XXI ENMC
ENCONTRO NACIONAL DE MODELAGEM COMPUTACIONAL
IX ECTM
ENCONTRO DE CIÊNCIA E TECNOLOGIA DE MATERIAIS

08 a 11 de Outubro de 2018
Instituto Federal Fluminense
Búzios - RJ

REDES NEURAIS NA CLASSIFICAÇÃO DE NEOPLASIAS MAMÁRIAS

Ana Gabriela da Silva Freitas[1] – a.na.freitas@hotmail.com
Pedro Marcio Ferreira[2] – pedroborgesengen@gmail.com
Robson Mariano da Silva[3] – rsmariano2010@gmail.com
[1 2 3] Universidade Federal Rural do Rio de Janeiro, Instituto de Ciências Exatas, Programa de Pós-Graduação em Modelagem Matemática e Computacional – Seropédica, RJ, Brasil

Resumo. *O presente artigo, trata do reconhecimento de tumores na mama, utilizando para isto, Redes Neurais artificiais, (RNAs) de percepton de múltiplas camadas, (MLP), para tal, foram utilizadas várias características, como: raio, concavidade, fractal, área, perímetro e textura. A rede MLP foi configurada com as entradas, 2 camadas ocultas (7,6), respectivamente, e suas saídas, os valores utilizados são do banco de dados do* Wisconsin Diagnostic Breast Cancer, *os resultados apontaram erro de 11.9% no conjunto de validação, mostrando um bom desempenho do modelo para reconhecimento de tumores de mama.*

Palavras chave: *Inteligência Computacional, Câncer de Mama, Rede Neural Artificial*

1. INTRODUÇÃO

De acordo com informações do Instituto Nacional de Câncer (INCA, 2016), o câncer de mama é uma doença proveniente da multiplicação de células anormais na mama, formadoras de um tumor. Atualmente, este é o segundo tipo de câncer que mais cresce no mundo entre mulheres, ficando atrás apenas dos cânceres de pele não melanoma. No Brasil, 28,1% da população afetada por câncer, apresenta câncer de mama. São vários os tipos dessa anomalia mamária, sendo que um dos fatores que serve para diferenciá-los é o seu tempo de desenvolvimento, enquanto alguns se proliferam com maior rapidez, outros são mais lentos. Segundo Ferreira e Pires (2013), cerca de 20.9% das mulheres em idade fértil (de 15 a 49 anos) no Brasil, vêm a óbito por neoplasias malignas, como o câncer de mama.

Em relatório, a Organização Mundial da Saúde (OMS, 2016), estima que no ano de 2012, 408 mil mulheres receberam um diagnóstico positivo de câncer de mama, e dessas, 92 mil vieram a óbito nas Américas. Para o ano de 2030, existe uma expectativa de que esse número aumente em 46%, tornando-o ainda mais alarmante. Desde então, criou se uma campanha de abrangência mundial, afim de que esses números sejam controlados. O outubro rosa (como é chamada a campanha) alerta mulheres de todo o mundo para uma detecção precoce dessa doença, facilitando sua cura e então reduzindo a mortalidade por essa causa.

Setenta por cento dos diagnósticos de câncer não é realizado por oncologistas, dado que mostra a importância dos médicos não-cancerologistas no controle dessa doença. A suposição de um diagnóstico é feita em várias etapas, dentre as quais há exigência de análise cuidadosa. Esse diagnóstico é realizado por exame clínico e exame de imagens (mamografia, ultrassom ou ressonância). Ao haver desconfiança do câncer, o médico mastologista encaminha a paciente a uma biópsia, método invasivo utilizado para identificar se a alteração é maligna ou benigna.

A fim de evitar este método invasivo, que pode ocasionar traumas físicos e psicológicos à paciente, métodos de Inteligência Computacional (IC) vêm sendo testados, para serem auxiliadores nesse diagnóstico. Ribeiro (2006) apresentou uma proposta de uma metodologia para classificar nódulos mamários por contorno, através das técnicas de Redes Neurais *Multi-Layer Perceptron* (MLP) e *Self-Organizing Map* (SOM), a fim de facilitar o diagnóstico dessa forma, já que existe uma difícil interpretação pelo contorno, já que há uma dificuldade de visibilização e o baixo contraste das imagens mamograficas. Com a Rede Neural MLP, com topologia de 20 neurônios de entrada, 40 na primeira camada intermediária, 20 na segunda e 5 na camada de saída, obteve se um falso negativo de 5% e um falso positivo de 7%. Glingani e Ambrósio (2014), com o intuito de estabelecer padrões, utilizaram uma Rede Neural Artificial MLP, com algoritmos de comparação e de retropropagação de erro no treinamento da rede. O desempenho do modelo foi avaliado com um teste de generalização, sempre deixando um de fora (*leave one out*), que demonstrou a capacidade da rede de aprender, generalizar e classificar microcalcificações mamárias. Lima *et al*. (2018) realizou a aplicação de Máquina de Vetor de Suporte não linear para classificar microcalcificações mamárias, identificadas em exames de imagem. Em 50 simulações realizadas, obtiveram uma acurácia superior a 88%, com sensibilidade de 86% e especificidade de 82%.

Com base nesses estudos, é possível perceber que esta área de pesquisa é bastante promissora, sendo assim, para este trabalho foi utilizada uma Rede Neural Artificial com o objetivo de identificar padrões de tumores cancerígenos através de informações conhecidas, presentes em bancos de dados.

2. REVISÃO BIBLIOGRÁFICA

2.1 Câncer de Mama

Segundo Haddad e Silva (2001) o perfil da morbimortalidade brasileira vem sofrendo uma mudança intensa, passando de doenças infecto-parasitárias a crônico-degenerativas, como o câncer, tendo como principais causas as mudanças nos hábitos de vida e no perfil epidemiológico da população.

O carcinoma mamário tem destaque fundamentado em diversos fatores, dentre os quais pode se citar sua alta incidência, elevado índice de mortalidade, dificuldade em estabelecer um diagnóstico precoce, além da escassez de informações quanto ao seu comportamento biológico.

O câncer de mama é tipo de câncer mais comum entre as mulheres no mundo e no Brasil, depois do de pele não melanoma, respondendo por cerca de 28% dos casos novos a cada ano. Sua maior incidência ocorre em mulheres de 35 a 50 anos e a cada 8 mulheres uma é diagnosticada com câncer de mama. Segundo dados do Instituto Nacional do Câncer, é estimado cerca de 59.000 novos casos de câncer de mama no Brasil nos anos de 2018-2019. (INCA 2018)

Segundo o Instituto Nacional do Câncer (INCA), o câncer de mama é uma doença causada pela multiplicação de células anormais da mama, que formam um tumor. Há vários

tipos de câncer de mama. Alguns tipos têm desenvolvimento rápido enquanto outros são mais lentos. Especificamente no Brasil, o percentual é elevado e chega a 28.1%.

Os principais sinais e sintomas de câncer de mama são nódulos na mama e/ou axila, dor mamária e alterações da pele que recobre a mama, como abaulamentos ou retrações com aspecto semelhante à casca de laranja. Os cânceres de mama localizam-se, principalmente, no quadrante superior externo, e em geral, as lesões são indolores, fixas e com bordas irregulares, acompanhadas de alterações da pele quando em estágio avançado.

Desde o início das pesquisas sobre o câncer de mama, a melhor maneira para cura da doença aproximadamente 95% é detecção precoce (Mavroforakis, 2005). A detecção pode ser realizada através do exame de mamografia, sendo o método mais eficaz para o diagnóstico do câncer de mama disponível hoje. A mamografia é uma forma particular de radiografia capaz de registrar imagens da mama com a finalidade de diagnosticar a presença ou ausência de estruturas que possam indicar a doença. Com esse tipo de exame pode-se detectar o tumor antes que ele se torne palpável.

Distorções de interpretação e classificação dessas lesões por especialistas, implicam em um número mais elevado de biópsias desnecessárias, ou seja, cerca de 65 a 85% das biópsias de mamárias são realizadas em lesões benignas, implicando em uma redução na relação custo benefício dos exames e a não detecção da doença, caracterizando um diagnóstico falso negativo do exame.

2.2 Redes Neurais Artificiais

Por ser um método computacional baseado no funcionamento do cérebro, as Redes Neurais Artificiais (RNA) têm como objetivo principal solucionar problemas com base em uma associação de informações conhecidas. Esta técnica trabalha como classificadora de padrões, com natureza estatística, onde as classes definidas são representadas por pontos em um espaço de decisão multidimensional. Seu processo de treinamento estima limites de decisão, e suas construções realizadas pela variabilidade estatística existente entre as divergentes classes. Várias unidades de processamento compõem esta rede, conhecidos como camadas, onde o valor calculado para cada uma dessas unidades de processamento é totalmente dependente dos valores de entrada e dos pesos atribuídos a si.

A definição de uma Rede Neural como sendo um processador maciço paralelamente distribuído, composto por elementos de processamento simples, que têm propensão natural de armazenamento do conhecimento experimental, a fim de torna-lo disponível ao uso, foi feita por Haykin (1994). Semelhantemente ao cérebro humano em dois pontos, o primeiro se refere ao conhecimento adquirido pela rede ser proveniente do seu ambiente com base em um processo de aprendizagem, e o segundo está relacionado aos pesos sinápticos, que são forças de conexão entre os neurônios, utilizados para armazenar o conhecimento adquirido.

Uma arquitetura de rede bastante utilizada nessas aplicações, são as redes do tipo *feedfoward*, também conhecidas como modelo *perceptron* de múltiplas camadas (*Multilayer Perceptron* – MLP). Estas não realizam *loops* de rede, compostas por uma camada de entrada, uma ou mais intermediárias (ou ocultas), e uma camada de saída, onde os neurônios artificiais trabalham com funções não lineares. O treinamento ocorre em um processo interativo de ajuste dos pesos associados às ligações, com o objetivo de minimizar a diferença entre a resposta esperada e a resposta obtida pela rede, adquirindo a capacidade de generalizar os resultados para respectivo problema na fase de testes da rede (Saheki, 2005). A Figura (1) representa uma rede *feedfoward*.

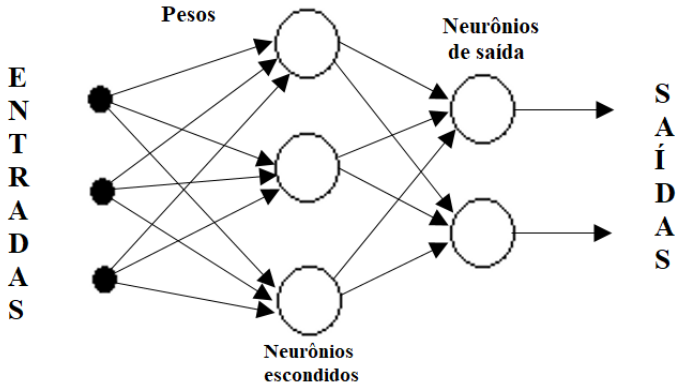

Figura 1: Rede *feedfoward (Adaptado de Azevedo, 2016)*

Na estrutura interna de cada neurônio, temos a função de ativação, que de acordo com a não-linearidade irá restringir a amplitude do intervalo de saída do neurônio, para este trabalho foi utilizada a função sigmoide ou logística, por sempre assumir valores positivos. Esta está representada na Figura (2).

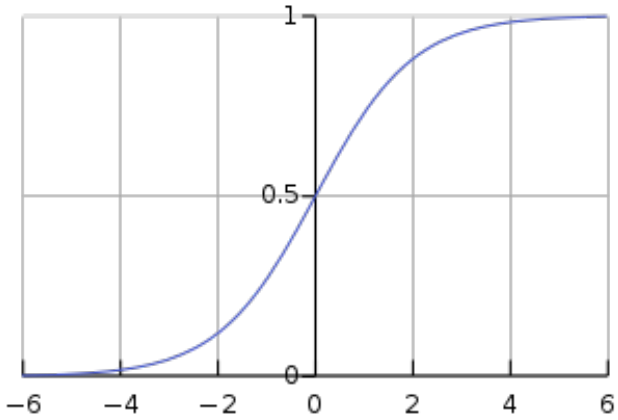

Figura 2: Função sigmoide ou logística (WIKIPÉDIA. Desenvolvido pela Wikimedia Foundation. Disponível em :< https://en.wikipedia.org/wiki/Logistic_function> Acessado em Setembro/2018)

Para que ocorra o ajuste de pesos dos neurônios durante o treinamento, de uma função não-linear é utilizada a Equação (1).

$$\phi(x) = \frac{1}{1+e^{-kx}} \tag{1}$$

3. MATERIAIS E MÉTODOS

A Rede Neural Artificial deste trabalho recebeu para seu treino e teste, o banco de dados de domínio público retirado do Wisconsin Diagnostic Breast Cancer do ano de 1995. Esta base é composta por 569 pacientes, separados em dois grupos, um de pessoas saudáveis e outro de portadores de carcinoma mamário. Dentre os pacientes do banco, 212 apresentaram diagnóstico positivo para a doença, sendo assim, os 357 restantes foram considerados saudáveis. Acompanhado do diagnóstico, o banco apresenta para cada um dos pacientes,

outras variáveis, das quais foram escolhidas seis para que completasse o sistema. As variáveis escolhidas foram raio do tumor encontrado, textura, perímetro, área, concavidade e dimensão fractal do mesmo.

A metodologia proposta é representada na Figura 3, onde o modelo computacional proposto é baseado em Redes Neurais no diagnóstico de tumores de mama.

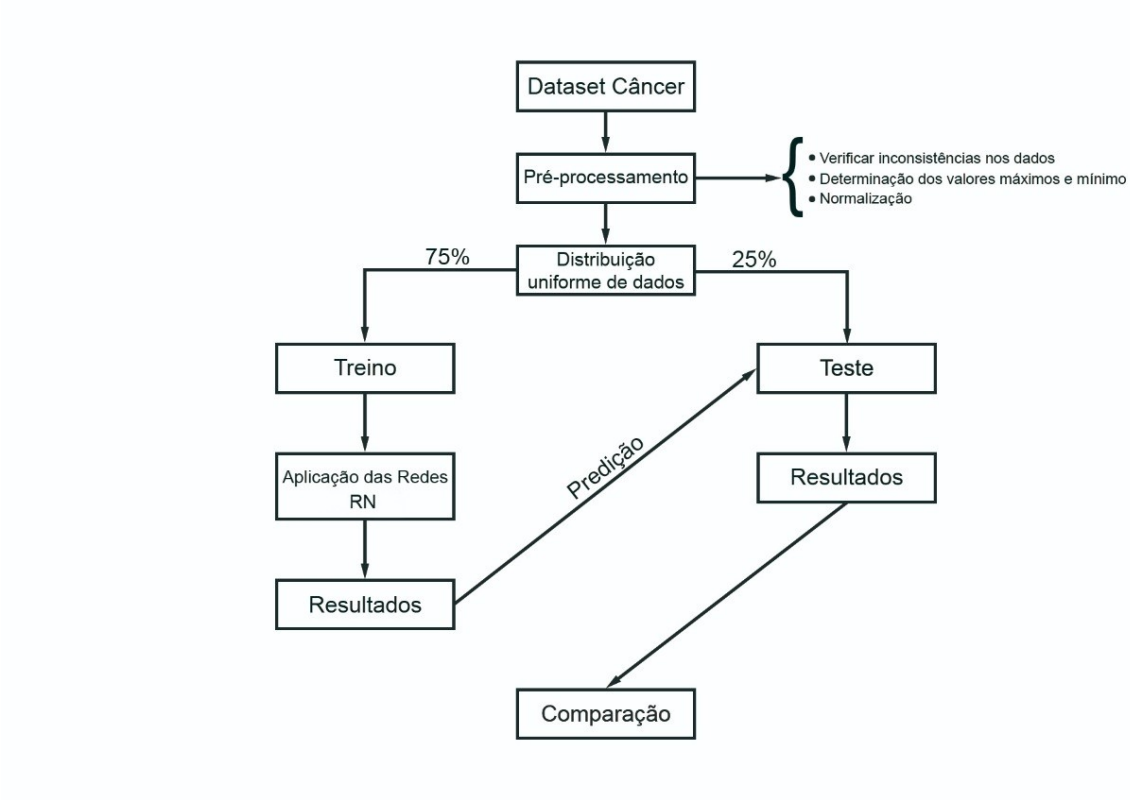

Figura 3: Fluxograma da metodologia proposta

O pré processamento dos dados consistiu na normalização dos atributos e análise exploratória. Sendo então a amostra dividida ao mero acaso em dois grupos independentes, o conjunto de treinamento é composto por 75% dos pacientes e o conjunto de teste com os 25% restantes. O modelo de rede que utilizamos foi do tipo MLP, em que a camada de entrada possui 6 neurônios, estes responsáveis por receber as variáveis do problema (“os impulsos”), foram utilizados 2 camadas ocultas no qual cada uma possui 7 e 6 neurônios, no qual elas tem o propósito de capturar/representar as diversas nuances que os dados de treinamento podem ter. Em teoria, quanto mais camadas ocultas tivermos, mais ajustada pode ficar a rede, especialmente em problemas não linearmente separáveis que é o caso do nosso trabalho, pois consegue diferenciar as pequenas nuances de um determinado dado entre as classes a que podem pertencer e por fim a camada de saída com um único neurônio, onde o resultado final é concluído e apresentado.

Utilizamos a função sigmóide como função de ativação, que descreve literalmente quando um neurônio deve ser ou não ativado para um dado valor de ativação e treinamos a rede com o algoritmo backpropagation para otimizar os erros. O modelo de RNA foi implementado no software R, com o auxílio dos pacotes neuralnet e MASS. Este que é um software livre, utilizado em aplicações de testes estatísticos, aplicações de modelos lineares e não lineares, análise de séries temporais, classificação e agrupamento de dados e técnicas gráficas expansíveis.

Um sumário estatístico foi gerado pelo próprio software para os conjuntos de treino e teste, no qual foram apresentadas informações como valor mínimo, média e valor máximo, além dos 1°, 2° e 3° quartis, do qual o segundo é chamado de mediana, para cada uma das variáveis utilizadas no modelo, onde é possível verificar nas Tabelas 1 e 2.

Tabela 1: Sumário de valores do conjunto de treino

	Raio	Área	Textura	Perímetro	Concavidade	Dim. Fractal
Valor mín.	6.98	143.5	10.38	43.79	0.0	0.049
1° Quartil	11.71	420.3	16.33	75.19	0.028	0.057
Média	13.21	538.4	18.83	85.42	0.061	0.061
Mediana	14.07	648.6	19.27	91.57	0.086	0.062
3° Quartil	15.75	777.2	21.79	103.70	0.125	0.065
Valor máx.	28.11	2499.0	33.81	188.50	0.426	0.097

Tabela 2: Sumário de valores do conjunto de teste

	Raio	Textura	Perímetro	Área	Concavidade	Dim. Fractal
Valor mín.	7.72	9.71	47.92	178.8	0.0	0.052
1° Quartil	11.68	15.74	74.91	418.0	0.033	0.057
Média	13.74	18.88	88.32	582.7	0.059	0.062
Mediana	14.30	19.35	93.18	673.7	0.094	0.063
3° Quartil	16.12	21.79	104.25	813.3	0.143	0.066
Valor máx.	27.42	39.28	186.90	2501.0	0.410	0.095

A acurácia garante o grau de confiabilidade do modelo, e é calculada pela Equação 2.

$$\text{Acurácia} = \frac{VP+VN}{N} = \frac{(Verdadeiro\ positivo + Verdadeiro\ negativo)}{Total\ lote} \quad (2)$$

O valor de sensibilidade demonstra a capacidade de o sistema reconhecer pacientes doentes. Este é calculado pela Equação 3.

$$\text{Sensibilidade} = \frac{VP}{VP+FN} = \frac{Número\ de\ resultados\ de\ testes\ verdadeiros\ positivos}{Todos\ os\ doentes\ afetados} \quad (3)$$

onde VP = Verdadeiro Positivo e FN = Falso Negativo.

4. RESULTADOS E DISCUSSÕES

Realizamos 30 simulações para o modelo proposto, ambas com um conjunto de treino formado por 75% dos dados, e um conjunto de teste com os 25% restantes. Com o *software* R foi utilizado Redes Neurais MLP para o reconhecimento dos tumores, e então, obtivemos como resposta do sistema o erro de treino, erro de teste, acurácia, sensibilidade e falso negativo.

Em todas as simulações do modelo treinamos em um grupo e validamos em outro. Comprovando assim sua funcionalidade para qualquer conjunto de valores e não apenas para

os já treinados, dessa forma os resultados de treino foram desprezados, considerando apenas os valores de teste.

Este modelo completo, em sua melhor simulação, obteve uma acurácia de 88%, garantindo que em sua simulação mais assertiva, a chance de 88% do modelo acertar no diagnóstico de um paciente. E em sua pior simulação, esse valor foi de 73.9%, o que não é um valor consideravelmente baixo.

Com valores elevados para sensibilidade, em sua melhor simulação tivemos que 96.3% dos diagnósticos positivos foram para pacientes realmente doentes.

Falso negativo é um fator de extrema importância quando se trata de resultados na área da biomedicina, principalmente em diagnósticos médicos, uma vez que este informa se o paciente recebeu diagnóstico negativo, mas o correto seria positivo. Este índice representa a porcentagem de diagnósticos assertivos, ou seja, menores valores representam menos erros de resultados de exames. Para este caso obtivemos um valor baixo, 5.2%, o que garante uma baixa probabilidade de erro.

Para escolha de melhor e pior simulação, escolhemos o valor de acurácia, por esse ser o real valor de avaliação de confiabilidade do modelo, portanto, na Tabela 3 serão apresentadas essas simulações com seus respectivos valores.

Tabela 3: Pior e melhor simulações do modelo completo

Simulações	Erro de *teste*	Acurácia	Sensibilidade	Falso Negativo
Pior	23.0%	76.7%	85.0%	15.0%
Melhor	11.9%	88.0%	96.3%	5.2%

5. CONCLUSÃO

A elevada taxa de incidência e mortes causadas pelo câncer de mama, atualmente no Brasil e no mundo, justifica o desenvolvimento de pesquisas científicas voltadas para estratégias de auxílio na detecção precoce da doença, fator determinante para o sucesso do tratamento. Dentro deste contexto, o presente artigo apresenta as Redes Neurais Artificiais MLP treinadas com o algoritmo backpropagation como uma ferramenta auxiliar na classificação de neoplasias mamárias.

Na análise dos resultados foi possível evidenciar o desempenho promissor do modelo proposto na caracterização de tumores, visto que o mesmo no conjunto das 30 simulações realizadas, obteve em seu melhor desempenho uma acurácia de 88% e sensibilidade superior a 96%. Para trabalhos futuros, pretendemos realizar algumas alterações que podem levar a um resultado final mais preciso, aumentando o conjunto de dados e a partir deles realizar o treinamento da rede garantido mais confiabilidade, ajustando o número de camadas escondidas e utilizando métodos automático para selecionar os hiperparâmetros da rede.

Sendo assim, o sistema proposto pode contribuir, como uma segunda opinião para os especialistas em uma melhor interpretação das imagens mamográficas e consequentemente, proporcionar um diagnóstico mais confiável ao paciente e evitando a realização de biópsias desnecessárias.

Agradecimentos

A CAPES pelo apoio financeiro.

REFERÊNCIAS

Azevedo, A.M.C. Redes Neurais – Rede Função de Base Radial(RBF). 2016. Disponível em : https://abil-ioazevedo.wordpress.com/2016/12/20/redes-neurais-rede-funcao-de-base-radial-rbf/- Acesso em Agosto/2018.

Câncer de Mama. Instituto Nacional de Câncer José Alencar Gomes da Silva, 2016. Disponível em: http://www.inca.gov.br/outubro-rosa/cancer-mama.asp – Acessado em Agosto/2018.

Câncer de Mama é a 2ª principal Causa de Morte Entre Mulheres nas Américas; Diagnóstico Precoce e Tratamento Podem Salvar Vidas. Organização Pan-Americana de Saude & Organização Mundial de Saúde Brasil, 2016. Disponível em: https://www.paho.org/bra/index.php?option=com_content&view=article&id=5273:cancer-de-mama-e-a-2a-principal-causa-de-morte-entre-mulheres-nas-americas;-diagnostico-precoce-e-tratamento-podem-salvar-vidas&Itemid=839 – Acessado em Agosto/2018.

Como diagnosticar e tratar câncer de mama. Grupo Editorial Moreira Jr. Disponível em: http://www.moreirajr.com.br/revistas.asp?fase=r003&id_materia=312 – Acessado em Agosto/2018.

Como realizar o diagnóstico do câncer de mama?. Hospital do Câncer de Barretos, 2012. Disponível em: https://www.hcancerbarretos.com.br/pesquisas/92-paciente/tipos-de-cancer/cancer-de-mama/163-como-realizar-o-diagnostico-do-cancer-de-mama - Acessado em Agosto/2018

Ferreira, D.L., Pires, V.A.T.N. Perfil de morbidade e mortalidade de mulheres em idade fértil na área de abrangência da microrregião de saúde de Ipatinga. Revista Enfermagem Integrada, v.6, n.1, p.1119-1132, 2013.

Glingani, F.A., Ambrósio, P.E. Sistema de análise computadorizada para auxilio à detecção de lesões de mama baseado em Redes Neurais Artificiais. Disponível em: http://telemedicina.unifesp.br/pub/sbis/CBIS2004/trabalhos/arquivos/553.pdf – Acessado em Agosto/2018.

Haddad N, Silva MB. Mortalidade por neoplasmas em mulheres em idade reprodutiva - 15 a 49 anos - no estado de São Paulo, Brasil, de 1991 a 1995. Rev Assoc Med Bras 2001.

Haykin, S. Redes Neurais, Princípios e Práticas. 2Ed. Bookman. Artmed Editora S.A. Porto Alegre, 2008.

Interpretação dos resultados dos testes. Thermo Scientific; 2012. Disponível em: http://www.phadia.com/pt-BR/Diagnostico-de-auto-imunidade/Saber-mais/Avaliacao-dos-Resultados-dos-Testes/#Sens%20Spec – Acessado em Agosto/2018.

Mavroforakis, M., Georgiou, H., Dimitropoulos, N., Cavouras,D., Theodoridis, S. *Significance analysis of qualitative mammographic features, using linear classifiers, neural networks and support vector machines. European Jounal of Radiology*. v.54, n.1, p.80-89. 2005.

Ribeiro, P.B. Classificação por análise de contornos de nódulos mamários utilizando redes neurais artificiais. Dissertação. Universidade de São Paulo. São Carlos, 2006.

Saheki, A.H. Construção de uma rede bayesiana aplicada ao diagnóstico de doenças. Dissertação. Universidade de São Paulo. São Paulo, 2005.

Santos, A.K.F., Camara, K.M., Bini, I.C. Cãncer de mama. Disponível em: http://www.vitrineacademica.dombosco.sebsa.com.br/index.php/vitrine/article/viewFile/115/116 - Acessado em Agosto/2018.

The R *Project for Statistical Computing*. *The* R *Foundation*; 2017. Disponível em: https://www.r-project.org/ - Acessado em Agosto/2018.