Universidade Federal de Uberlândia
Faculdade de Engenharia Elétrica
Graduação em Engenharia Biomédica

DOUGLAS MARSICANO DUNGA

SISTEMAS INTELIGENTES PARA DIAGNÓSTICOS AUTOMATIZADOS DO CÂNCER DE MAMA

Uberlândia
2019

DOUGLAS MARSICANO DUNGA

SISTEMAS INTELIGENTES PARA DIAGNÓSTICOS AUTOMATIZADOS DO CÂNCER DE MAMA

Trabalho apresentado como requisito parcial de avaliação na disciplina Trabalho de Conclusão de Curso de Engenharia Biomédica da Universidade Federal de Uberlândia.

Orientador: Fernando Pasquini Santos

_____

Assinatura do Orientador

Uberlândia
2019

AGRADECIMENTOS

Agradeço aos meus pais Edson e Mariângela, que sempre acreditaram em mim e nos meus sonhos, me apoiando em todas as etapas do caminho, se entristecendo com meus tropeços e vibrando com minhas conquistas.

A minha avó Aparecida, minha irmã Nádia e toda família, pelo apoio, amizade e companheirismo.

A todos os amigos de faculdade, por terem compartilhado os anseios e alegrias, me ajudando a tornar possível essa trajetória e não apenas isso, por terem feito valido apena até os piores momentos. Sobretudo ao Eduardo, Gabriella, Otávio, Iago, Nyalla, Paulo, Hugo, Bisson, Fernanda e Lucas.

Aos professores pela dedicação, sugestão e instruções que não apenas me capacitaram profissionalmente como também me fizeram progredir pessoalmente. Em especial ao professor Cadu, mais chamado por mim apenas de Tio Du, que não apenas tio e professor foi meu principal orientador.

À Universidade Federal de Uberlândia (UFU) e à Faculdade de Engenharia Elétrica (FEELT) por todo o conhecimento atribuído e pela formação de qualidade.

Enfim, a todos que contribuíram direta ou indiretamente para a realização deste trabalho e em minha jornada até aqui, o meu muito obrigado.

RESUMO

Neoplasias de mama são casos frequentes dentre as diversas anomalias relacionadas ao crescimento desordenado das células humanas. O câncer de mama é um dos tipos mais frequentes em mulheres e apresenta alto índice de mortalidade. Por este motivo o diagnóstico precoce torna-se fundamental para o combate à doença.

Utilizando a base de dados WDBC e a plataforma *WEKA* de Aprendizado de Máquina, um sistema para diagnóstico automatizado do câncer de mama foi proposto. Cinco diferentes algoritmos classificadores foram configurados e aplicados na obtenção um sistema confiável no apoio ao diagnóstico de câncer de mama.

A base de dados foi devidamente tratada e filtrada antes da aplicação dos algoritmos e o resultado obtido pôde ser analisado pela acurácia média do sistema: 15% superior à metodologia de diagnóstico utilizada atualmente (BI-RADS).

A técnica de AM mostrou-se promissora e com potencial de melhorar drasticamente o nível de assertividade no diagnóstico do câncer de mama. Observou-se a necessidade de reproduzir o experimento e analisar seus resultados comparando os níveis de desempenho aplicando novas bases de dado afim de encontrar o algoritmo e a configuração mais adequada garantindo que o sistema se torne cada vez mais robusto e confiável.

ABSTRACT

Breast neoplasms are frequent cases among the various abnormalities related to the disordered growth of human cells. Breast cancer is one of the most common types in women and has a high mortality rate. For this reason, early diagnosis becomes essential to combat the disease.

Using the WDBC database and WEKA Machine Learning platform, a system for automated breast cancer diagnosis has been proposed. Five different classification algorithms have been configured and applied to obtain a reliable system to support breast cancer diagnosis.

The database was properly treated and filtered before the algorithms were applied and the result obtained could be analyzed by the average system accuracy: 15% higher than the current diagnostic methodology (BI-RADS).

The breastfeeding technique has shown promise and has the potential to dramatically improve the level of assertiveness in the diagnosis of breast cancer. It was necessary to replicate the experiment and analyze its results by comparing performance levels by applying new databases in order to find the most appropriate algorithm and configuration, ensuring that the system became increasingly robust and reliable

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

AD - Árvores de Decisão

AM - Aprendizado de Máquina

AMS - Aprendizado de Máquina Supervisionado

CBI - Classificação Baseada em Instâncias

GIGO- *"Garbage in, garbage out"*

IA - Inteligência Artificial

INCA - Instituto Nacional de Câncer

NB - Naive Bayes

NCI - Instituto Americano do Câncer

PAAF - Punção aspirativa por agulha fina

PD - Processamento de Dados

RBC - Raciocínio Baseado em Casos

RBF - Função de Base Radial

RNA – Rede Neural Artificial

RNP - Rede Neural Probabilística

SUS - Sistema Único de Saúde

SVM – "Support Vector Machine" Máquina de Vetor Suporte

WDBC - Wisconsin Diagnostic Breast Cancer

SUMÁRIO

1. INTRODUÇÃO

Neoplasias de mama são casos frequentes dentre as diversas anomalias relacionadas ao crescimento desordenado das células humanas. Segundo o Instituto Americano do Câncer (NCI), o número de novos casos de câncer na região mamária nos Estados Unidos foi superior a 268 mil no ano de 2018, sendo que destes, o número de incidentes que culminaram em mortes excedeu os 41.400 (NCI, 2019). Já no Brasil os números para este mesmo ano são igualmente surpreendentes. Segundo pesquisa relatada pelo Instituto Nacional de Câncer (INCA), o câncer de mama é o tipo que mais acomete mulheres no país e perde somente para o câncer de pele quando envolve os dois gêneros. Ainda segundo o mesmo instituto, o câncer de mama liderou as estatísticas de novos casos da doença no ano de 2018. São, ao todo, 59.700 (29,5%) casos registrados neste período, dos quais 15.403 (16,2 %) levaram ao óbito (INCA, 2019).

O câncer de mama é o crescimento descontrolado de certas células da mama que adquiriram características anormais (células dos lobos, células produtoras de leite, ou dos ductos, por onde é drenado o leite), anormalidades estas causadas por uma ou mais mutações no material genético da célula. A Figura 1 destaca a anatomia da região mamária bem como os diferentes conjuntos celulares que a compõem. A doença ocorre majoritariamente com mulheres, porém, indivíduos do sexo masculino também estão sujeitos ao desenvolvimento de carcinomas do gênero (ACS, 2018).

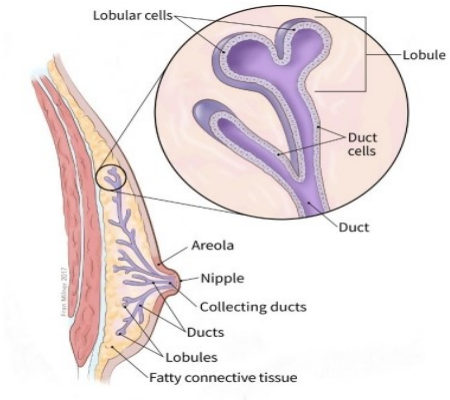

Figura 1 - Anatomia da mama (ACS, 2018)

Existem vários tipos de câncer de mama, a maioria manifesta-se como nódulo, porém nem todos o fazem. Vale destacar, que a maior incidência de nódulos na mama é de "lesões benignas". Os denominados tumores benignos de mama são crescimentos anormais organizados e bem delimitados e, portanto, não se disseminam. Todavia, alguns nódulos

benignos podem aumentar o risco de se contrair câncer de mama (ACS, 2017); (INSTITUTO ONCOGUIA, 2017).

Quanto mais cedo detectada a doença mais eficaz poderá ser o de tratamento fornecido ao paciente. Consequentemente, as chances de cura são aumentadas drasticamente. Assim a detecção precoce do distúrbio é estratégia fundamental na diminuição das taxas de mortalidade em decorrência desse tipo de câncer (GONÇALVES, 2017). Além de possibilitar terapias mais simples e efetivas e contribuir para a redução do estágio de apresentação do câncer. A maioria dos tipos de câncer de mama é passível de diagnóstico precoce mediante avaliação e encaminhamento oportunos após os primeiros sinais e sintomas (MS, 2015), (INCA, 2013).

O principal e mais comum exame para rastreamento do câncer de mama é a através da mamografia. No Brasil, o Sistema Único de Saúde (SUS) oferece o serviço a mulheres acima de 40 anos. O procedimento é realizado através de um equipamento que emite raios X para estabelecer um mamograma do paciente, ou seja, determinar a morfologia, anatomia e patologias gerais na região mamária (COSTA, 2002), (INCA, 2019).

Diagnóstico, em medicina, é o processo analítico de que se vale o especialista ao exame de uma doença ou de um quadro clínico, para chegar a uma conclusão. É também o nome dado à conclusão em si mesma (SIMPSON, 1994), ou seja, o diagnóstico visa identificar pessoas com sinais e sintomas de uma determinada doença. Em virtude da natureza da doença em questão, quanto antes o diagnóstico nestes casos, maior a qualidade de tratamento aplicado a um paciente (WANG, ZHENG, *et al.*, 2017).

Um dos principais desafios contra a detecção é como classificar os tumores em malignos (cancerígenos) ou benignos (não cancerígenos). Um tumor é considerado maligno se as células crescerem nos tecidos circundantes ou se espalharem para áreas distantes do corpo. Um tumor benigno não invade tecidos próximos nem se espalha para outras partes do corpo da mesma maneira que os tumores cancerígenos. Apesar disso, tumores benignos requerem atenção dos oncologistas devido ao potencial de pressionar estruturas vitais, como vasos sanguíneos ou nervos (LAMIDI, 2018).

Uma vez que, até o momento, o câncer não pode ser evitado, a melhor forma da diminuição das taxas de morbidade e de mortalidade é pela detecção precoce (SIMPSON, 1994). A palpação das mamas e a mamografia são os procedimentos mais utilizados e mais eficientes para se determinar alguma anomalia (AL., 2003) (MS, 2015). Quando isso acontece e alguma uma anormalidade suspeita é detectada pela mamografia de diagnóstico, exames adicionais como ultrassom e ou biópsias podem ser requisitados (SBIB - ALBERT EINSTEIN).

Este último método, a biópsia cirúrgica, é capaz de confirmar a malignidade da doença com alto grau de sensibilidade, fato este, que o torna confiável o suficiente tanto para diagnosticar, ou seja, determinar se o tumor é benigno ou maligno, quanto para prognosticar, ou seja antever a progressão/regressão da neoplasia com base nos indícios apontados (MAGLOGIANNIS, ZAFIROPOULOS, *et al.*, 2007) (PROGNÓSTICO, 2003-2019).

Por outro lado, a técnica não é vista com bons olhos se consideramos que ela possui alto custo operacional e, por ser invasiva, é mais desconfortável, requer longos períodos para recuperação e aumenta os riscos de infecção, podendo em alguns casos, provocar deformações indesejadas da mama, impactando negativamente o psicológico dos pacientes (MAGLOGIANNIS, ZAFIROPOULOS, *et al.*, 2007) (MALKIN, 2019) (SUSAN G. KOMEN, 2019).

No intuito de reduzir os aspectos negativos de um procedimento cirúrgico sem abdicar de alta precisão de diagnóstico, a técnica de Punção aspirativa por agulha fina (PAAF) ganha destaque neste cenário (LITHERLAND, EVANS, *et al.*, 2005). O método utiliza agulhas muito finas (calibres entre 0,455 a 0,573 mm) e uma seringa para retirar por aspiração uma pequena quantidade do tumor. Em caso de o tumor se encontrar em camadas mais profundas, a biopsia por agulha pode ser guiada por um exame de imagem, por exemplo, ultrassom ou tomografia computadorizada (MYVMC, 2005). A Figura 2 ilustra como a técnica é aplicada.

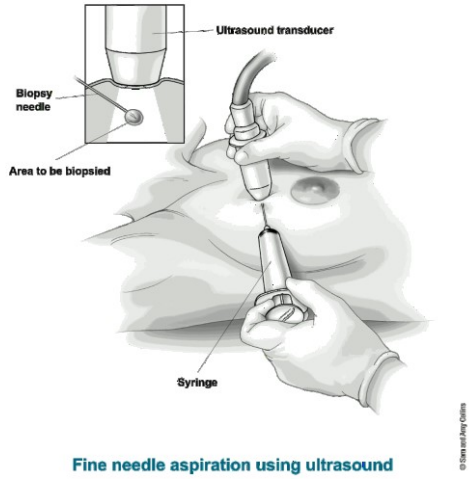

Figura 2 - Exemplo de PAAF (ACS, 2017)

Medidas como Dimensão Fractal e Técnicas Texturais vêm sendo aplicadas na extração de características das células do corpo humano para alimentar algoritmos classificadores de Aprendizado de Máquina (AM) tais como o SVM, do inglês, *Support Vector Machine*, ou Máquina de Vetor Suporte; e o RNA termo empregado para Redes Neurais Artificiais (MAGLOGIANNIS, ZAFIROPOULOS, *et al.*, 2007).

Este estudo teve por objetivo comparar a acurácia de diagnóstico de sistemas inteligentes de classificação por Aprendizado de Máquina (AM) para nódulos mamários extraídos por biópsias por punção aspirativa por agulha fina (PAAF).

No tópico que segue será abordada uma revisão bibliográfica para embasamento teórico dos pontos chaves deste estudo, como conceitos de extração de características, aprendizado supervisionado e aprendizado de máquina, definição das principais técnicas de AM e seu uso na classificação do diagnóstico do câncer de mama.

2. REVISÃO BIBLIOGRÁFICA

Nos últimos anos, houve um grande aumento no interesse pelas técnicas de Processamento de Dados e pelo Raciocínio Baseado em Casos (RBC) (PERNER, HOLT e RICHTER, 2005). Paralelamente, a especialidade da Informática Clínica cresceu aceleradamente e ganha cada vez mais destaque no ramo da Informática Médica (SABBATINI, 2019). Com isso, pesquisas vêm sendo incentivadas para acelerar o desenvolvimento de ferramentas de diagnóstico no intuito de apoiar a decisão dos profissionais médicos. Algoritmos de Aprendizado de Máquina são empregados para fornecer segundas opiniões no diagnóstico e prognóstico de dados médicos com a integração de estratégias de RBC.

Especialmente na área de imagens médicas bidimensionais, técnicas de registro, segmentação, análise e classificação de padrões de imagens médicas compõe a vanguarda de interesse (VIERGEVER, 1998). No caso da segmentação por exemplo, abordagens baseados em aprendizado são usadas para resolver problemas como detecção de bordas (TSOTSOS, 1985) e crescimento de regiões e, desse modo, detectar os limites de objetos relevantes (MARTELLI, 1988).

Além disso, modelos deformáveis (TAVARES, 2003) podem ser empregados para rastrear a deformação de órgãos como o coração e modelar vários parâmetros funcionais de órgãos, enquanto fatias sucessivas de varredura podem ser usadas para reconstruir o volume 3D de um órgão interno (ARAUJO, 2010). Na maior parte das ferramentas de análise de imagens, o processo de segmentação de imagens é seguido por um módulo de extração de atributos, capaz de realizar medições em um pixel, ou em uma matriz específica de pixels de modo a representar numericamente um objeto relevante segmentado, permitindo inclusive extração de características não visíveis sem o auxílio dessa tecnologia. Propriedades relacionadas em cores e bordas também podem ser calculadas e usadas para classificação (SILVA, 2007).

Vários estudos comprovaram a eficácia dos algoritmos descritores de bordas (semelhantes ao algoritmo empregado pela base de dados que alimentou este estudo) para a detecção de volumes malignos em métodos de avaliação computacionais de várias categorias de diagnóstico por imagem, como mamografia (BOCCHI e NORI, 2006), ultrassom (AMORES e RADEVA, 2005), raios-x (SAMEER, LEE, *et al.*, 2004), ressonância magnética (MARTENS, 2002) e câncer de pele (MAGLOGIANNIS, PAVLOPOULOS e KOUTSOURIS, 2005). As informações sobre textura de imagem também se mostram interessante, especialmente para os casos em que a densidade do tecido é mapeada para a intensidade de pixel, ou seja, nos casos

de imagens obtidas através de ultrassom ou então imagens de câncer por microscopia (MORRIS, 1988) .

O último estágio de uma ferramenta de diagnóstico computacional é a classificação, que resulta na tomada de decisão (NÓBRAGA, 1992). Utilizado como ferramenta de análise nesta última etapa, o aprendizado de máquina é um ramo da pesquisa em inteligência artificial que emprega uma variedade de ferramentas estatísticas, probabilísticas e de otimização para "aprender" com exemplos anteriores e depois usar esse treinamento prévio para classificar novos dados, identificar novos padrões ou prever novas tendências (MITCHELL, 2000).

Como o intuito do estudo é classificar cada caso no espaço de descrição, classificação parametrizada, o tipo de AM adequado neste caso é o Aprendizado de Máquina Supervisionado (AMS), uma vez que um algoritmo de AMS toma "decisões baseadas na experiência contida em exemplos solucionados com sucesso"[1] proporcionado assim, pesos e parâmetros capazes de calibrar os níveis de assertividade conforme o modelo (BARANAUSKAS, 2002).

A técnica de AMS envolve principalmente duas fases: Treinamento e Teste. Durante a fase de treinamento, são extraídos valores típicos de uma sequência de dados cuja classificação é conhecida para que sejam calculadas certas regras de classificação que satisfaça a regra de agrupamento utilizada. Existe uma ampla seleção de técnicas de AMS classificativas na literatura, variando desde Modelos Discriminativos Estatísticos até ferramentas mais sofisticadas de IA, como Redes Neurais Artificiais (RNA). Já durante a fase de testes, outro conjunto de dados é usado para testar o processo de classificação e avaliar o desempenho de um classificador afim de se obter um cálculo da assertividade do algoritmo para resolução de um problema. As técnicas de classificação e os métodos de avaliação utilizados no contexto deste estudo são analisadas em maior detalhe nos tópicos a seguir.

[1] (WEISS, 1991)

3. METODOLOGIA

3.1 Base de Dados

O presente estudo utilizou a Base de Dados *Wisconsin Diagnostic Breast Cancer (WDBC)* (WOLBERG, STREET, *et al.*, 1995), obtida através de trabalhos realizados no Hospital da Universidade de Wisconsin para o diagnóstico tumores de mama. Esta base de dados apresenta atributos calculados a partir da digitalização de imagens por PAAF. A Figura 3 exemplifica um caso de estudo que deu origem ao banco.

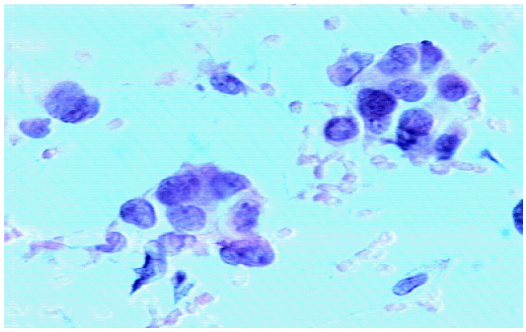

Figura 3 – Imagem extraída da técnica de PAAF utilizada para formar a base WDBC (WOLBERG, STREET, et al.*, 1995)*

O conjunto de dados WDBC é composto por 569 casos de estudo. Cada caso representa as medições obtidas por PAAF para um tipo de diagnóstico. Da totalidade de sujeitos, 357 foram classificados (previamente) como tumores benignos e 212 como malignos. A Figura 4 retrata como os dados são disponibilizados.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	842302,00000	M	17,99000	10,38000	122,80000	1001,00000	0,11840	0,27760	0,30010	0,14710	0,24190	0,07871	1095,00000	0,90530	8589,00000	153,40000	0,00640	0,04904	0,05373	0,01587	0,03003
2	842517,00000	M	20,57000	17,77000	132,90000	1326,00000	0,08474	0,07864	0,08690	0,07017	0,18120	0,05667	0,54350	0,73390	3398,00000	74,08000	0,00523	0,01308	0,01860	0,01340	0,01389
3	84300903,00000	M	19,69000	21,25000	130,00000	1203,00000	0,10960	0,15990	0,19740	0,12790	0,20690	0,05999	0,74560	0,78690	4585,00000	94,03000	0,00615	0,04006	0,03832	0,02058	0,02250
4	84348301,00000	M	11,42000	20,38000	77,58000	386,10000	0,14250	0,28390	0,24140	0,10520	0,25970	0,09744	0,49560	1156,00000	3445,00000	27,23000	0,00911	0,07458	0,05661	0,01867	0,05963
5	84358402,00000	M	20,29000	14,34000	135,10000	1297,00000	0,10030	0,13280	0,19800	0,10430	0,18090	0,05883	0,75720	0,78130	5438,00000	94,44000	0,01149	0,02461	0,05688	0,01885	0,01756
6	843786,00000	M	12,45000	15,70000	82,57000	477,10000	0,12780	0,17000	0,15780	0,08089	0,20870	0,07613	0,33450	0,89020	2217,00000	27,19000	0,00751	0,03345	0,03672	0,01137	0,02165
7	844359,00000	M	18,25000	19,98000	119,60000	1040,00000	0,09463	0,10900	0,11270	0,07400	0,17940	0,05742	0,44670	0,77320	3,18000	53,91000	0,00431	0,01382	0,02254	0,01039	0,01369
8	84458202,00000	M	13,71000	20,83000	90,20000	577,90000	0,11890	0,16450	0,09366	0,05985	0,21960	0,07451	0,58350	1377,00000	3856,00000	50,96000	0,00881	0,03029	0,02488	0,01448	0,01486
9	844981,00000	M	13,00000	21,82000	87,50000	519,80000	0,12730	0,19320	0,18590	0,09353	0,23500	0,07389	0,30630	1002,00000	2406,00000	24,32000	0,00573	0,03502	0,03553	0,01226	0,02143
10	84501001,00000	M	12,46000	24,04000	83,97000	475,90000	0,11860	0,23960	0,22730	0,08543	0,20300	0,08243	0,29760	1599,00000	2039,00000	23,94000	0,00715	0,07217	0,07743	0,01432	0,01789
11	845636,00000	M	16,02000	23,24000	102,70000	797,80000	0,08206	0,06669	0,03299	0,03323	0,15280	0,05697	0,37950	1187,00000	2466,00000	40,51000	0,00403	0,00927	0,01101	0,00759	0,01460
12	84610002,00000	M	15,78000	17,89000	103,60000	781,00000	0,09710	0,12920	0,09954	0,06606	0,18420	0,06082	0,50580	0,98490	3564,00000	54,16000	0,00577	0,04061	0,02791	0,01282	0,02008
13	846226,00000	M	19,17000	24,80000	132,40000	1123,00000	0,09740	0,24580	0,20650	0,11180	0,23970	0,07800	0,95550	3568,00000	11,07000	116,20000	0,00314	0,08297	0,08890	0,04090	0,04484
14	846381,00000	M	15,85000	23,95000	103,70000	782,70000	0,08401	0,10020	0,09938	0,05364	0,18470	0,05338	0,40330	1078,00000	2903,00000	36,58000	0,00977	0,03126	0,05051	0,01992	0,02981
15	84667401,00000	M	13,73000	22,61000	93,60000	578,30000	0,11310	0,22930	0,21280	0,08025	0,20690	0,07682	0,21210	1169,00000	2061,00000	19,21000	0,00643	0,05936	0,05501	0,01628	0,01961
16	84799002,00000	M	14,54000	27,54000	96,73000	658,80000	0,11390	0,15950	0,16390	0,07364	0,23030	0,07077	0,37000	1033,00000	2879,00000	32,55000	0,00561	0,04240	0,04741	0,01090	0,01857
17	848406,00000	M	14,68000	20,13000	94,74000	684,50000	0,09867	0,07200	0,07395	0,05259	0,15860	0,05922	0,47270	1,24000	3195,00000	45,40000	0,00572	0,01162	0,01998	0,01109	0,01410
18	84862001,00000	M	16,13000	20,68000	108,10000	798,80000	0,11700	0,20220	0,17220	0,10280	0,21640	0,07356	0,56920	1073,00000	3854,00000	54,18000	0,00703	0,02501	0,03188	0,01297	0,01689
19	849014,00000	M	19,81000	22,15000	130,00000	1260,00000	0,09831	0,10270	0,14790	0,09498	0,15820	0,05395	0,75820	1017,00000	5865,00000	112,40000	0,00649	0,01893	0,03391	0,01521	0,01356
20	8510426,00000	B	13,54000	14,36000	87,46000	566,30000	0,09779	0,08129	0,06664	0,04781	0,18850	0,05766	0,26990	0,78860	2058,00000	23,56000	0,00846	0,01460	0,02387	0,01315	0,01980
21	8510653,00000	B	13,08000	15,71000	85,63000	520,00000	0,10750	0,12700	0,04568	0,03110	0,19670	0,06811	0,18520	0,74770	1383,00000	14,67000	0,00410	0,01898	0,01698	0,00649	0,01678
22	8510824,00000	B	9504,00000	12,44000	60,34000	273,90000	0,10240	0,06492	0,02956	0,02076	0,18150	0,06905	0,27730	0,97680	1909,00000	15,70000	0,00961	0,01432	0,01985	0,01421	0,02027
23	8511133,00000	M	15,34000	14,26000	102,50000	704,40000	0,10730	0,21350	0,20770	0,09756	0,25210	0,07032	0,43880	0,70960	3384,00000	44,91000	0,00679	0,05328	0,06446	0,02252	0,03672
24	851509,00000	M	21,16000	23,04000	137,20000	1404,00000	0,09428	0,10220	0,10970	0,08632	0,17690	0,05278	0,69170	1127,00000	4303,00000	93,99000	0,00473	0,01259	0,01715	0,01038	0,01083
25	852552,00000	M	16,65000	21,38000	110,00000	904,60000	0,11210	0,14570	0,15250	0,09170	0,19950	0,06330	0,80680	0,90170	5455,00000	102,60000	0,00605	0,01882	0,02741	0,01130	0,01468

Figura 4 – WDBC (Fonte própria)

Para esse conjunto de dados, cada instância (linha) possui 32 atributos (colunas), dos quais os dois primeiros atributos correspondem ao número de identificação exclusivo e ao status do diagnóstico (B-benigno / M-maligno). As outras 30 características são o resultado da

computação e extração de dez características com valor real (apresentadas na Tabela 1), juntamente com suas médias, erro padrão e a média dos três maiores valores para os núcleos celulares estudados, respectivamente.

Tabela 1 - WDBC Características extraídas dos núcleos celulares por PAAF (WOLBERG, STREET, et al., 1995)

1	*Raio (distância média entre o centro aos pontos do perímetro)*
2	*Textura (derivação padrão dos valores da escala de cinza)*
3	*Perímetro*
4	*Área*
5	*Suavidade (variação local no comprimento do raio)*
6	*Compacidade* $[\frac{perímetro^2}{área} - 1]$
7	*Concavidade (severidade das porções côncavas do contorno)*
8	*Pontos côncavos (número de pontos côncavos do contorno)*
9	*Simetria*
10	*Dimensão Fractal (diz o quão densamente um conjunto ocupa um espaço[2])*

A base WDBC mostra-se de boa qualidade principalmente pelo fato de ser numerosa e bem caracterizada. Apresenta 30 atributos e 569 casos de estudo com poucas informações faltantes é, portanto, completa e adequada para aplicações do tipo do problema em questão.

3.2 Técnicas de classificação automatizada

Existem vários tipos de abordagens dentre os algoritmos de AMS e, consequentemente, a classificação diferencia-se conforme as complexidades e metodologias aplicadas. A abordagem mais apropriada para a resolução de um caso está relacionada com a natureza do problema em si e em com como cada código utiliza o conjunto de dados fornecido. Da mesma forma, alguns métodos podem ter suposições ou requisitos de dados que os tornam inaplicáveis para resolução do problema em questão.

[2] (Cálculo da Dimensão Fractal - Método de BoxCounting)

Apesar da escolha do algoritmo mais adequado basear-se em conceitos bem definidos e aceitos, como acurácia, sensibilidade, interpretabilidade, demanda computacional e especificidade (ALMALIKI, 2019), determinar qual o melhor método para solucionar um certo problema ainda não é uma tarefa inerentemente bem estabelecida. Por este motivo, torna-se importante aplicar e comparar mais de um método de AM em qualquer conjunto de treinamento estudado (CRUZ, 2006).

Neste contexto, cinco algoritmos de AMS se destacam e são os alvos do estudo em questão. Os subtópicos a seguir discorrem resumidamente a respeito da base lógica e dos conceitos que orientam cada um dos algoritmos classificadores empregados.

3.2.1 Naive Bayes (NB)

O algoritmo Naive Bayes é um classificador probabilístico baseado no "Teorema de Bayes", criado por Thomas Bayes (1701 - 1761).

A principal característica do algoritmo, e o motivo de receber "naive" (ingênuo) no nome é que o modelo supõe que os atributos são independentes, dada uma classe específica, o que, na prática, é uma estimativa simplista.

Assim, desconhecendo a correlação entre as variáveis de estudo, o cálculo pode ser simplificado de modo que cada atributo seja independente e então a probabilidade pode ser obtida pelo produto das probabilidades condicionais individuais de cada atributo.

Apesar da ingenuidade contida no nome, o modelo funciona muito bem quando testado em conjuntos de dados reais (MAGLOGIANNIS, ZAFIROPOULOS, *et al.*, 2007), particularmente quando combinado com procedimentos de seleção de atributos que eliminam atributos redundantes e, portanto, não independentes (serão citados no subtópico a seguir).

As redes bayesianas são um conjunto de métodos para cálculos probabilísticos na maioria dos problemas caracterizados por incerteza. Funções de densidade de probabilidade dos atributos contidos na base de dados são aproximadas como uma distribuição normal durante a fase de treinamento. Em seguida, são calculados os valores médios e de variação correspondentes para cada distribuição. Uma Rede Bayesiana simples para diagnóstico médico está representada na Figura 5, onde é demonstrada a dependência entre a doença e os *n* sintomas.

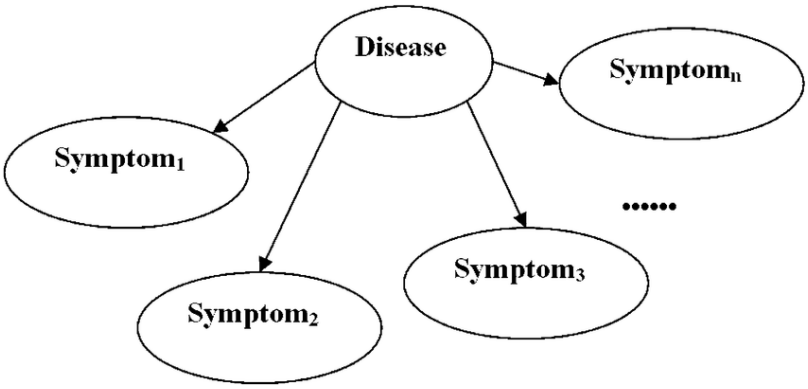

Figura 5 - Rede Bayesiana Simples para Diagnóstico Médico (MAGLOGIANNIS, ZAFIROPOULOS, et al.*, 2007)*

3.2.2 Máquinas de Vetor Suporte (SVM)

O conceito de Máquina de Vetores de Suporte, do inglês Support Vector Machine (SVM) foi introduzido inicialmente por Vapnik e Chervonenkis em 1965. Consiste em uma técnica de aprendizado supervisionado capaz de classificar os dados a partir da criação de hiperplanos de separação que diferenciem as classes.

Resumidamente, o objetivo deste método é encontrar o melhor hiperplano capaz de maximizar a margem de separação entre as classes, ou seja, gerar o modelo capaz de separar classes com o menor erro (número de elementos classificados erroneamente) possível.

Para encontrar o hiperplano ideal, o qual maximiza a margem entre as classes são utilizados Vetores de Suporte, que são as amostras do conjunto de treinamento de ambas as classes que estão mais próximas (BURBIDGE e TROTTER, 2001).

Assim, se os dados forem linearmente separáveis é possível selecionar dois hiperplanos paralelos, fazendo com que a distância entre as duas classes seja máxima (BURBIDGE, TROTTER, et al., 2001). A região entre os dois hiperplanos que separam as duas classes é chamada de margem. A equação desses hiperplanos pode ser definida como:

Em alguns casos, quando existem algumas amostras das classes que estão muito próximas, pode não ser possível uma separação sem nenhum erro. Assim, torna-se necessário acrescentar parâmetros que permitam com que o modelo se adeque aos dados, tornando alguns erros toleráveis ao modelo de SVM.

Assim, adiciona-se parâmetros ao modelo original capazes de definir a máxima tolerância para o erro. Encontrar os valores corretos para cada parâmetro não é uma tarefa e há uma vasta literatura sobre como escolher os melhores valores (HSU, CHANG e LIN, 2003). Geralmente, a técnica mais usada é encontrá-las por tentativa e erro.

O algoritmo de SVM ganha popularidade por ser capaz de lidar bem também com problemas um pouco mais complexos como separações não lineares. Para tal, a ferramenta abre mão da utilização da chamada *Função Kernel*, a qual permite aumentar a dimensionalidade da solução. Essa elevação na dimensão dos dados tem a intensão de tornar os problemas que antes eram impossíveis de serem separados linearmente, em problemas com uma possível solução linear a partir de hiperplanos gerados nessa nova dimensão. Isso nos permite classificar melhor os exemplos, gerando modelos que se adequam melhor ao conjunto de dados. Essa característica da técnica de SVM aumenta sua usabilidade e, consequentemente, aumenta sua complexidade, já que agora novos parâmetros como o tipo de função Kernel a ser utilizada são adicionados (DUDA, 2002).

3.2.3 Classificação Baseada em Instâncias (CBI)

Este método classificatório assume que as instâncias podem ser representadas como pontos em um espaço Euclidiano. Este tipo de classificação, portanto, assume o caráter não paramétrico, ou seja, considera a utilização com distribuições arbitrárias e sem suposições e conhecimentos prévios sobre densidades de distribuição (KOERICH, 2006).

Dessa forma, a aprendizagem consiste em armazenar o conjunto de treinamento na memória para serem utilizados na classificação do conjunto de testes.

Assim, treinado um conjunto de dados, o algoritmo deverá informar o quão similar o novo ponto testado se aproxima dos pontos memorizados. Por fim, detalha-se um conjunto de instâncias, ou medidas de similaridade, entre os módulos formados entre os vetores. O ponto testado é classificado da mesma forma que a instância mais similar a ele (KOERICH, 2006).

Existem vários algoritmos que trabalham com esta lógica. Foi utilizado o mais comum para aplicações científicas, o *kNN - lazy IBk,* ou do inglês "*k* Nearest Neighbor" tradução literal de "*k* Vizinhos mais Próximos" que classifica um ponto atribuindo a ele o rótulo representado mais frequentemente dentre as *k* amostras mais próximas e utilizando um esquema de votação (BROWNLEE, 2019).

3.2.4 Aprendizagem por Árvores de Decisão (AD)

Como o próprio nome intitula, a técnica consiste na construção gerar uma árvore de decisão baseada em um conjunto de dados de treinamento, sendo este modelo usado para classificar as instâncias no conjunto teste.

O algoritmo J48, proposto por (QUINLAN, 1993), é considerado o que apresenta o melhor resultado na montagem de árvores de decisão, a partir de um conjunto de dados de treinamento. E por este motivo é bastante conhecido no meio científico. Para a montagem da árvore, o algoritmo utiliza a abordagem de dividir-para-conquistar, fragmentando o problema em subproblemas mais simples. Funciona aplicando recursivamente a mesma estratégia a cada subproblema, dividindo o espaço definido pelos atributos em subespaços, associando-se a eles uma classe, por mais complexo que o problema inicial seja (LIBRELOTTO, 2013).

3.2.5 Redes Neurais Artificiais (RNA)

Redes neurais artificiais (RNAs) são originárias da engenharia e duas de suas principais áreas de aplicação são problemas de classificação e decisão. Uma RNA é um padrão de processamento de informações inspirado na maneira como os sistemas nervosos biológicos, como o cérebro, processam informações. O elemento chave do conceito é a reestruturação dos sistemas de processamento de informações, que são compostos por vários elementos de processamento interconectados (neurônios) trabalhando para resolver problemas específicos.

Assim como o aprendizado nos sistemas biológicos, o aprendizado de RNAs também envolve ajustes nas conexões sinápticas que existem entre os neurônios o que fazem deste algoritmo uma boa escolha por razões computacionais, pois, uma vez treinado, opera rapidamente (WITTEN e FRANK, 2005).

Para o problema de diagnóstico, empregou-se uma Rede Neural Probabilística (RNP), uma vez que esse tipo de rede apresenta alta capacidade de generalização e não requer grande quantidade de dados para treinamento. Assim, o classificador neural decidirá se a instância de entrada corresponde a um caso benigno ou maligno.

A topologia para este caso consiste em um sistema de três camadas: 31-568-2. A camada de entrada consiste em 31 nós, que correspondem ao status do diagnóstico, seguidos pelos 30 valores calculados (média, erro padrão e valor “pior”) da imagem digitalizada de cada instância.

A segunda camada é a camada padrão, que organiza o treinamento configurado de tal maneira que um elemento de processamento individual represente cada vetor de entrada normalizado, consistindo em 568 nós. Essa quantidade de nós corresponde à quantidade total de casos de estudo do conjunto de dados, exceto a que resta para testar cada época de treinamento, de acordo com o método de teste do canivete. Finalmente, a rede possui uma camada de saída composta por 2 nós, representando a decisão sobre malignidade ou não do tumor.

3.3 Software de Aplicação

A ferramenta Weka, Waikato Environment for Knowledge Analysis, (HALL, 2009) foi utilizada para aplicação dos algoritmos na versão 3.8. O software consiste em um pacote para aplicações em IA, do tipo open source, desenvolvido em Java dentro das especificações da GPL (General Public License) e se consolidou como a ferramenta de AMS mais utilizada por estudantes e professores de universidades. A interface de apresentação pode ser observada na Figura 6.

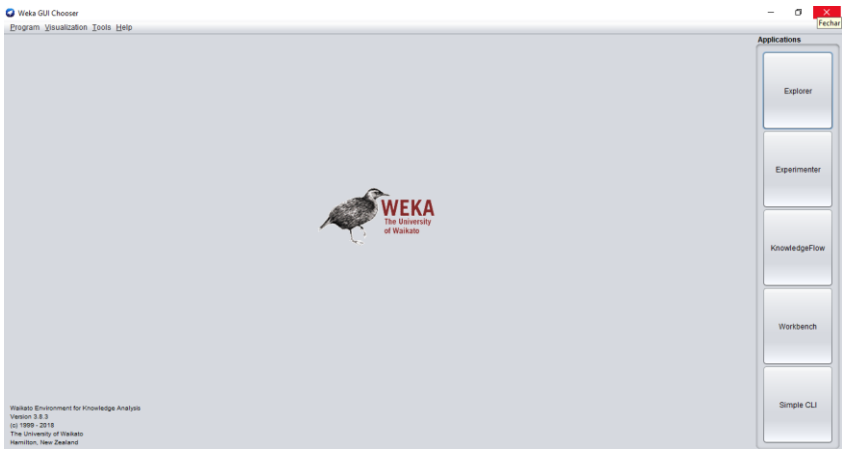

Figura 6 - Interface da plataforma Weka (Fonte própria)

O programa funciona como uma completa ferramenta para pesquisas do gênero deste estudo. Dispõe de uma vasta biblioteca de algoritmos, permite a configuração de vários parâmetros intrínsecos a cada um deles e, além disso, considera as técnicas de seleção de características e avaliação de classificadores descritas a seguir.

3.4 Seleção de características

Um grande conjunto de características não necessariamente pode ajudar a resolver um dado problema. Além de aumentar o tempo para o processamento dos dados, um grande conjunto de características pode apresentar valores redundantes ou até mesmo insignificantes para o problema em mãos, especialistas em AM denominam esse fenômeno de *Maldição da Dimensionalidade*. O princípio relata que a quantidade de dados de que se necessita para alcançar o conhecimento desejado, impacta exponencialmente o número de atributos necessários e, consequentemente, na qualidade do classificador, bem como ilustrado na Figura 7 (LENINE, 2017).

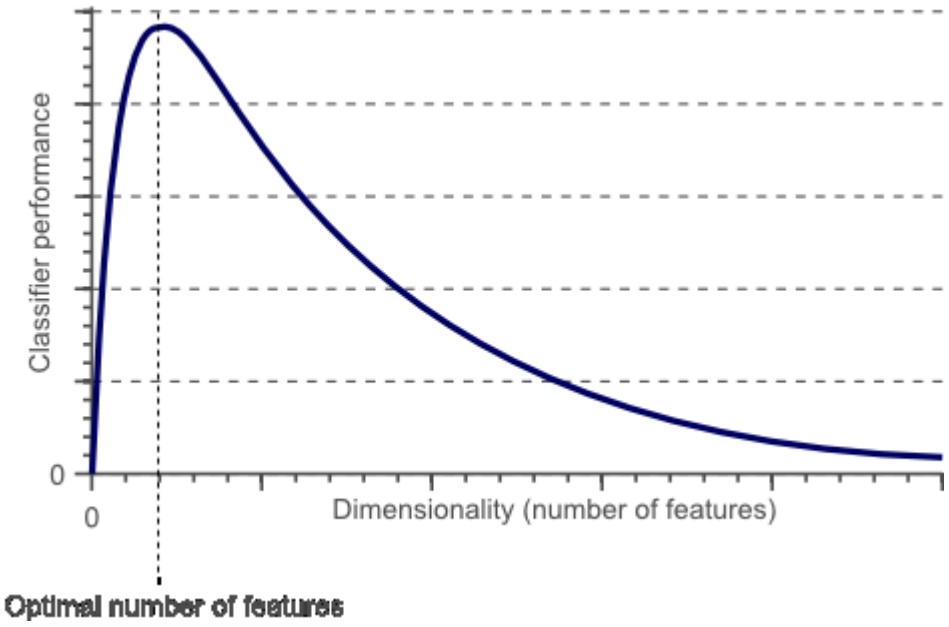

Figura 7 – Gráfico representativo da Maldição da Dimensionalidade (LENINE, 2017)

Como não se sabe a priori quais os conjuntos de dados são mais relevantes para a acurácia da classificação, utiliza-se algum método (algoritmo) capaz de buscar um conjunto "ideal" dentro do conjunto dados originalmente disponível.

A seleção de atributos utilizando a modalidade Filtro é realizada a priori, basicamente, fazendo uso do método de Seleção Baseada em Correlação (SBC), método que identifica os atributos melhores correlacionados com uma determinada classe e correlacionados entre si (MARTELLI, 1988), para executar uma busca nos atributos, ou seja, é posto por exemplo um algoritmo de pesquisa que identifica os atributos mais relevantes, e em sequência são encaminhados ao algoritmo de aprendizado os atributos selecionados.

3.5 Técnicas de avaliação dos classificadores

O último estágio do desenvolvimento de um algoritmo de AM envolve a avaliação. A eficiência do modelo proposto é calculada usando um procedimento predefinido. Os métodos dominantes apresentados na literatura são a Validação Cruzada *K-fold* e a Validação Leave-One-Out, ou *Deixar um de fora*, em tradução literal. (SCHREIBER, BESKOW, *et al.*, 2017).

3.5.1 Validação cruzada

O método de Validação Cruzada K-fold, consiste em uma técnica computacional intensiva, que usa todas as amostras disponíveis como amostras de treinamento e teste (DUCHESNE e RÉMILLARD, 2005).

Dado uma base de dados hipotética em que conste *N* registros, e definindo o k por um número inteiro, a base de dados será dividida igualmente em *k* subconjuntos. Após a divisão em subconjuntos, será utilizado um subconjunto, para ser utilizado na validação do modelo e os conjuntos restantes são utilizados como treinamento. O processo de validação cruzada é então repetido *K* vezes, de modo que cada um dos *k* subconjuntos sejam utilizados exatamente uma vez como teste para validação do modelo.

Por exemplo para *k=10*, existirão 10 subconjuntos de dados *B1, B2... B10* o primeiro passo do *k-Fold* é utilizar de *B1*a *B2* para treino e o *B10* servirá para teste. No segundo passo, *B9* é utilizado para teste e todo o restante para treino, incluindo *B10* que foi usado para teste no primeiro passo, no terceiro passo até o décimo será aplicada a mesma lógica sucessivamente. O resultado da validação *k-Fold* é o desempenho médio do classificador nos K testes. O objetivo de repetir os testes diversas vezes tem a intenção de aumentar a confiabilidade da estimativa da precisão do classificador.Por conseguir chegar a resultados mais precisos em relação a outros métodos de validação além de exigir um menor poder de processamento de recursos computacionais (DUCHESNE e RÉMILLARD, 2005), este foi o método aplicado no estudo em questão. Os valores para o parâmetro *k* mais empregados na literatura são 3, 10 e 20 (MAGLOGIANNIS, ZAFIROPOULOS, *et al.*, 2007). Neste estudo empregou-se *k=10* em todas as aplicações.

4. RESULTADOS E DISCUSSÃO

Essa seção exibirá e discorrerá sobre os resultados obtidos com a plataforma WEKA para o problema do diagnóstico automatizado de casos benignos vs. malignos de câncer de mama no caso dos dados contidos na base WDBC.

4.1 Seleção de características

O primeiro passo desta análise foi implementado para selecionar as características mais relevantes para este estudo. O método de Seleção Baseada em Correlação (SBC) filtrou o conjunto de dados WDBC para contornar a problemática da Maldição da Dimensionalidade dos Dados. Assim, a filtragem indicou os atributos mais relevantes para o problema de classificação. O ranqueamento dos melhores atributos pode ser observado na Tabela 2.

Tabela 2 - Atributos mais relevantes pelo método da SBC

Ranqueamento (médio)	***Mérito (médio)***	***Atributo***
1 +- 0	*0.794*	*30 worstConcavePoints*
2.1 +- 0.3	*0.783*	*25 worstPerimeter*
2.9 +- 0.3	*0.777*	*10 meanConcavePoints*
4.1+- 0.3	*0.743*	*5 meanPerimeter*
4.9+- 0.3	0.734	*26 worstArea*
6.2 +- 0.4	0.709	*6 meanArea*
6.8 +- 0.4	0.697	*9 meanConcavity*

De maneira análoga, o ranqueamento dos piores atributos obteve a ordenação encontrada na Tabela 3.

Tabela 3 - Atributos menos relevantes pelo método da SBC

Ranqueamento (médio)	***Mérito (médio)***	***Atributo***
29.8 +- 1.54	*0.014*	*14 seTexture*

29.4+- 1.56	*0.021*	*29 worstConcavity*
29.3 +- 1.1	*0.017*	*21 seSymmetry*
28.6 +- 0.92	*0.021*	*12 meanFractalDimension*
27.6 +- 1.28	0.040	*1 Id*
24.9 +- 1.22	0.067	*17 seSmoothness*
24.7 +- 1.1	0.079	*22 seFractalDimension*

Selecionados os atributos mais relevantes, pôde-se obter a distribuição fracionada por classes de acordo com a Figura 8. Analisando os resultados a olho nu, torna-se mais fácil a distinção visual da distribuição para os atributos nas classes de melhor mérito (ou seja, mais relevantes).

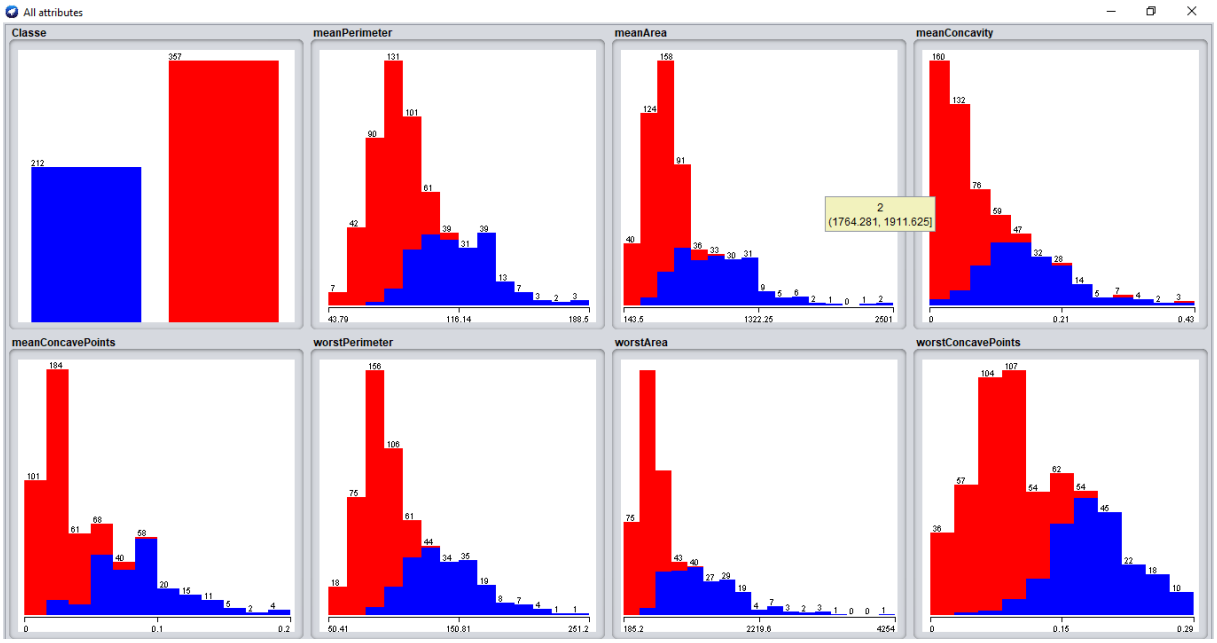

Figura 8 - Distribuições dos melhores atributos WDBC filtrados e da classe (Fonte própria)

A comportamento relatado é reiterado ao observar a distribuição do atributo menos relevante observado na Figura 9.

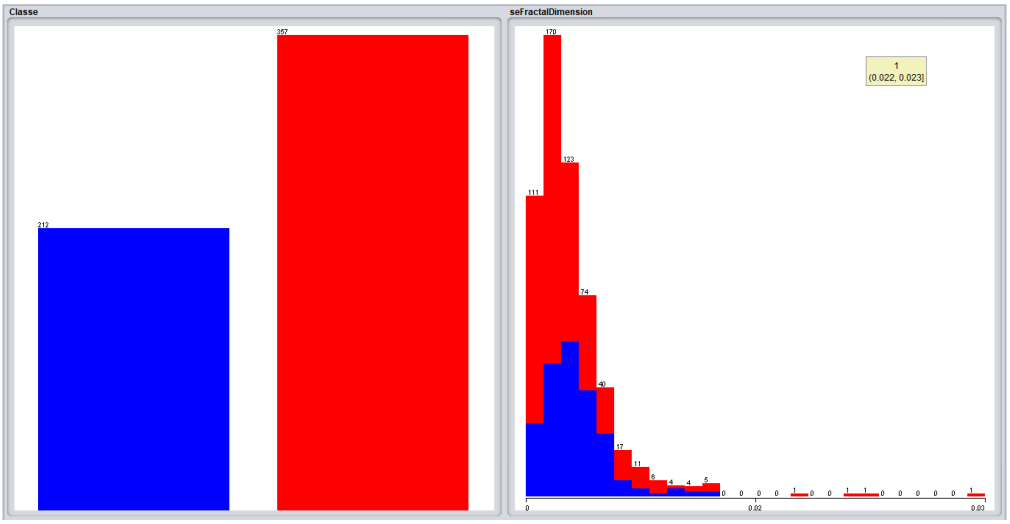

Figura 9 - Distribuições do pior atributo WDBC filtrado e da classe (Fonte própria)

4.2 Implementação dos Classificadores

Os cinco algoritmos apresentados foram aplicados na base WDBC e os resultados obtidos foram observados e interpretados a partir do número de erros e a da acurácia (taxa do número de acertos sobre o número de erros) final obtida como medida de desempenho, bem como a Tabela 4 apresenta.

Tabela 4 - Resultados de classificação obtidos

Classificador	Erros	Acurácia
Naive Bayes Classifier	34	94,025%
SVM (Linear)	47	91,740%
SVM (RBF)	12	97,891%
IBK	33	94,200%
Árvore de Decisão (J48)	42	92,619%
RNA (Multilayer Perceptron)	24	95,782%

O primeiro classificador da Tabela 4, o Naive Bayes (NB), apesar de desconsiderar a correlação entre as características, apresentou índice de acerto de 94,025%.

Já o SVM, por possuir inúmeras variáveis configuráveis, se mostrou o algoritmo mais trabalhoso de se conseguir resultados satisfatórios. A primeira vez que foi implementado, obteve índices de acurácia próximos a aleatoriedade (algoritmo de ZeroR). Por consequência as duas funções kernel que obtiveram os melhores resultados foram apresentadas na tabela 4. A função linear obteve 91,740%, o pior desempenho com 47 erros. Enquanto a Função Radial

alcançou 97,891% de apresentando a melhor taxa de classificação. Para as duas aplicações os outros parâmetros de implementação foram configurados pelo método da tentativa e erro para se obter a maior taxa de classificação.

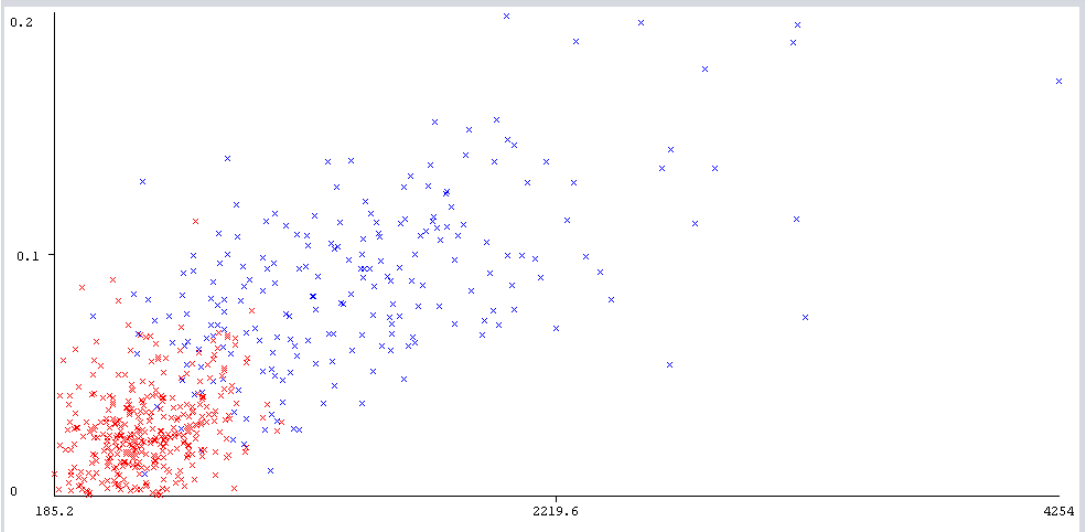

Figura 10 - Exemplo de gráfico utilizado para cálculo do SVM (Fonte própria)

A Figura 10 exemplifica um dos resultados plotados para consolidação da técnica de SVM. O eixo *X* representa os valores para característica *26 worstArea*. O eixo *Y* relata a disposição do atributo *10 meanConcavePoints*. Os pontos vermelhos são benignos e os azuis malignos. Observando a distribuição dos pontos, é possível observar como a função kernel pode auxiliar na melhora dos índices de classificação da técnica.

O algoritmo de IBK, classificou o conjunto de dados acertadamente 94,200%. A imagem exibida na Figura 11 foi extraída da plataforma WEKA e exemplifica um dos gráficos utilizados para o resultado para o cálculo.

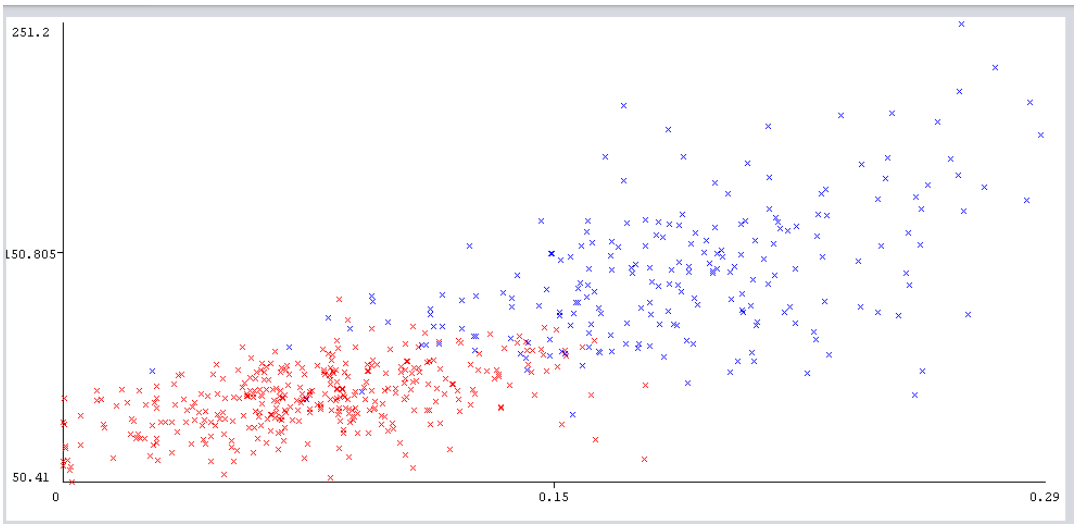

Figura 11 - Exemplo de gráfico utilizado para cálculo do IBK (Fonte própria)

O eixo X representa os valores para característica *30 worstConcavePoints*. O eixo Y relata a disposição do atributo *25 worstPerimeter*. Os pontos vermelhos são benignos e os azuis malignos. Observando a tendência consegue-se observar a correlação entre os parâmetros.

O algoritmo de árvore de decisão, apresentou o segundo pior índice de acurácia com 92,619% de acertos. Como uma das vantagens e justificativa para que esta técnica seja utilizada é o fato de o conhecimento adquirido ser representado por meio de regras. Essas regras podem ser expressas em etapas de linguagem natural, facilitando a interpretação. A regra usada como lógica de classificação criada pelo algoritmo J48 é observada pela Figura 12.

```
worstArea <= 880.8
|   worstConcavePoints <= 0.1357: B (337.0/7.0)
|   worstConcavePoints > 0.1357
|   |   worstConcavePoints <= 0.175
|   |   |   worstArea <= 719.8: B (13.0/1.0)
|   |   |   worstArea > 719.8
|   |   |   |   meanPerimeter <= 93.86: M (10.0)
|   |   |   |   meanPerimeter > 93.86: B (11.0/4.0)
|   |   worstConcavePoints > 0.175: M (15.0)
worstArea > 880.8
|   meanConcavity <= 0.0716
|   |   worstPerimeter <= 127.1
|   |   |   meanPerimeter <= 103.4
|   |   |   |   meanConcavity <= 0.05441: M (5.0)
|   |   |   |   meanConcavity > 0.05441: B (4.0/1.0)
|   |   |   meanPerimeter > 103.4: B (5.0)
|   |   worstPerimeter > 127.1: M (5.0)
|   meanConcavity > 0.0716: M (164.0)

Number of Leaves  :     10

Size of the tree :      19
```

Figura 12 - Lógica de decisão J48 (Fonte própria)

O algoritmo J48 selecionou cada nó da árvore com base no atributo dos dados que mais particiona o seu conjunto de amostras em subconjuntos tendendo a uma das categorias. O critério de partição adotado é o ganho de informação normalizado calculado pela diferença em Entropia. O atributo com maior ganho de informação normalizado escolhido para tomar a primeira decisão foi o *26 worstArea*, dividindo então a amostra em dois grupos onde o valor "*880.8"*foi usado como ponto de divisão de dois grupos, já que apresentou a melhor partição do conjunto. O algoritmo então repete a etapa anterior nas partições menores sucessivamente até que todas as características sejam utilizadas.

Dessa forma, cada atributo divisor de ramos é chamado por *Leave* (Folha). O tamanho da árvore criada é expressa pela soma da quantidade total de folhas com a quantidade de terminações classificatórias. A Figura 13 abaixo mostra a árvore de decisão criada onde observamos as folhas e os ramos.

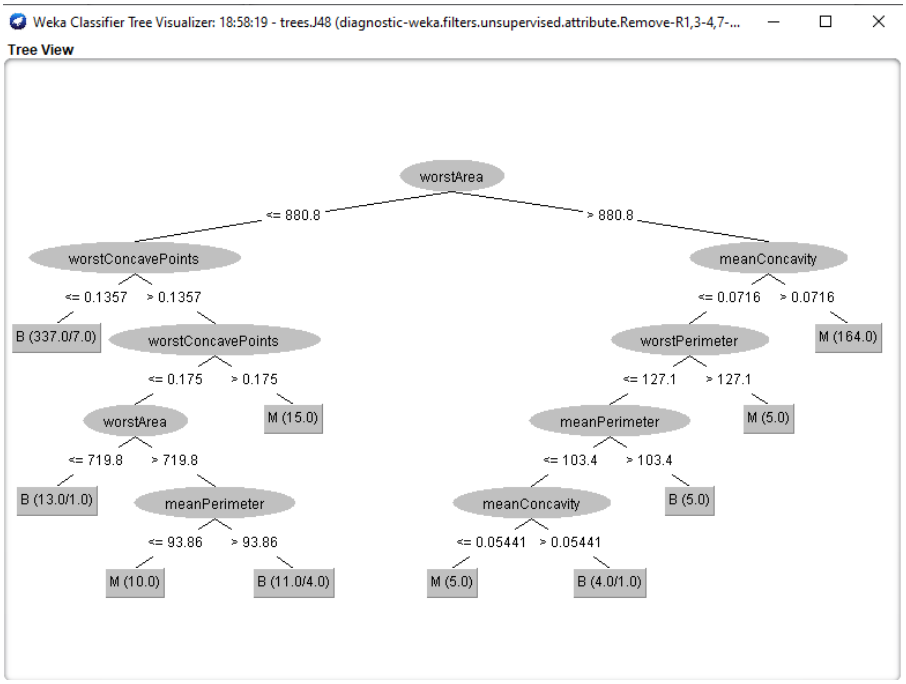

Figura 13 - Árvore de decisão J48 (Fonte própria)

O último algoritmo, a Rede Neural Artificial utilizando a lógica de Multilayer (Multicamadas) organiza mais de uma camada de neurônios em alimentação direta. Esse tipo de rede é composto por camadas de neurônios ligadas entre si por sinapses com pesos. A plataforma WEKA possibilita o acesso a estes pesos (*Weights*) que podem ser observados na Figura 14.

```
=== Classifier model (full training set) ===

Sigmoid Node 0
    Inputs    Weights
    Threshold    8.030808590914244
    Node 2    -0.13406507655896177
    Node 3    -5.459930450148593
    Node 4    -6.960318594320319
    Node 5    -4.956530193063318
Sigmoid Node 1
    Inputs    Weights
    Threshold    -8.03178822985839
    Node 2    0.1365590502463848
    Node 3    5.450510692424991
    Node 4    6.957505869352774
    Node 5    4.961100362839802
Sigmoid Node 2
    Inputs    Weights
    Threshold    -2.257597395607125
    Attrib meanPerimeter    0.05958790552196743
    Attrib meanArea    0.36938905537058764
    Attrib meanConcavity    0.6432596184000492
    Attrib meanConcavePoints    0.14984419161632145
    Attrib worstPerimeter    -0.7610370310566649
    Attrib worstArea    -0.2664028256407721
    Attrib worstConcavePoints    -0.7520444288443154
Sigmoid Node 3
```

Figura 14 - RNA Detalhamento dos pesos dos neurônios (Fonte própria)

O resultado obtido com a lógica, permitiu identificar a criação de seis nós, ou seja, seis camadas de neurônios na implementação. Da qual foi possível obter o segundo maior valor de acurácia de 95,782%.

Analisando o desempenho dos algoritmos de AM propostos neste trabalho, verificamos que o algoritmo de SVM proporcionou, com seus dois tipos de aplicação, o melhor e o pior desempenhos na classificação. Isso devido à complexidade de configuração dos parâmetros dessa ferramenta, neste caso, a escolha adequada da função kernel foi fundamental para a determinação. Já com relação aos outros classificadores, a RNA obteve o segundo melhor resultado, enquanto a árvore J48, o segundo menos assertivo.

A acurácia média obtida com todos os algoritmos empregados no estudo foi de 94,376%, desemprenho pode ser classificado como consideravelmente satisfatório levando em consideração que avaliações mamárias por mamografia apresentam acurácia média de 75% para problemas de diferenciação entre lesões benignas e malignas utilizando a classificação BI-RADS (Breast Imaging Reporting and Data System) (NASCIMENTO, SILVA, *et al.*, 2010),

Não foram analisados outros fatores de desempenho dos classificares como tempo de classificação e nível de processamento, pois a natureza curta da base de dados e do problema binário, não permitiram estudar as diferenças mais afundo.

Não foi o caso deste estudo, porém não foi incomum encontrar artigos em que a estatística convencional se provasse ser mais poderosa ou mais precisa que a técnica de AM (CRUZ, 2006). Em muitos destes casos, esse fato pode ser explicado pelas considerações iniciais do usuário da ferramenta de AM sobre interdependência e não linearidade de dados estarem equivocadas. Isso não revela, necessariamente, uma fraqueza da ferramenta, mas sim provoca a atenção para um cuidado adicional para escolha adequada da ferramenta de aplicação.

5. CONCLUSÃO

O aprendizado de máquina se mostrou uma ferramenta útil como apoio ao diagnóstico do câncer de mama. O sistema apresentado neste estudo obteve acurácia superior aos diagnósticos utilizados atualmente pelos colégios brasileiro e americano de radiologia.

Isso posto, é animador que o conceito de AM possa ser disponibilizado como uma forma de apoio ao diagnóstico da equipe de mastologistas. Para tal, a ferramenta deve ser testada em novas bases de dados e aprimorada a fim de se obter a melhor acurácia e de torná-la menos sensível à origem da base de dados possível.

No ramo do AM, a classificação correta não é sempre garantida. Como qualquer outro método, um bom entendimento do problema e o reconhecimento das limitações dos dados é importante, bem como o entendimento das premissas e limitações dos próprios algoritmos aplicados.

Neste sentido, se um experimento de AM é projetado acertadamente, e ele aprende corretamente as regras do problema e produz como resultado, classificações robustas e válidas e assim suas chances de sucesso são aprimoradas. Fica claro também, a estrita relação do sistema com a natureza dos dados que o alimentam. Se os dados forem de má qualidade, os resultados também o serão. A expressão GIGO- “Garbage in, garbage out” é empregada usualmente no ramo da IA (CRUZ, 2006).

A técnica de AM pode melhorar drasticamente o nível de diagnóstico do câncer de mama. Pesquisas mostram que médicos experientes auxiliados por mamógrafos de ponta podem detectar câncer com acurácia de 79%, enquanto um nível de assertividade de 94%, como o deste estudo, (às vezes até 97% (LAMIDI, 2018) pode ser alcançado usando técnicas de aprendizado de máquina.

Os resultados apresentados neste estudo tiveram seus parâmetros detalhados a fim de facilitar uma futura comparação com outros estudos ou mesmo verificação dos dados, visto que muito se comenta sobre a dificuldade de reprodutibilidade no meio da IA.

Uma vez que a intenção é oferecer uma futura aplicação concreta dos conceitos utilizados aqui como uma ferramenta de apoio em um equipamento de diagnóstico do tipo, como biópsia guiada, mamografia, ou mesmo em um aparelho de ressonância magnética. Para isso é necessário saber realmente um sistema de IA estar fazendo ou fará, e isso é um grande risco quando você usa Inteligência Artificial para qualquer trabalho crítico desta natureza.

Dado o exposto, observa-se a necessidade de reproduzir o experimento e analisar seus resultados tanto quanto for possível. Dessa forma, irá se garantir que o sistema se torne cada vez mais robusto e confiável.

6. REFERÊNCIAS

ACS. Breast Cancer Signs and Symptoms. **NIH**, 2017. Disponivel em: <https://www.cancer.org/cancer/breast-cancer/about.html>. Acesso em: nov. 2018.

ACS. What Is Breast Cancer in Men? **American Cancer Society**, 2018. Disponivel em: <https://www.cancer.org/cancer/breast-cancer-in-men/about/what-is-breast-cancer-in-men.html>. Acesso em: 07 jun. 2019.

AL., M. L. E. Análise das oportunidades de diagnóstico para cancer de mama. **Revista da Associação Medica Brasileira**, São Paulo, 2003.

AMORES, J.; RADEVA, P. Registration and retrieval of highly elastic bodies using contextual information. **Pattern Recognition Letters**, Belaterra, Abril 2005.

ARAUJO, A. **Novas Metodologias para Análise de órgãos e músculos a partir de imagens médicas**. Universisade do Porto FEUP. Porto, p. 15-19. 2010.

BOCCHI, L.; NORI, J. Shape analysis of microcalcifications using Radon transform. **Medical Engineering & Physics**, Florença, Jan 2006.

COSTA, E. T. E. A. Unidades Radiográficas para Mamografia. In: COSTA, E. T.; GIRÃO ALBUQUERQUE, J. A. **Equipamentos Médico-Hospitalares e o Gerenciamento da Manutenção**. 1. ed. [S.l.]: Ministério da Saúde, 2002. p. pp. 30, pp.653-682.

GONÇALVES, C. B. **Detecção de câncer de mama utilizando imagens termográficas**. Universidade Federal de Uberlândia. Uberlândia. 2017.

INCA. **CONTROLE DOS CÂNCERES DO COLO DO ÚTERO E DA MAMA**. 2ª. ed. Cadernos de Atenção Básica: INCA, v. n° 13, 2013.

INCA. **Atualização em mamografia para técnicos em radiologia**. 2ª. ed. Rio de Janeiro: Ministério da Saúde, 2019.

INCA. Estatísticas de câncer. **Instituto nacional do Cancer**, 2019. Disponivel em: <https://www.inca.gov.br/numeros-de-cancer>. Acesso em: 2 maio 2019.

INSTITUTO ONCOGUIA. Tipos de Câncer de Mama. **Oncoguia**, 2017. Disponivel em: <http://www.oncoguia.org.br/cancer-home/cancer-de-mama/20/12/>. Acesso em: mar. 2019.

LAMIDI, A. Breast Cancer Classification Using Support Vector Machine (SVM). **Towards Data Science**, nov. 2018.

LITHERLAND, J. C. et al. The impact of core-biopsy on pre-operative diagnosis rate of screen detected breast cancers. **Elsevier**, Breast Screening Training Centre, Nottingham City Hospital, Nottingham, UK, 15 jun. 2005.

MAGLOGIANNIS, I. et al. An intelligent system for automated breast cancer diagnosis. **Springer Science+Business Media**, Karlovasi, 12 Julho 2007. 24–36.

MAGLOGIANNIS, I.; PAVLOPOULOS, S.; KOUTSOURIS, D. An integrated computer supported acquisition, handling and characterization system for pigmented skin lesions in

dermatological images. **IEEE transactions on information technology in biomedicine**, p. 9(1):86–98, 2005.

MALKIN, D. D. Surgical biopsy. **Canadian Cancer Society**, 2019. Disponivel em: <https://www.cancer.ca/en/cancer-information/diagnosis-and-treatment/tests-and-procedures/surgical-biopsy/?region=on>. Acesso em: abr. 2019.

MARTELLI, A. An application of heuristic search methods to edge and contour detection. **Commun ACM**, 1988. 19:73–83.

MARTENS, H. Sensory analysis for magnetic resonance-image analysis: using human perception and cognition to segment and assess the interior of potatoes. **Wiss Technology**, p. 35(1):70–79, 2002.

MORRIS, D. An evaluation of the use of texture measures for tissue characterization of ultrasound images of in vivo human placenta. **Ultrasound Medical Biologic**, p. 14(1):387–395Morris, 1988.

MS. **Diretrizes para a Detecção Precoce do Câncer de Mama no Brasil**. 1ª. ed. Rio de Janeiro: INCA, 2015. Disponivel em: <http://www.saude.gov.br/atencao-especializada-e-hospitalar/especialidades/oncologia/diagnostico>. Acesso em: jun. 2019.

MYVMC. Fine Needle Aspiration Biopsy (FNA). **Virtual Medical Center**, 2005. Disponivel em: <https://www.myvmc.com/investigations/fine-needle-aspiration-biopsy-fna/>. Acesso em: 06 abr. 2019.

NCI. Annual Report to the Nation on the Status of Cancer. **National Cancer Institute**, 2019. Disponivel em: <https://seer.cancer.gov/statistics/>. Acesso em: 11 fev. 2019.

NÓBRAGA, E. G. **Ferramentas computacionais para a implementação de sistemas de monitoramento e diagnostico de maquinas rotativas**. Campinas, SP: Universidade Estadual de Campinas, Faculdade de Engenharia Eletrica, v. [180]f, 1992.

PERNER, P.; HOLT, A.; RICHTER, M. Image Processing in Case-Based Reasoning. **The Knowledge Engineering Review**, set. 2005. 20(3):311-314.

PROGNÓSTICO. **Dicionário infopédia da Língua Portuguesa**, 2003-2019. Disponivel em: <https://www.infopedia.pt/dicionarios/lingua-portuguesa/progn%C3%B3stico>.

SABBATINI, R. Informática: Uma Nova Especialidade Médica? **Academia Médica**, 2019. Disponivel em: <https://academiamedica.com.br/blog/informatica-medica>. Acesso em: 05 jul. 2019.

SAMEER, A. et al. Evaluation of shape similarity measurement methods for spine X-ray images. **J Vis Commun Image Represent**, p. 15(3):285–302, 2004.

SBIB - ALBERT EINSTEIN. Programa de Acompanhamento de Pacientes com Câncer de Mama. **SBIB - Albert Einstein**. Disponivel em: <https://www.einstein.br/especialidades/oncologia/atendimento-consulta/programa-de-acompanhamento-de-pacientes-com-cancer-de-mama>. Acesso em: maio 2019.

SILVA, C. Y. D. **Extração de Características de Imagens médicas utilizando wavelets para mineração de imagens e auxílio ao diagnóstico**. Instituto de Ciências Matemáticas e de Computação - ICMC - USP. São Paulo. 2007.

SIMPSON, W. R. . E. J. W. S. **System Test and Diagnosis**. http://gregstanleyandassociates.com/whitepapers/FaultDiagnosis/faultdiagnosis.htm. ed. Boston,Mass: Kluwer Academic Publishers, 1994. Disponivel em: <http://gregstanleyandassociates.com/whitepapers/FaultDiagnosis/faultdiagnosis.htm>. Acesso em: jun. 2019.

SUSAN G. KOMEN. Susan G. Komen. **SURGICAL BIOPSIES**, 2019. Disponivel em: <https://ww5.komen.org/BreastCancer/SurgicalBiopsies.html>. Acesso em: 10 ago. 2019.

TAVARES, J. **Modelos Deformáveis em Iamgem Médica**. INEB – Instituto de Engenharia Biomédica, Laboratório Sinal e Imagem. Porto. 2003.

TSOTSOS, J. Knowledge organization and its role in representation and interpretation for time-varying data: the ALVEN system. **Comput Intell**, 1985. 1(1):16–32.

VIERGEVER, M. A survey of medical image registration. **Med Image Anal**, 1998. 2(1):1–16.

WANG, et al. A support vector machine-based ensemble algorithm for breast cancer diagnosis. **European Journal of Operational Research**, n. Elsevier, 2017. ISSN doi 10.1016/j.ejor.2017.12.001.

WOLBERG, W. H. et al. Wisconsin Diagnostic Breast Cancer (WDBC). **Machine Learning for Cancer Diagnosis and Prognosis**, Wisconsin, 1995. Acesso em: 15 ago. 2018.