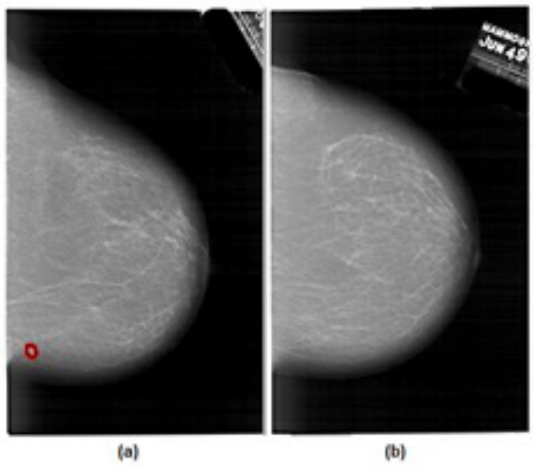
5 RESULTADOS

Nesta seção são apresentados e discutidos os resultados de testes realizados para validação da metodologia proposta na seção 4. Primeiramente os pares de regiões são determinados como suspeitos e não suspeitos. Posteriormente, as regiões dos pares suspeitos são submetidas individualmente à segunda fase para classificação em lesão e não lesão.

5.1 Pares de regiões suspeitas

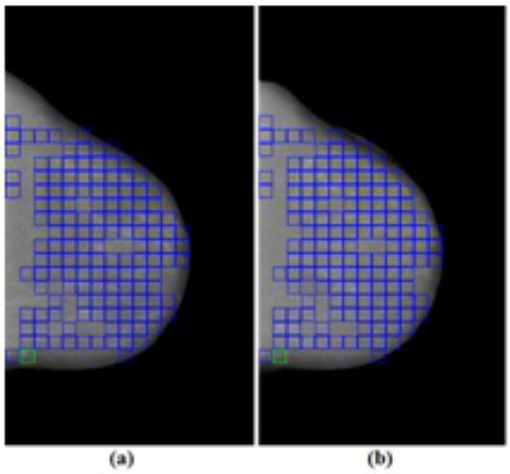
Foram selecionados 60 pares de mamografias como casos de testes, sendo que 30 deles apresentavam algum tipo de lesão em uma das mamas e os outros 30 casos apresentavam diagnóstico normal, todos na projeção crânio-caudal. Todas as imagens passaram pelas etapas de pré-processamento, registro e divisão em regiões. São discutidos nesta seção os resultados de quatro pares de imagens.

Figura 11 – Caso A1639 do banco DDSM



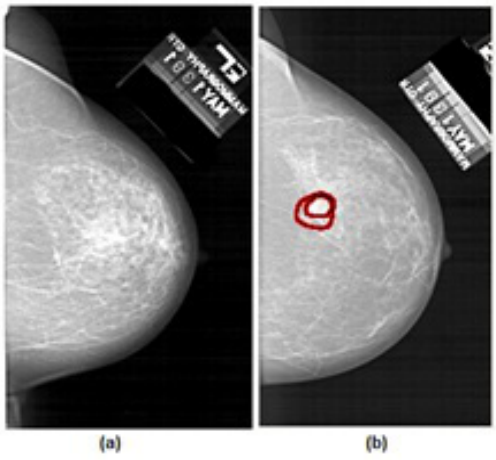
(a) Mama esquerda com área da lesão destacada em vermelho. (b) Mama direita. Fonte: Adaptado do banco DDSM (HEATH et al., 2001)

Figura 12 – Regiões classificadas como suspeitas no caso A1639 do banco DDSM



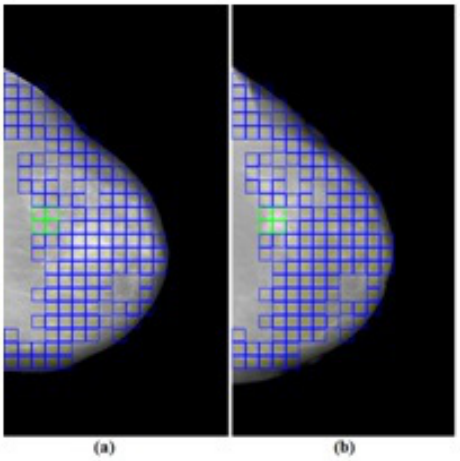
(a) Mama esquerda. (b) Mama direita. Destacam-se os pares de regiões considerados suspeitos em azul e verde. O par de regiões verde corresponde à lesão. Fonte: elaborado pelos autores.

Figura 13 – Caso A1134 do banco DDSM



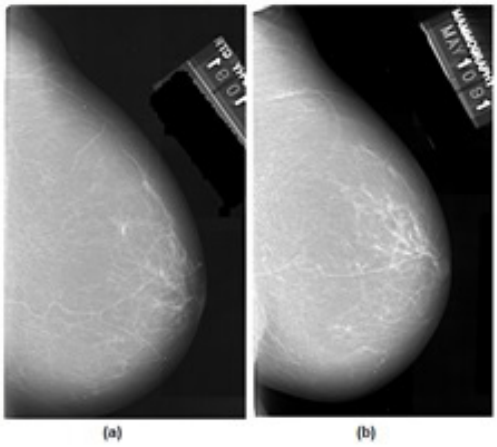
(a) Mama esquerda. (b) Mama direita com área lesionada destacada em vermelho. Fonte: Adaptado do banco DDSM (HEATH et al., 2001)

Figura 14 – Regiões classificadas como suspeitas no caso A1134 do banco DDSM



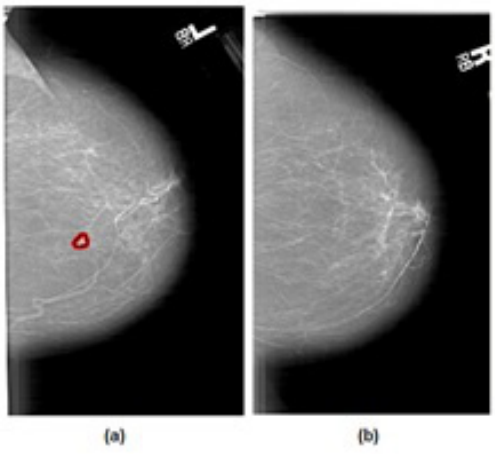
(a) Mama esquerda. (b) Mama direita. Destacam-se os pares de regiões considerados suspeitos em azul e verde. Os pares de regiões em verde correspondem à lesão. Fonte: Elaborado pelos autores

Figura 15 – Caso A0074 do banco DDSM



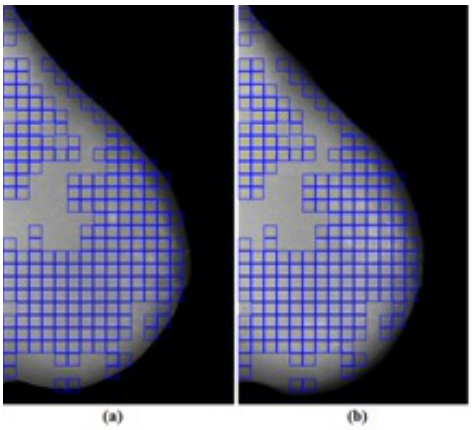
(a) Mama esquerda. (b) Mama direita. Fonte: Adaptado do banco DDSM (HEATH et al., 2001)

Figura 16 – Caso A1627 do banco DDSM



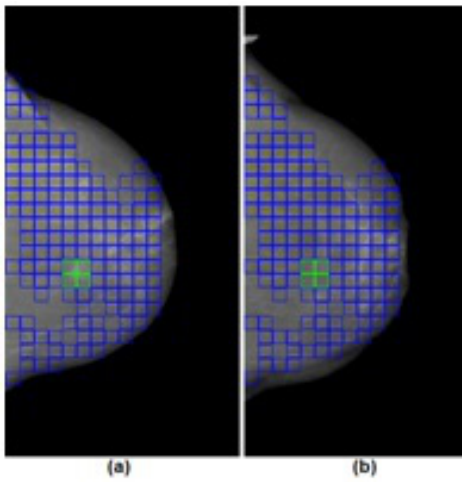
(a) Mama esquerda com área lesionada destacada em vermelho. (b) Mama direita. Fonte: Adaptado do banco DDSM (HEATH et al., 2001)

Figura 17 – Regiões classificadas como suspeitas no caso A0074 do banco DDSM



(a) Mama esquerda. (b) Mama direita. Destacam-se os pares de regiões considerados suspeitos em azul. Fonte: Elaborado pelos autores

Figura 18 – Regiões classificadas como suspeitas no caso A1627 do banco DDSM



(a)Mama esquerda. (b)Mama direita. Destacam-se os pares de regiões considerados suspeitos em azul e verde. Os pares de regiões em verde correspondem à lesão. Fonte: Elaborado pelos autores

5.1.1 Pares de regiões suspeitas: teste 1

O par de mamografias A1639 do banco DDSM é exibido na figura 11, com a área lesionada marcada em vermelho pelo especialista na mama esquerda.

A figura 12 exibe os pares de regiões considerados mais assimétricos (suspeitos), incluindo a área marcada como lesão no banco DDSM.

5.1.2 Pares de regiões suspeitas: teste 2

Outro par de imagens usado como teste foi o caso A1134 do banco DDSM. A figura 13 mostra o par de mamografias, com a lesão na mama direita marcada pelo especialista em vermelho. Já a figura 14 mostra o par com as regiões suspeitas marcadas.

5.1.3 Pares de regiões suspeitas: teste 3

Este é o caso A0074 (Figura 15) e possui diagnóstico normal no banco DDSM. Na figura 16 são mostradas as regiões consideradas suspeitas.

5.1.4 Pares de regiões suspeitas: teste 4

O quarto teste corresponde ao caso A1627 do banco DDSM. Na figura 17 pode ser vista a área marcada como lesão pelo especialista e na figura 18 identificam-se as regiões consideradas suspeitas pela metodologia.

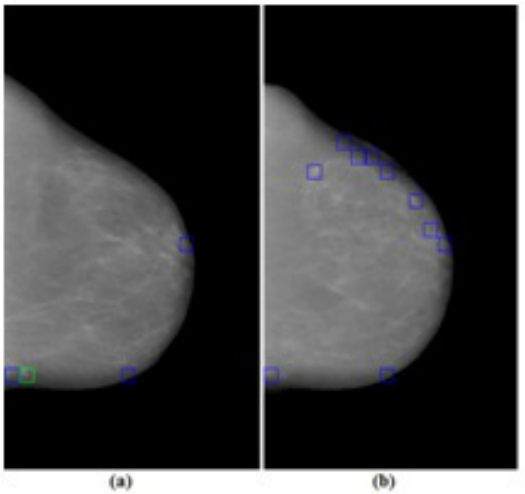
5.2 DETECÇÃO DE REGIÕES COM LESÃO

A partir de outros 439 pares de mamografias foi montada uma base de vetores de características. Foram extraídos 5424 vetores de características de regiões com e sem lesão, na proporção de 1:2, respectivamente. Com estes vetores foi criado um modelo de treinamento para a SVM. Os mesmos casos usados na detecção de regiões suspeitas foram submetidos (agora cada mama individualmente) à SVM treinada para classificação das regiões suspeitas em lesão e não lesão.

5.2.1 Detecção de regiões com lesão: teste 1

As regiões consideradas suspeitas do caso A1639 do banco DDSM foram submetidas à SVM treinada para classificação em "lesão" e "não lesão". O resultado pode ser visto na figura 19. As regiões em azul e verde foram classificadas como lesão, sendo que a região em verde corresponde à lesão marcada no DDSM

Figura 19 – Classificação final do caso A1639

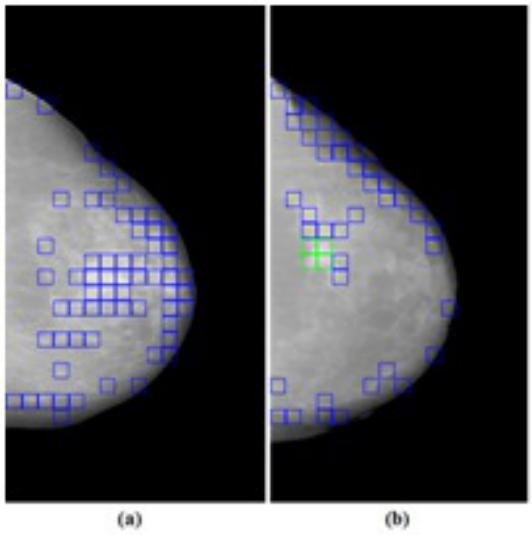


(a)Mama esquerda.(b) Mama direita. Regiões classificadas como "lesão" em azul e verde. A região em verde correspondente à lesão marcada no DDSM. Fonte: Elaborado pelos autores

5.2.2 Detecção de regiões com lesão: teste 2

Na figura 20 pode ser visto o resultado da submissão à SVM treinada das regiões suspeitas do caso A1134.

Figura 20 – Classificação final do caso A1134

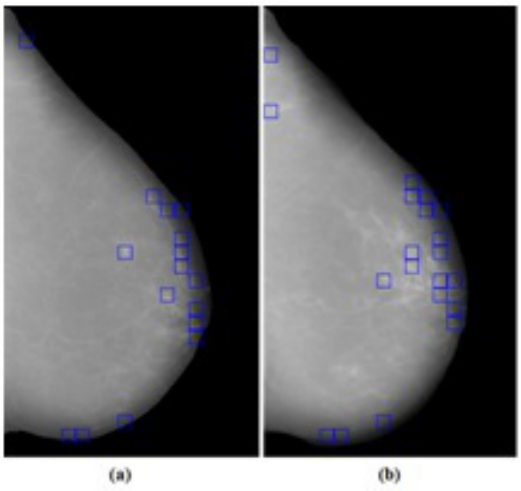


(a)Mama esquerda. (b)Mama direita. Regiões classificadas como "lesão" em azul e verde. As regiões em verde correspondem à lesão no DDSM. Fonte: Elaborado pelos autores

5.2.3 Detecção de regiões com lesão: teste 3

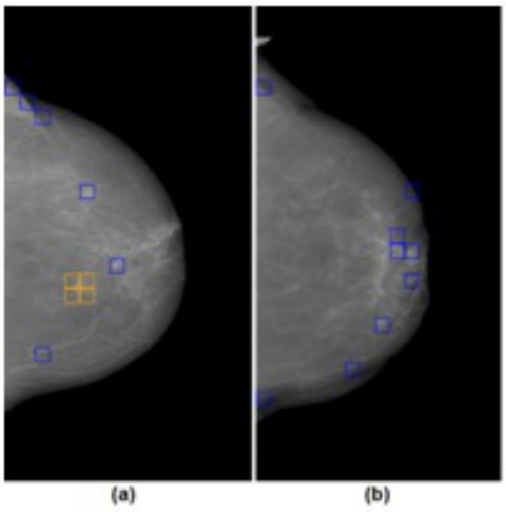
Na figura 21 é mostrada a classificação final do caso A0074 do banco DDSM, que possui diagnóstico normal. Como este caso não possui lesão em nenhuma das mamas, todas as regiões destacadas neste par de mamografias foram classificadas erroneamente como "lesão" pela SVM.

Figura 21 – Classificação final do caso A0074



(a)Mama esquerda.(b)Mama direita. Regiões classificadas como "lesão" em azul. Não é indicada lesão no banco DDSM. Fonte: Elaborado pelos autores

Figura 22 – Classificação final do caso A1627



(a) Mama esquerda. (b) Mama direita. Regiões classificadas como "lesão" em azul. Regiões com lesão no DDSM, porém não detectadas, em laranja.
Fonte: Elaborado pelos autores

5.2.4 Detecção de regiões com lesão: teste 4

No caso A1627, a lesão não foi detectada (regiões destacadas em laranja). Na Figura 22, pode-se observar que a lesão está localizada na fronteira entre 4 regiões. Como estas regiões possuem, cada uma, pequena parte da lesão, ficaram pouco caracterizadas como área lesionada, provocando uma classificação errada pela SVM.

5.3 ANÁLISE DOS RESULTADOS

A metodologia visa a classificar, de modo geral, regiões em lesão e não lesão a partir da busca prévia de regiões assimétricas. Foi submetido à metodologia um total de 23730 regiões internas de mamografias (total de regiões individuais internas às mamas dos 60 pares de mamografias), onde 23549 possuíam diagnóstico normal e 181 possuíam algum tipo de lesão (de acordo com o DDSM). A metodologia demonstrou 80,11% de sensibilidade, 84,41% de especificidade e 84,38% de acurácia. A análise pode ser conferida na tabela 1.

6 CONSIDERAÇÕES FINAIS

Este trabalho apresentou o desenvolvimento de uma metodologia de detecção de regiões lesionadas em mamografias, utilizando Índice de Getis-Ord na sua forma geral, partindo da determinação inicial de pares de regiões suspeitos através do uso conjunto das medidas CCC e DE entre as regiões de cada par. Foi percebido que a metodologia apresentou 80,11% de sensibilidade, 84,41% de especificidade e 84,38% de acurácia. Com o intuito de melho-

Tabela 1 – Resultados gerais da metodologia

INTERNAS	NORMAIS	LESÃO	FP	FN	VP	VN	SE(%)	ES(%)	AC(%)
23730	23549	181	3669	36	145	19880	80,11	84,41	84,38

INTERNOS: regiões internas às mamas.
NORMAIS: regiões normais, segundo DDSM.
LESÃO: regiões com lesão, segundo DDSM.
FP: falso-positivos (regiões classificadas erroneamente como lesão).
FN: falso-negativos (regiões classificadas erroneamente como não lesão).
VP: verdadeiro-positivos (regiões classificadas corretamente como lesão).
VN: verdadeiro-negativos (regiões classificadas corretamente como não lesão).
SE: sensibilidade (capacidade de identificar regiões com lesão).
ES: especificidade (capacidade de identificar regiões normais).
AC: acurácia (taxa de acerto).

Fonte: Elaborado pelos autores

rar os resultados, trabalhos futuros podem considerar o uso de outras medidas estatísticas em conjunto com as utilizadas ou separadamente, tanto na determinação de regiões suspeitas quanto na detecção de lesões com um classificador, assim como a distinção entre áreas de massa e não massa, ou quanto à determinação da natureza de lesões (benignas ou malignas).

REFERÊNCIAS

ACS – American Cancer Society. *What are the key statistics about breast cancer?*, 2012. Disponível em: <http://www.cancer.org/Cancer/BreastCancer/DetailedGuide/breast-cancer-key-statistics >. Acesso em: 31 maio 2012.

ALMEIDA, T. S. et al. Algoritmo para detecção de boca em faces humanas usando matriz de Co-ocorrência e SVM. *Cadernos de Pesquisa*, v. 19, n. esp., jul. 2012.

BRADSKI, G; KAEHLER, A. Learning OpenCV: computer vision with the openCV library, O'Reilly Press, Oct., 2008.

BRAZ JUNIOR, G. *Classificação de regiões de mamografias em massa e não massa usando estatística espacial e máquina de vetores de suporte*. Dissertação (Mestrado em Engenharia de Eletricidade) - Universidade Federal do Maranhão, São Luís, 2008.

CHANG, C. C; LIN, C.J. LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

ERICEIRA, D. R. *Detecção de regiões suspeitas e classificação de massas em mamografias digitais utilizando descrição espacial com função variograma*. Dissertação (Mestrado em Engenharia de Eletricidade) - Universidade Federal do Maranhão, São Luís, 2011.

GETIS, A.; ORD, J. K. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, v. 24, n. 3, p. 189–206, July 1992.

GONZALEZ, R. C.; WOODS,R. E. *Digital image processing*. Prentice Hall. Ed. 2, 2002.

HEATH, M. et al. The digital database for screening mammography. In: Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed., 212-218, *Medical Physics Publishing*, 2001.

INSTITUTO NACIONAL DO CÂNCER. *Detecção precoce do câncer de mama*, 2012. Disponível em: <http://www.inca.gov.br/conteudo_view.asp?id=1932>. Acesso em: 31 maio 2012.

_____. *Estimativa* 2012: Incidência de câncer no Brasil, 2011 Disponível em: <http://www1.inca.gov.br/estimativa/2012/>. Acesso em: 31 janeiro 2013.

ITK – *The Insight Segmentation and Registration Toolkit*. Disponível em: <www.itk.org>. Acesso em: 31 maio 2012.

KEERTHI, S. S.; LIN C. J. Asymptotic behaviors of support vector machines with Gaussian Kernel. *Neural computation*, v.15, n. 7, p. 1667-1689, July 2003.

LEE, K. A. *A mammographic registration method based on optical flow and multiresolution computing*. Thesis, School of Engineering of the Air Force Institute of

Technology, United States, 1997.

MITCHELL, H. B. *Image fusion*: theories, techniques and applications. Springer. 2010.

ROCHA, S. V. et a. Detecção e diagnóstico de massas em mamografia: revisão bibliográfica. *Cadernos de Pesquisa*, v. 18, n. esp., dez 2011. Disponível em : <http://www.periodicoseletronicos.ufma.br/index.php/cadernosdepesquisa/article/view/735>. Acesso em: 31 maio 2012.

RODRIGUES, E. P. *Avaliação de métricas para o corregistro não rígido de imagens médicas*. 2010. Tese (Doutorado em Física Aplicada à Medicina e Biologia) - Universidade de São Paulo, Ribeirão Preto, 2010.

SALES, A. M. V. *Detecção de regiões com lesão em mamografias usando coeficiente de correlação cruzada, distância euclidiana e índice de Getis-Ord*. Monografia (Curso de Ciências da Computação) - Universidade Federal do Maranhão, São Luís, 2012.

SCUTT, D.; LANCASTER, G. A.; MANNING, J. T. Breast asymmetry and predisposition to breast cancer. *Breast Cancer Research*, 8:R14, licensee BioMed Central Ltd., 2006.

THIRION, J.P. Fast Non-Rigid Matching Of 3D Medical Image. Technical report, *Research Report* RR-2547, Epidure Project, INRIA Sophia, May 1995.