Análise e comparação de algoritmos de classificação para o diagnóstico de câncer de mama.

Jonas F. Silva, Tácito Neves
Núcleo de Ciências Exatas(NCEx)– Universidade Federal de Alagoas Arapiraca, AL – Brasil
jonassilvaferreira@gmail.com, tácito.neves@arapiraca.ufal.br

Abstract. *The medical diagnosis based on images presents several computational challenges, including the phases of acquisition, pre-processing, segmentation and classification of the images. This work starts from the analysis of a database of magnetic resonance images of breast tumors (mammograms) and implementation of machine learning algorithms for the comparison of the best classifying model for the diagnosis of breast cancer.*

Resumo. *O diagnóstico médico baseado em imagens apresenta diversos desafios computacionais, incluindo as fases de aquisição, de pré-processamento, segmentação e classificação das imagens. este trabalho parte da análise de um banco de dados de imagens de ressonância magnética de tumores de mama (mamografias) e implementação de algoritmos de aprendizado de máquina para a comparação do melhor modelo classificador para o diagnóstico do câncer de mama.*

1. Introdução

O câncer de mama é o tipo que mais acomete as mulheres em todo o mundo, sendo 2,09 milhões de novos casos e 627 mil mortes por ano, de acordo com a Organização Pan-americana da Saúde [OPAS 2018]. A proporção entre homens e mulheres é de 1:100. Ou seja, para cada 100 mulheres com câncer de mama, um homem terá a doença. No Brasil, o Ministério da Saúde, a partir do INCA, estima 59.700 casos novos em um ano [INCA 2019]. Segundo dados da Sociedade Brasileira de Mastologia [SBM 2019], 1 em cada 12 mulheres terão um tumor nas mamas até os 90 anos de idade. O diagnóstico pode ser feito por meio de mamografia, ressonâncias magnética, ecografia, entre outros exames.

O meio mais comum de diagnóstico é o uso da mamografia sendo realizado quase que completamente de forma manual. Utilizando as imagens das mamografias, o profissional procura identificar os nódulos. Assim a ressonância magnética é um instrumento usado frequentemente no diagnóstico de tumores em tecidos moles, como é o caso do tecido mamário. Por meio da análise visual, os especialistas são capazes de identificar a existência de tumores nas imagens e, com base em informações visuais pertencentes à atividade biológica do tumor, como seu tamanho ou coloração, definir o seu tipo, natureza, características, e, em especial, concluir se o nódulo se trata de um tumor maligno ou benigno.

Com base nisto, e em esforços anteriores de classificação de tumores em tecidos moles em geral, o presente trabalho se propõe a mostrar uma comparação de desempenho

de diferentes técnicas e algoritmos criados para classificação individuais e combinados dos nódulos.

2. **Banco de dados, pré-processamento e extração de features.**

O banco de dados usado para este trabalho foi o DDSM, Digital Database for Screening Mammography[FLORIDA 2020]. O banco de dados contém 2620 casos contendo diversas imagens de mamografia de tecidos sem tumores e com tumores malignos e benignos. Para fins de classificação, foram apenas utilizadas imagens com tumores confirmados benignos ou malignos.

No início do desenvolvimento do projeto foi utilizado imagens disponibilizadas no dataset DDSM e máscaras para segmentação dos tumores para isolar as regiões de interesse (ROIs) demarcadas por especialistas.

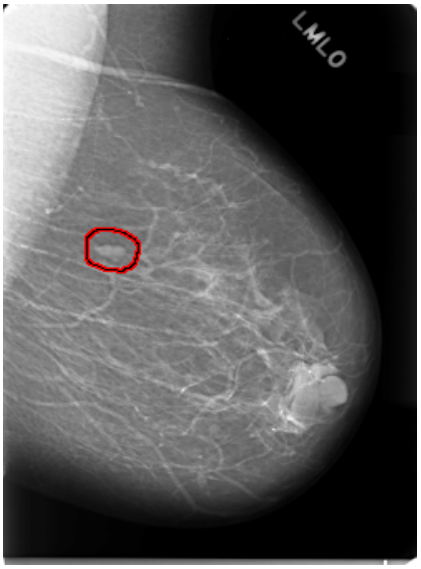


Figura 1 – imagem do dataset a área de interesse marcada por especialistas

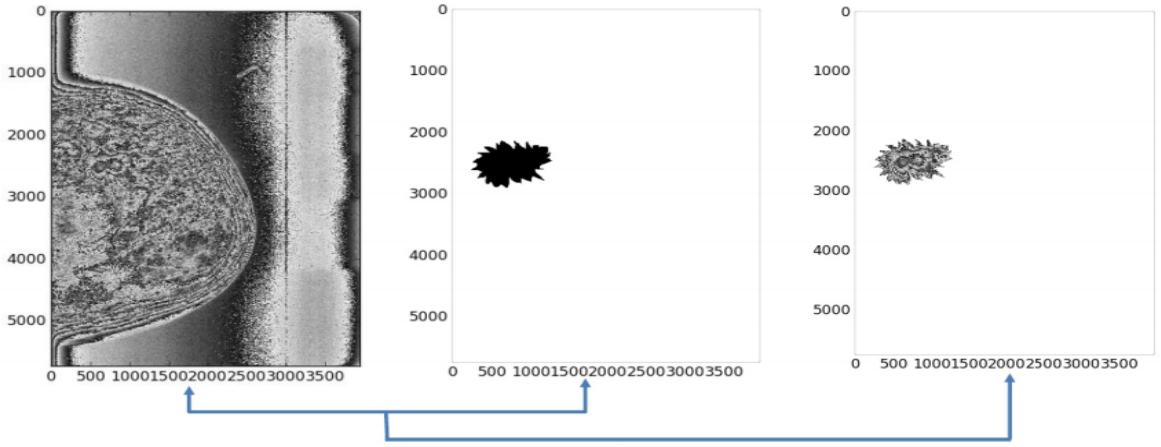
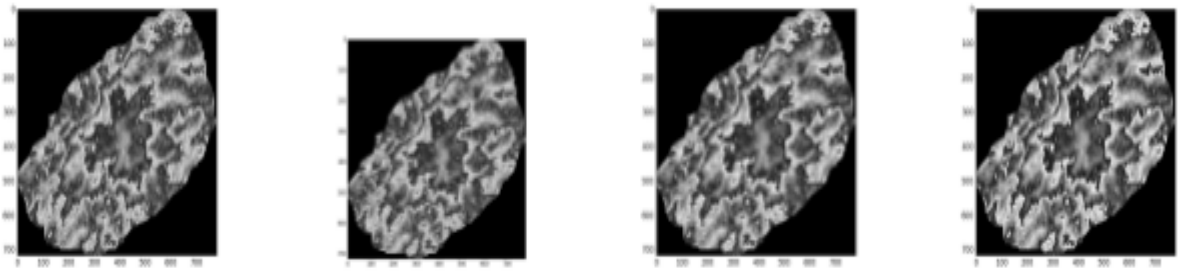


Figura 2 – isolamento das áreas de interesse

Entretanto, devido à qualidade das imagens, que são produtos de um procedimento de digitalização de imagens analógicas, foi preciso o emprego de técnicas de redução de ruído e de aumento de contraste das regiões de interesse como etapa prévia à extração de características a serem utilizadas no processo de aprendizado.



(a) Original (b) Média 1 (c) Média 3 (d) Média 5

Figura 3 – Imagens de um tumor maligno, original e filtradas

As imagens disponibilizadas no dataset foram transformadas do formato LJPEG (Lossless JPEG) para bitmaps de 16 bits por pixel antes de começar a extração de features. posteriormente, através do emprego de máscaras, as áreas de interesse das imagens foram isoladas, redimensionadas, e passaram por um filtro de redução de ruído (filtro de média 3x3) e, por fim, foi efetuado um ajuste de histograma com o objetivo de aumentar o contraste de diferentes partes da imagem. Neste último processo, foi utilizada a técnica de equalização de histograma CLAHE (Contrast Limited Adaptive Histogram Equalization).
Após este processo, as imagens foram convertidas para profundidade de 8 bits e são extraídas características de textura para classificação das regiões de interesse. Foi optado por empregar as características Haralick[HARALICK, 1979]., após experimentar com algumas outras características, tais como o LBP[OJALA; HARWOOD 1996], que não forneceram bons resultados.
Foram extraídos dois grupos de features das imagens pré-processadas: um conjunto de features de primeira ordem baseado no histograma de tons de cinza da imagem; e o conjunto de features Haralick[HARALICK 1979]. A extração de features foi realizada, em períodos diferentes, nas imagens completas, cortadas pelo tamanho da ROI e também após a aplicação de filtros de mediana para diminuir a granularidade das imagens.

3. Algoritmos utilizados.

Foram considerados e comparados diversos algoritmos de classificação na realização deste projeto, sendo que alguns já foram implementados e testados que são os que serão explanados aqui e outros ainda serão feitos e treinados e testados.

Os classificadores produzidos são versões implementadas da biblioteca scikit learn [SCIKIT-LEARN 2020]. O primeiro é uma máquina de vetores de suporte(SVM), usando uma função de kernel RBF. Os parâmetros de custo(C) e coeficiente γ foram definidos de forma experimental.

O segundo é uma árvore de decisões implementada com algoritmo C4.5, cujos critérios de avaliação e divisão de galhos foram definidos de maneira experimental.

Também foi utilizado o algoritmo Random Forests, que se serve de muitas árvores de decisão (uma floresta) para atingir resultados melhores utilizando-se apenas uma árvore, o que pode ocorrer desde que exista variação entre as árvores [HERBRICH, 2015].

Também será utilizado um modelo de CNN para a classificação, porém no momento ainda não foi implementado no projeto.

4. Desenvolvimento do projeto

Todo o desenvolvimento do projeto, se deu em uma máquina contendo um core i5 de sétima geração, 8Gb de memória RAM e uma placa de vídeo NVidia GTX 1050ti de 4gb.

4.1. Imagens

O banco de dados usado neste trabalho foi sucessivamente alterado e filtrado de forma a tentar aperfeiçoar a taxa de acerto dos algoritmos de classificação obtidos com base nele.

O primeiro passo foi a conversão dos arquivos de imagem e utilização de uma máscara de bits representando a ROI. Imagens sem uma máscara equivalente foram descartadas como inutilizáveis.

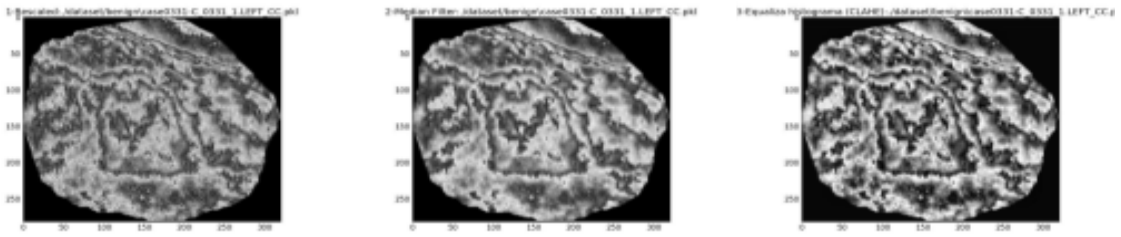
O primeiro conjunto de features extraídos da imagem foi baseado no histograma da imagem, resultando em um vetor de 10 features por instância.

Em seguida foi utilizado o conjunto de features de Haralick, resultando na obtenção de em 13 features por instância, conforme mostrado na tabela abaixo.

Conjunto de features	Descrição
Primeira Ordem	Média, Mínimo, Variância, *Skeweness*, *Kurtosis*, Percentis em 1%, 10%, 50%, 90% e 99%
Haralick	Momento Secundário Angular, Correlação, Soma de Quadrados, Momento do Inverso da Diferença, Medidas de Correlação de Informação, Média, Contraste, Variância e Entropia do Somatório, Entropia, Variância e Entropia da Diferença, Coeficiente de Correlação, Máximo

A fim de tentar aperfeiçoar a classificação e padronizar as imagens, os testes seguintes foram realizados com imagens oriundas do mesmo modelo de scanner (neste caso, o LUMISYS). Isto gerou uma redução no tamanho do dataset, mas a melhora nos resultados não foi proporcional à redução.

O próximo passo foi a utilização de filtros de média com coeficiente variável como mostrado anteriormente na figura 2, de forma a diminuir as variações pixel a pixel da imagem e reforçar a detecção de regiões contíguas. Em seguida, as imagens foram redimensionadas e foi aplicado um filtro CLAHE (Contrast Limited Adaptive Histogram Equalization), de forma a limitar a amplificação de ruído normalmente decorrente do processo de equalização.



(a) Original (b) Média (c) Média + CLAHE

Figura 4 – Imagens de tumores antes e após aplicação de CLAHE

4.2. Treinamento e testes.

A rotina de treinamento e teste se deu a princípio, a partir dos conjuntos de dados que foram separados em duas partições, uma para treinamento e outra para testes, de forma que a partição de testes apresentou aproximadamente 15% do dataset inteiro. Esta divisão foi feita após consecutivas reordenações aleatórias, de forma a garantir que ambas as partes tenham uma relação aproximadamente igual de casos positivos e negativos.

Antes de iniciar a averiguação de valores ótimos para os parâmetros configuráveis, foi feita uma validação cruzada, para avaliar a acurácia média e máxima do algoritmo em relação à base de treinamento.

A investigação de parâmetros foi feita de forma a explorar diferentes conjuntos de valores para os parâmetros variáveis que conseguissem afetar o modelo simulado por cada algoritmo quando exposto às bases de treino, afetando assim a performance do classificador. Os fatores testados foram:

- **SVM:** Parâmetros C e γ da função kernel;
- **Árvore de Decisão C4.5:** altura máxima da árvore, critério de divisão (impureza Gini ou entropia) e critério de avaliação (melhor divisão ou aleatório);
- **Random Forests:** profundidade máxima, número de classificadores, mínimo de amostras para divisão.

Após efetuada a validação cruzada em todo o espaço de possibilidades, foi guardado o melhor conjunto de parâmetros, de acordo com a acurácia média encontrada. Este mesmo conjunto de parâmetros foi então utilizado para uma avaliação iterativa do impacto do tamanho da base de treinamento na taxa de erro estimado e experimental do modelo de classificação.

5. Resultados

Para cada instância pré-processada do dataset, foram executados experimentos e extraídos dados estatísticos acerca do modelo eleito experimentalmente como o melhor para uma dada iteração. Foi dado maior foco a duas estatísticas: de acerto estimado é de acerto experimental.

O acerto estimado foi calculado por meio de execução repetitiva de validação cruzada no mesmo dataset de treinamento, variando-se apenas a ordem dos dados na entrada. Os dados de validação foram armazenados e a média aritmética simples foi calculada em cima destes dados.

O acerto experimental foi calculado usando-se o dataset de teste separado no início do experimento, e para esta medida foi usado apenas o melhor modelo encontrado dentre os resultados das repetidas iterações. Este modelo então é julgado de acordo com sua performance atuando sobre o dataset de teste.

De modo geral, o algoritmo de SVM teve um desempenho melhor que os dos demais, enquanto o random forests teve um desempenho bastante notável em alguns casos.

Descrição do Dataset	Árvore de Decisão	Random Forests	SVM	CNN
Primeira Ordem	*65,49%*	–	*86,26%*	–
Haralick	*71,63%*	–	*87,77%*	–
LUMISYS	*76,39%*	*79,00%*	*86,35%*	–
LUMISYS(ROI)	*71,39%*	–	*79,19%*	–
Filtro Média (3)	*70,30%*	*76,95%*	*81,22%*	–
Rescaled + CLAHE	*72,04%*	–	*82,11%*	–

Melhores resultados obtidos por datasets

Descrição do Dataset	Árvore de Decisão	Random Forests	SVM	CNN
Primeira Ordem	*66,68%*	–	*88,16%*	–
Haralick	*72,21%*	–	*89,14%*	–
LUMISYS	*74,94%*	*83,67%*	*86,27%*	–
LUMISYS(ROI)	*69,89%*	–	*81,29%*	–
Filtro Média (3)	*70,38%*	*81,20%*	*82,52%*	–
Rescaled + CLAHE	*73,44%*	–	*83,26%*	–

Estimativas médias de treinamento por datasets

6. Considerações finais

Embora os resultados obtidos até o momento não tenham sido totalmente satisfatórios, pois tiveram uma taxa de acertos mediana, cumpre observar que a base de imagens trabalhadas foi grande e diversa, assim dificultando o trabalho com ela nos algoritmos já implementados. Porém, como esperado, o tratamento das imagens realizado antes mesmo do início foi um fator impactante no desempenho final dos algoritmos de classificação.

Espera-se que com a implementação e testes do modelo de rede neural convolucional (CNN) a taxa de acertos seja maior e tenha um resultado mais satisfatório.

7. References

OPAS 2018, ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE: Folha informativa - câncer. Disponivel em: https://www.paho.org/bra/index.php

INCA 2019, INSTITUTO NACIONAL DE CÂNCER. Disponível em : https://www.inca.gov.br/tipos-de-cancer/cancer-de-mama

FLORIDA, U. of S. DDSM: Digital Database for Screening Mammography. Disponível em: http://www.eng.usf.edu/cvprg/Mammography/Database.html .

SBM 2019, SOCIEDADE BRASILEIRA DE MASTOLOGIA. Disponível em : http://www.sbmastologia.com.br/

HARALICK, R. M. Statistical and structural approaches to texture. Proc. IEEE, vol. 67, no. 5, pp. 786-804, 1979., 1979.

HERBRICH, G. Machine Learning: An Algorithmic Perspective. second edition. [S.l.]: Chapman & Hall/CRC, 2015.

OJALA, M. P. T.; HARWOOD, D. A comparative study of texture measures with classification based on feature distributions. Pattern Recoonition, Vol. 29, No. l, pp. 51 59, 1996., 1996.