UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
PROGRAMA DE PÓS-GRADUCÃO EM ENGENHARIA MECÂNICA

YAGO DE MIRANDA AGUIAR

DESENVOLVIMENTO E APLICAÇÃO DE METODOLOGIA DE APRENDIZAGEM DE MÁQUINA PARA CLASSIFICAÇÃO DE IMAGENS TERMOGRÁFICAS NA ÁREA MÉDICA

Recife

2021

YAGO DE MIRANDA AGUIAR

DESENVOLVIMENTO E APLICAÇÃO DE METODOLOGIA DE APRENDIZAGEM DE MÁQUINA PARA CLASSIFICAÇÃO DE IMAGENS TERMOGRÁFICAS NA ÁREA MÉDICA

Dissertação apresentada ao programa de Pós-Graduação em Engenharia Mecânica, PPGEM, da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Engenharia Mecânica. Área de concentração: Energia

Orientador (a): Profª. Drª.Rita de Cássia Fernandes de Lima

Coorientador (a): Prof. Dr. Marcus Costa de Araújo

Recife

2021

Catalogação na fonte:
Bibliotecário Josias Machado, CRB-4 / 1690

A282d Aguiar, Yago de Miranda.
Desenvolvimento e aplicação de metodologia de aprendizagem de máquina para classificação de imagens termográficas na área médica. / Yago de Miranda Aguiar. – 2021.
104 f.: il., figs., tabs., abrev. e sigl.

Orientadora: Prof.ª Dr.ª Rita de Cássia Fernandes de Lima.
Coorientadora: Prof. Dr. Marcus Costa de Araújo.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia Mecânica, Recife, 2021.
Inclui referências.

1. Engenharia mecânica. 2. Termografia 3. Aprendizagem de máquina. 4. Orange Canvas. 5. Câncer de mama. I. Lima, Rita de Cássia Fernandes de (Orientadora). II. Araújo, Marcus Costa de (Coorientador). III. Título.

UFPE

621 CDD (22. ed.) BCTG/2022-126

YAGO DE MIRANDA AGUIAR

DESENVOLVIMENTO E APLICAÇÃO DE METODOLOGIA DE APRENDIZAGEM DE MÁQUINA PARA CLASSIFICAÇÃO DE IMAGENS TERMOGRÁFICAS NA ÁREA MÉDICA

Dissertação apresentado ao programa de Pós-Graduação Em Engenharia Mecânica, PPGEM, da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Engenharia Mecânica. Área de concentração: Energia

Aprovado em: 10 de setembro de 2021

BANCA EXAMINADORA

_____

Profa. Dra. Rita de Cassia Fernandes de Lima (Orientadora)
Universidade Federal de Pernambuco

_____

Prof. Dr. Marcus Costa de Araújo (Coorientador)
Universidade Federal de Pernambuco

_____

Prof. Dr. Álvaro Antônio Ochoa Villa (Examinador Externo)

_____

Profa. Dra. Renata Maria Cardoso Rodrigues de Souza
(Examinador Externo)

AGRADECIMENTOS

Agradeço primeiramente a minha família pelas oportunidades, suporte e apoio que me deram nessa caminhada. À minha orientadora Professora Rita de Cássia Fernandes de Lima por todo direcionamento e conhecimento compartilhado na elaboração deste trabalho, e que, mesmo em meio as dificuldades que a pandemia trouxe, semprese fez presente quando eu precisava contacta-la para dúvidas. Ao meu coorientador, Professor Marcus Costa de Araújo, por ter me dados muitas dicas e ter me introduzidoao mundo da aprendizagem de máquina. E por fim, à todos que estão comigo no dia-a-dia e que de alguma forma me ajudaram a chegar até aqui.

RESUMO

O presente trabalho tem como objetivo aplicar uma metodologia de classificação de imagens termográficas mamárias, por meio de uma plataforma computacional de código aberto (Orange Canvas), e avaliar o impacto nos resultados pelo uso de diferentes formas de segmentação de imagem e técnicas de redução de desbalanceamento edimensionalidade. Foram avaliadas duas diferentes bases de dados de imagens termográficas de pacientes do Hospital das Clínicas da UFPE. Utilizou-se cinco algoritmos declassificação, que aliados às técnicas de SMOTE e PCA/Rank/PSO, obtiveram comoresultado: 96,2% de Acurácia e 99,5% de Sensibilidade ao Maligno para classificação binária (Câncer x Não-Câncer), e 65,6% de Acurácia e 92,2% de Sensibilidade aoMaligno para classificação em quatro classes (Maligno, Benigno, Cisto e Normal).

Palavras-chave: termografia; aprendizagem de máquina; orange canvas; câncer de mama.

ABSTRACT

The present work aims to apply a methodology for classification of breast thermographic images using two open source computational platforms (Matlab and Orange Canvas), and to evaluate the impact on the results by using different forms of image segmentation, techniques of balancing and dimensionality reduction. Two databases of thermographic images of patients from the Hospital das Clínicas of UFPE were evaluated. Five classifi- cation algorithms were used, which combined with the SMOTE and PCA/Rank/PSO techniques, obtained as a result: 96.2% Accuracy and 99.5% Sensitivity to Malignant for binary classification (Cancer x Non-Cancer), and 65.6% Accuracy and 92.2% Sensitivity to Malignant for classification into four classes (Malignant, Benign, Cyst and Normal).

Keywords: thermography; machine learning; orange canvas; breast cancer.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACS	*Academy Cancer Society*
AG	Algoritmo Genético
ANN	*Aritificial Neural Network*
AUC	*Area Under Curve*
BPSO	*Binary Particle Swarm Optimization*
CAD	*Computer-Aided Design*
CNN	*Condensed Nearest Neighbor*
FP	Falso Positivo
FN	Falso Negativo
GR	*Gain Ratio*
IG	*Information Gain*
INCA	Instituto Nacional de Câncer
KNN	*K-Nearest Neighbor*
OMS	Organização Mundial de Saúde
PSO	*Particle Swarm Optimization*
PCA	*Principal Component Analysis*
RF	*Random Forest*
ROC	*Receiver Operating Characteristics*
ROI	*Region of interest*
RL	Regressão Logística
SRM	*Structural Risk Minimization*
SVM	*Support Vector Machine*
SUS	Sistema Único de Saúde

SMOTE	*Synthetic Minority Oversampling Technique*
TA	Transformação de Atributos
TIR	Termografia Infravermelho
UFPE	Universidade Federal de Pernambuco
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo

LISTA DE SÍMBOLOS

β	Letra grega Beta
δ	Letra grega Delta
σ	Letra grega Sigma
C_1	Constante de aceleração individual
C_2	Constante de aceleração global
G_{best}	Melhor solução global
k	Número de vizinhos
MD	Mama direita
ME	Mama esquerda
max	Máximo
min	Mínimo
n	Quantidade de amostras
P_{best}	Melhor solução individual
Rand	Valor randômico entre 0 e 1
Sp^2	Variância amostral
V_{id}	Vetor velocidade
W	Peso de inércia

SUMÁRIO

1 INTRODUÇÃO

O presente capítulo tem como objetivo trazer informações que visem explicar o porquê do desenvolvimento deste trabalho, bem como o mesmo foi estruturado.

1.1 Contexto e Motivação

Atualmente, a engenharia e biologia vêm se fundindo cada vez mais para desenvolver inúmeros dispositivos e aplicações que auxiliam na manutenção do corpo humano em um estado saudável. Nos primórdios a medicina tinha como objetivo apenas curar os pacientes que já estavam enfermos. Entretanto, com o advento deum maior poder tecnológico, as perspectivas de um correto diagnóstico prévio à uma possibilidade de doença futura se tornaram reais.

Hoje estamos vivendo na era dos dados e uma grande quantidade de dadosclínicos de pacientes, pode ser utilizada pela biociência para estudar a fundo o corpo humano (ARAÚJO *et al.*, 2008). Por isso, formulações matemáticas juntamente com estruturas computacionais podem, de forma muito mais rápida e precisa, detectar ouprever enfermidades, ajudando o especialista a fornecer um melhor diagnóstico.

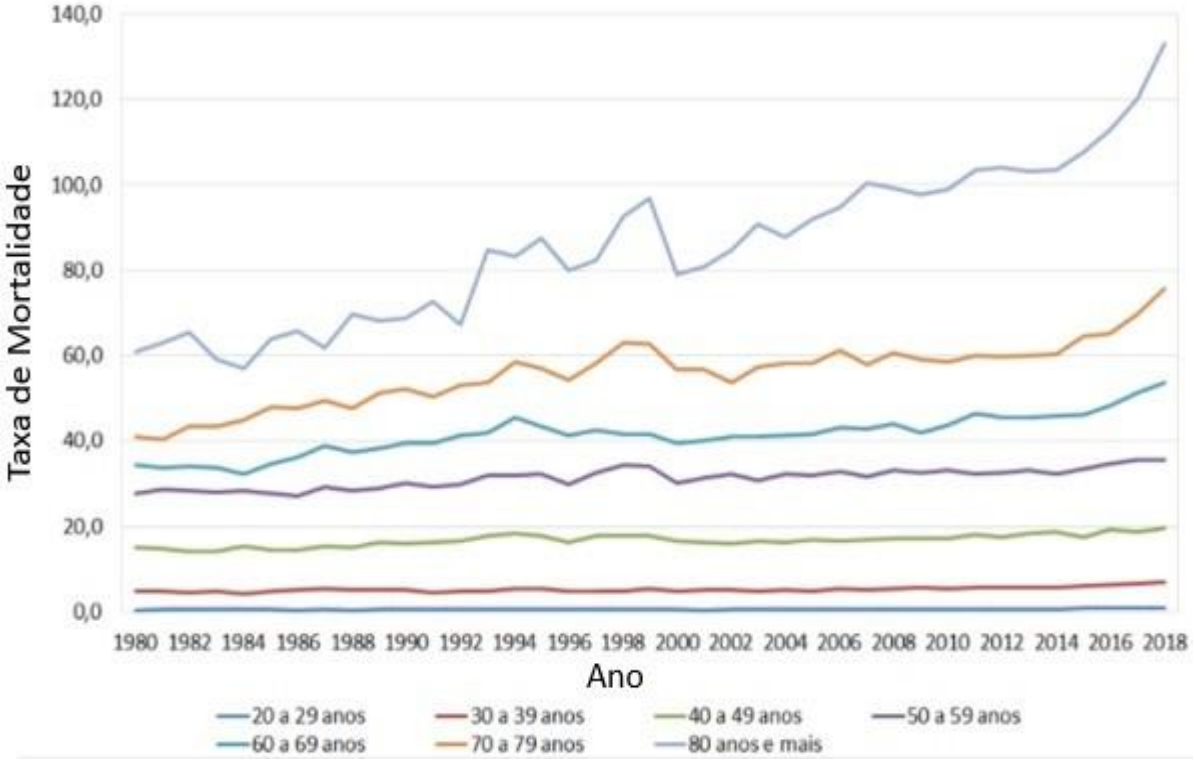
Algumas enfermidades, pelo fato de terem maior incidência e apresentarem uma maior taxa de mortalidade, têm maior destaque no tocante ao desenvolvimento do diagnóstico prévio com auxilio das técnicas de engenharia e de computação. Dentre elas, é possível apontar o câncer, em suas diferentes formas, como sendo uma das mais investigadas nos últimos anos.

No Brasil, excetuando-se os tumores de pele não-melanoma, o câncer de mama é o de maior ocorrência em todas regiões do país. Para o ano de 2020, foram estimados 66.280 casos, isso representa um taxa de mais de 40 casos para cada 100.000 mulheres (INCA, 2020).

O câncer de mama é avaliado como sendo de fácil prognóstico, porém as taxas de mortalidade se mantêm altas. Esse fato tem ligação direta com o estágiomais avançado da doença quando ela é diagnosticada. Em países desenvolvidos ataxa de sobrevida cinco anos após o diagnóstico chega a 85%, enquanto em países em desenvolvimento essa mesma taxa permanece na casa dos 60%, o que explicaa importância de um investimento tecnológico para detecção precoce da doença (INCA, 2020). Estudos afirmam que quando diagnosticado cedo, o paciente podechegar a 95% de se curar, isso se torna o maior estímulo

para o desenvolvimento detécnicas que possam detectar cada vez mais cedo a patologia (H. DO CANCER DEBARRETOS, 2015 *apud* BAFFA; LATTARI, 2018). Na Figura 1, é possível ver com clareza a importância da descoberta da doença em idades mais jovens.

Figura 1- Taxa de mortalidade por câncer de mama em mulheres no Brasil, específicas por faixas etárias, por 100.000 mulheres.



Fonte: (INCA, 2020)

As principais técnicas difundidas utilizadas para a detecção do câncer de mamasão a mamografia, ultrassonografia e ressonância magnética. A mamografia, que é ofertada pelo Sistema Único de Saúde (SUS) à todas mulheres acima de 40 anos, é a mais indicada forma de detecção desse tipo de câncer. Entretanto, como essa técnica faz uso da emissão de feixes de Raios X para gerar a imagem da mama, o exame podeter sua acurácia prejudicada para pacientes jovens, pois as mesmas têm a mama mais densa, dificultando na geração do contraste dos Raios X (KOAY *et al.*, 2004). Com a mama sendo composta predominantemente por tecido glandular denso, há também mais suscetibilidade à absorção da radiação do exame, elevando a chance de contrairo próprio câncer (HAYWARD, 1987 *apud* ARAÚJO, 2014).

A busca para preencher as lacunas da detecção que ainda existem na ma- mografia, fez aumentar o interesse na termografia infravermelha (TIR), quanto aoseu desenvolvimento e aplicação. Dentre suas vantagens estão o seu baixo custo, a não utilização de qualquer

radiação e de procedimentos invasivos (NG, 2009). Seu funcionamento está baseado na medição, através de uma câmera infravermelha, da radiação emitida pelo corpo humano, quantificando a temperatura do mesmo, emespecial da mama (KANDLIKAR *et al.*, 2017). A imagem termográfica gerada pela câmera permitirá observar, através de uma matriz de temperaturas, as alterações eventualmente presentes devido a alguma anormalidade mamária.

Devido à alta atividade metabólica, quando há alguma anomalia no tecido mamário a região tem sua temperatura modificada. Essa alteração faz que diversos pontos e a vizinhança da área mamária tenham sua temperatura aumentada. Esse grande indício de anormalidade é a chave para a detecção na termografia (GOLESTANI *et al.*, 2014).

Embora os trabalhos pioneiros com o uso da TIR para detecção em câncer de mama tenham sidos iniciados na década de 50, eles só voltaram ao foco de estudo a partir dos anos 2000. A grande ação impulsionadora foi a notável melhoria tecnológica das câmeras infravermelhas (ARAÚJO, 2009). O uso das câmeras em associação com análises estatísticas e recursos de aprendizagem de máquina, tornou-se uma forma ainda mais potente de otimização dessa técnica (NG *et al.*, 2006). De acordo com Ng (2009) a termografia tem capacidade de detectar tumores malignos com dez anos de antecedência se comparada com outras técnicas, como a mamografiapor exemplo. Além disso, a TIR aliada à mamografia e ao exame clínico prévio pode chegar a ter 98% de potencial de detecção (NG; SUDHARSAN, 2004). No Brasil, para o diagnóstico do câncer de mama, essa técnica ainda não é utilizada nem como mecanismo complementar, sobretudo por algumas instabilidades no diagnóstico final pelas diferentes formas da análise das imagens termográficas.

Tendo conhecimento das vantagens e das grandes chances de uma detecção prévia da TIR com auxílios de ferramentas de aprendizagem de máquina, e buscandotorná-la de fácil acesso para uso diário à profissionais de outra área, esse trabalho investiga formas de classificação de imagens termográficas na forma binária e multiclasses utilizando uma plataforma computacional de código aberto e não muito explorada em outros trabalhos. Pretende-se apresentar técnicas que visam melhorar problemas comuns na classificação de imagens, como o desbalanceamento e tempo de processamento.

O presente estudo é parte de um projeto de pesquisa intitulado “Análise da viabilidade do uso de câmera termográfica como ferramenta auxiliar no diagnóstico decâncer de mama em hospital público localizado em clima tropical", que foi aprovado pelo Comitê de Ética da Universidade Federal de Pernambuco (UFPE) – Brasil e registrado no Ministério da Saúde sob número CEP/CCS/UFPE N0279/05.

O presente trabalho encontra-se incluído na pesquisa de um grupo, podendo serdestacado

a seguinte produção bibliográfica, referente ao tema aqui estudado: Neto eLima (2015), NOVA (2017), Vasconcelos *et al.* (2018), Araújo *et al.* (2021), Queiroz *et al.* (2021), FREITAS (2021).

1.2 Objetivos

O objetivo geral deste trabalho foi o desenvolvimento de uma metodologia que possa propiciar o uso de ferramentas que empregam aprendizagem de máquina e otimização, de forma que possa facilitar a detecção precoce do câncer de mama,com o intuito de permitir a realização de triagens rápidas e até auxiliar o médico no diagnóstico.

Objetivos específicos

Para atingir o objetivo deste trabalho, foram seguidos os seguintes objetivos específicos:

- Extração de características das regiões de interesse das imagens termográficas.
- Utilização de uma plataforma computacional aberta e amigável, que permitaprofissionais de áreas distintas seguirem a metodologia do trabalho.
- Seleção e transformação das características extraídas das imagens.
- Aplicação de métodos para tratar de desbalanceamento da base de dados.
- Seleção de algoritmos de classificação.
- Realização de classificações, binárias e multiclasses, investigando os melhoresparâmetros para os algoritmos de classificação.
- Utilização de testes estatísticos para avaliar os melhores resultados oriundos dos experimentos.

1.3 Organização do trabalho

O presente trabalho está dividido em seis capítulos. No primeiro capítulo foram introduzidas as informações básicas sobre o câncer de mama, seus mais recentesdados na população e a apresentação, com contexto histórico, da importância da termografia infravermelha como ferramenta auxiliar para detecção do câncer de mama.

O Capítulo 2 traz informações relevantes de outros trabalhos: Tópicos como termografia na área médica de um modo geral; segmentação de imagens; métodos de balanceamento; classificação via algoritmos e da termografia no auxílio ao diagnósticodo câncer de mama propriamente dito.

O Capítulo 3, através da fundamentação teórica, abordou mais detalhadamente os conceitos como: o câncer de mama; disgnóstico precoce; termografia; segmentação de imagens; extração de características; aprofundamento dos algoritmos utilizados e as métricas de avaliação.

No Capítulo 4 é mostrado o passo-a-passo das metodologias utilizadas no presente trabalho. De forma que foram detalhados os seguintes passos: Aquisição e entendimento das duas bases de dados utilizadas; segmentação manual e automática;extração das características; técnicas de redução de desbalanceamento; técnicas de redução de dimensionalidade; classificação e análise estatística.

No Capítulo 5 os melhores resultados alcançados, com a aplicação da metodologia, são mostrados tanto para uma classificação em quatro classes como para uma binária. Além disso, é exibido o teste de hipótese que verificou a significância estatística dos resultados entre as diferentes segmentações.

No Capítulo 6 são apresentadas as conclusões dos resultados obtidos, levando em consideração os objetivos propostos inicialmente. Também são indicadas sugestõespara trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Neste capítulo foram compiladas algumas pesquisas que se relacionam com o temas de desenvolvimento do trabalho, para que houvesse a possibilidade de exploração e comparação com as técnicas utilizadas.

2.1 Termografia na área médica

Alterações de temperatura do corpo humano provocadas por doenças foramrelatadas desde a antiguidade. Em 480 a.C., Hipócrates relatou que a temperaturade uma região infeccionada era maior que outras partes parecidas do mesmo corpo (SANCHES, 2009). Séculos depois, Galileu inventou o termoscópio, equipamento que possibilita a visualização da temperatura com auxilio de grânulos (ASSIS, 2015).

Em 1800, Willian Herschel descobriu a tecnologia infravermelha, até então desconhecida pelo meio científico, tendo feito essa descoberta por acaso, enquanto ele fazia experimentos com vidros. Ele foi percebendo que algumas das amostras passavam calor do sol e outras bem pouco. O filho de Herschel, John Herschel, deucontinuidade às pesquisas do pai, e em 1840, tornou-se a primeira pessoa a criar umaimagem térmica, com auxilio de películas de petróleo (LAHIRI *et al.*, 2012).

Quase 100 anos depois, em 1934, Hardy apresentou o papel fisiológico da emissão infravermelha pelo corpo humano, e estabeleceu a importância do diagnóstico médico com técnicas de medição de temperatura. Dois anos depois comprovou que anomalias poderiam ser descobertas através de imagens termográficas (LAHIRI *et al.*, 2012; HARDY; MUSCHENHEIM, 1936). Entretanto, somente em 1960 a técnica da termografia infravermelha na medicina foi relatada de fato. Esse intervalo grande de tempo se deu devido à falta de tecnologia para os equipamentos necessários no processo (LAHIRI *et al.*, 2012; RING, 2010).

Desde então, a TIR vem se desenvolvendo com câmeras termográficas cada vezmais sensíveis, e aplicando a técnica de detecção e auxilio ao diagnóstico de anomalias em diversos campos da medicina. Atualmente, essa técnica vem sendo cada vez maisutilizada em áreas como: neuropatia diabética, desordem vascular, estudo de termo regulação, dermatologia, dor muscular, câncer de mama, ginecologia, imageamento cerebral, transplante renal e odontologia (LAHIRI *et al.*, 2012; RING; AMMER, 2000).

A neuropatia diabética consiste em várias manifestações clínicas em que a perda de sensibilidade aparece com frequência. É observado também que temperaturas altas na parte abaixo do pé, juntamente com a perda da sensibilidade podem ser um indicador de ulceração do pé. Bagavathiappan *et al.* (2010) fizeram uso da TIR para investigar acorrelação desse aumento de temperatura abaixo do pé com a diabete neuropática. Imagens termográficas da parte de baixo do pé de 112 pacientes com diabetes do tipo2 foram feitas. Em 33 desses pacientes a neuropatia diabética estava presente. Foram percebidas diferenças significativas de temperatura entre pacientes com a neuropatia e sem. Os que continham tiveram a temperatura do pé entre 32°C e 35°C, e os que não continham, entre 27°C e 30°C. Além disso, a média da temperatura dos pés, quefoi calculada em 6 diferentes regiões dos mesmos, também mostrou ser maior para pacientes com a neuropatia. Dessa forma, foi constatado que imagens termográficas podem servir como ferramente auxiliar à avaliação de diabetes no pé.

Bouzida *et al.* (2009) estudaram a termo regulação do corpo humano com auxílio da termografia. O trabalho analisou dois mecanismos aplicados ao corpo humano paraentender como ele se comporta com relação à geração e perda de calor. O primeiro experimento foi um estímulo de modulação do fluxo sanguíneo em um antebraço eo segundo manteve uma das mãos do voluntário num ambiente mais frio. Foi observado que o sistema regulatório humano imediatamente trabalhou para equilibrar atemperatura corporal nos dois casos. Além disso, as imagens termográficas também serviram para visualizar as características das veias durante o processo de contraste de temperaturas.

O intervalo de pós-morte é uma das questões de maior dificuldade que a medicina forense encontra. Eles utilizam a temperatura da cabeça como referência para estimativa do tempo passado, essa parte é escolhida por ser uma região menos afetada, que o restante do corpo, à mudanças de temperatura causadas por vestuário, peso corporal ou atividade física. O método mais comum para medir essa temperatura é coma perfuração da membrana timpânica dos ouvidos, entretanto essa técnica é bastanteinvasiva. Cattaneo *et al.* (2009) propuseram o uso da termografia infravermelha para a medição dessas temperaturas nos dois ouvidos. O estudo utilizou 25 cadáverespara comparar a medição feita pela TIR e pelo método já utilizado. Como resultado eles obtiveram que a média das medições não possuía diferenças estaticamente significativas, com isso puderam constatar que a técnica usando a TIR pode ser levadaem conta para esse tipo de análise.

Benko *et al.* (1996) examinaram, com auxílio da TIR, o comportamento da absorção da Radiação Beta na temperatura corporal humana em seis pacientes que estavam fazendo uso da radioterapia. Como resultado eles observaram que houve diferenças significativas das

temperaturas da região que sofrera a radiação, tanto no dia do tratamento como também dias depois. Foi verificado que essas diferenças se tornaram maiores com o passar dos períodos de tratamento, chegando a 0,9°C de diferença no 16°dia de tratamento. Com essas evidências, os autores recomendaramo uso da TIR para acompanhamento de pacientes que virão a se submeter a radiaçãobeta para tratamentos, e até mesmo aqueles que serão expostos à radiação com doses mais baixas.

Visando analisar o uso da TIR em uma área de poucas aplicações na odontologia, Kasprzyk-Kucewicz et al. (2020) examinaram a viabilidade do uso de imagens termográficas na localização de inflamações e monitoramento de efeitos pós retirada do terceiro molar. O estudo consistiu de 27 pacientes que tiveram imagens termográficas feitas da região onde foi realizada o procedimento. Essas imagens foram registradas antes, logo após, quatro e sete dias após o procedimento. A partir das comparações das temperaturas da região nos quatro momentos, foi constatado que houve modificações significativas na temperatura da região onde houve o reparo, com variações de até 2,5°C. Logo, foi concluído que a TIR é uma ferramenta útil na detecção da localização de processos inflamatórios em tecidos periodontais.

De forma geral, é notável que a TIR tem um espaço cada vez maior nas diferentes áreas médicas, sendo as imagens termográficas uma alternativa a mais, e menos custosa, a diversos processos já estabelecidos na medicina. É importante ressaltar,que um assertivo diagnóstico será feito apenas com uma correta interpretação dasimagens termográficas, assistidas por ferramentas de classificação, e com simulações numéricas de perfis de temperatura (QUEIROZ *et al.*, 2016).

2.2 Segmentação de imagens

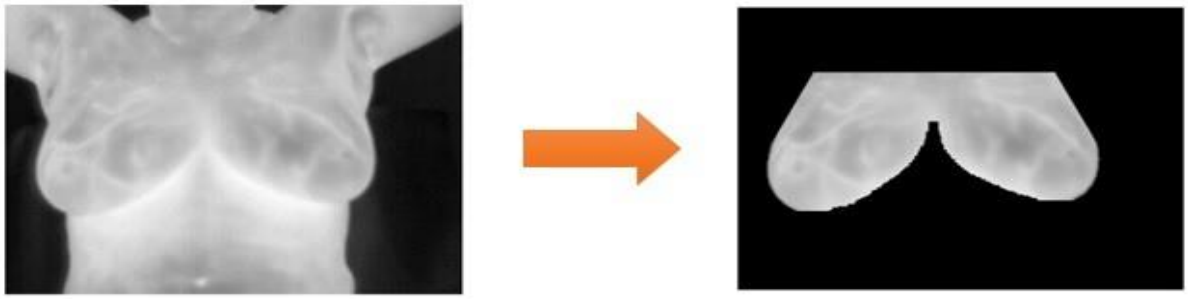
A segmentação de imagens é necessária quando se planeja extrair características de uma região específica da imagem. Para isso, várias formas de segmentações vêm sendo desenvolvidas. A segmentação da região de interesse (ROI, *region of interest*), no caso da detecção do câncer de mama, é considerada difícil pela falta depadrão no formato das mamas. A literatura tem buscado diferente níveis de automaçãopara alcançar uma segmentação certeira e de simples execução (GONÇALVES *et al.*, 2017).

Dourado (2014) elaborou uma aplicação para segmentação de imagens ter- mográficas de forma automática. Esse método é feito com uma rotina na plataforma Matlab, onde as matrizes de temperatura são tratadas como imagens em tons de cinza,sendo os tons mais claros correspondentes às temperaturas mais altas. Os passos se seguem com a definição dos limites

superior e inferior, retiradas das regiões laterais superiores residuais e da borda inferior. O resultado final da segmentação automáticapode ser visto na Figura 2.

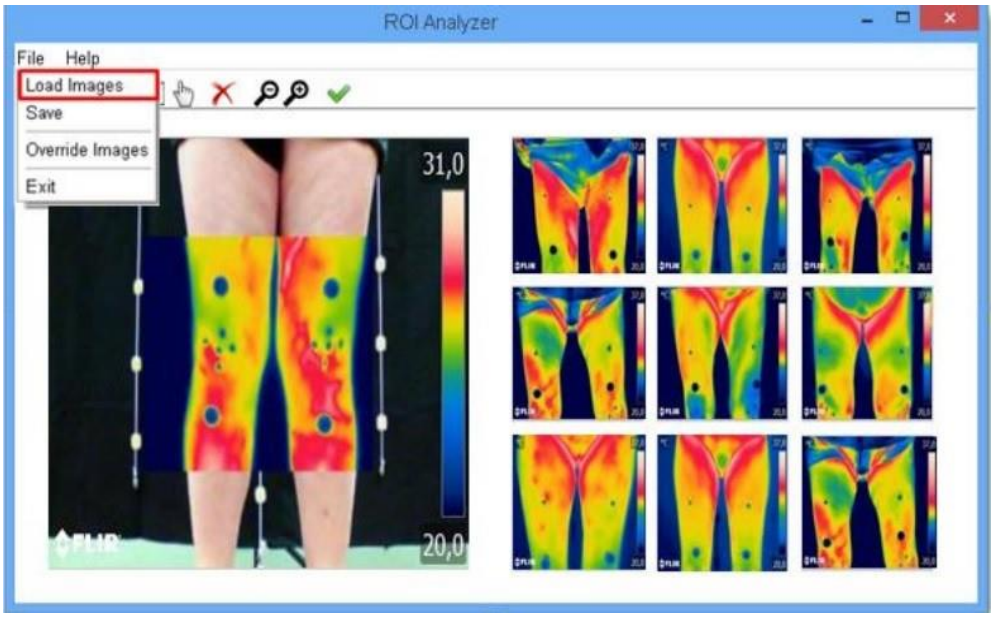
Silva *et al.* (2015) desenvolveram um sistema inteligente para segmentação semi-automática de imagens termográficas. Nessa aplicação, o usuário segmenta uma imagem termográfica padrão, com mais formas geométricas que as plataformascomerciais oferecem, e essa segmentação é sobreposta nas outras imagens. Todos osdados estatísticos são recolhidos e processados para serem apresentados através de gráficos para que o usuário consiga ter uma percepção e visão global do problema. Um exemplo do carregamento das imagens dessa aplicação é ilustrada na Figura 3.

Figura 2– Exemplo da segmentação automática



Fonte: Dourado (2014)

Figura 3– Carregamento de imagens termográficas no ROI Analyzer



Fonte: Silva *et al.* (2015)

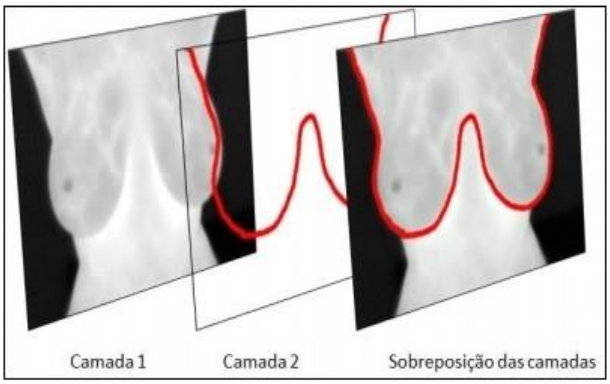
Araújo (2014) desenvolveu uma técnica de segmentação manual utilizando um

programa na plataforma Matlab, que importa os valores da matriz de temperaturas daimagem escolhida, e gera uma outra imagem, usando de pseudo-cores relacionadas às medidas de temperatura. A segmentação da ROI é feita com o auxílio de elipses geradas manualmente que, por sua vez, são sobrepostas à imagem original para asegmentação final das mamas.

Sabendo que a forma mais simples para se determinar uma ROI é delimitando com papel e caneta, Marques *et al.* (2012) propuseram uma ferramenta de segmentação para tablets com *touchscreen*. Dentre suas vantagens estão a mobilidade e flexibilidade promovidas pela execução do aplicativo em um dispositivo móvel. Entretanto, ainda háuma certa dificuldade nessa tipo de segmentação devido à uma não delimitação da região superior das mamas.

A Figura 4 mostra o funcionamento da ferramenta na delimitação da ROI.

Figura 4– Camadas da delimitação da ROI pela aplicação via tablet



Fonte: Marques *et al.* (2012)

2.3 Métodos de balanceamento

A aprendizagem através de padrões mediante base de dados desbalanceadas é uma tarefa muito difícil, pois a classe majoritária irá enviesar o resultado do processo.Esse problema é algo que ocorre com certa frequência em diagnósticos médicos ou em fraudes de cartão de crédito. O número de pacientes doentes ou transações fraudulentas é bem menor que o oposto (POZZOLO *et al.*, 2013). Entretanto, diversas técnicas vêm sendo utilizadas na literatura para diminuir o impacto que o desbalanceamento causa.

Stoll *et al.* (2020) estudaram a influência nos resultados que as técnicas de desbalanceamento de *oversampling*, SMOTE, e *undersampling*, NCR, tiveram numa base desbalanceada de alunos que se formaram na universidade de Harvard e noMIT. Eles observaram que após aplicação das técnicas os resultados melhoraramconsideravelmente,

tendo a especificidade aumentado de 40% para 97% com SMOTEe 93% com NCR utilizando o algoritmo KNN.

Han *et al.* (2005) desenvolveram duas técnicas adaptadas da técnica SMOTE. Nesses novos procedimentos, chamados de *Bordeline*-SMOTE1 e *Bordeline*-SMOTE2, há uma diferença em relação a quais objetos da classe minoritária serão usados para sintetização de novos elementos. Apenas são utilizados os selecionados da região de *bordeline*. Experimentos indicaram uma melhora nas métricas de classificação utilizando o método proposto.

Eitrich *et al.* (2007) investigaram a classificação de uma base desbalanceada de drogas utilizando a técnica de custo sensitivo. O pior que pode acontecer para essa base é um falso negativo. Assim, os autores impuseram, através de um fator, uma penalidade maior caso fosse constatado um falso negativo, com intuito de o próprio algoritmo equilibrar o enviesamento causado pela desproporção entre as classes.

Milaré *et al.* (2009) utilizaram uma abordagem híbrida de *undersampling* para avaliação em dez base de dados desbalanceadas. Nessa abordagem, a classe majoritária é randomicamente subdividida em subconjuntos que contêm a mesma quantidade de elementos que a classe minoritária. Com isso, haverá um número de subconjuntosde bases balanceadas para fazer a aplicação do algoritmo de classificação.

2.4 Métodos de redução de características

A redução de características traz diversas vantagens a modelos de classificação.Ela pode ajudar a melhorar as métricas de avaliação, diminuir o tempo de processamento e a chance de *over-fitting* (GENG *et al.*, 2007). A retirada de atributos que não agregam mais informação para o problema, ou até mesmo a diminuição da quantidade total de atributos feita com a transformação deles, é realizada com auxílio de diversastécnicas estudadas pela literatura.

Muštra *et al.* (2012) utilizaram em seu trabalho de classificação o método *BestFirst* com pesquisas do tipo: *forward*, *backward*, bidirecional e randômica (com 25% dos atributos iniciais) para reduzir 419 atributos extraídos de imagens mamográficas. Como conclusão, obtiveram uma melhora na acurácia da classificação na faixa de 3% a12% após a redução dos atributos, com as pesquisas *forward* e *backward* alcançando melhores resultados.

Dara e Banka (2014) propuseram o uso do BPSO (*Binary Particle Swarm Optimization* - Otimização por Enxame de Partículas Binário), um algoritmo que vem da

computação evolutiva, para selecionar as melhores características de base de dados de câncer, leucemia, cólon e linfoma. Após o uso da técnica, que ao fim do processo atribui valor 1 às características que se mantêm e 0 às que devem ser excluídas, foiobservado um aumento significativo na acurácia da classificação das bases de dados.

Diferentemente dos outros métodos de redução, o PCA (*Principal Component Analysis*) diminui a quantidade das características por meio de um transformação. Song*et al.* (2010) fizeram uso desse método para melhorar o processo de reconhecimento facial. Como resultado, foi visto que o uso da técnica de PCA produziu uma menor taxa de erro que a base original, com uma queda de 50% para 31,9%, e com uma reduçãode 2000 características para 600.

Uma outra forma de selecionar atributos é por meio de um filtro que utiliza algum atributo calculado das características para ser o comparativo da seleção. Pereira *et al.*(2015) empregaram o *Info Gain*, com adaptações: para lidar com dados multirrótulo diretamente; para fazer a seleção das características em onze base de dados; para comparar com e sem o uso de outras técnicas. Os experimentos indicaram que a técnica proposta foi bastante competitiva comparada às outras em relação aos resultado, e para base de dados muito grandes, ela alcançou resultados bem melhores.

2.5 Classificação via algoritmos

A aprendizagem de máquina supervisionada é realizada com o auxílio de di-ferentes tipos de algoritmos. Como resultado do procedimento há uma classificação, mensurada por métricas, para os objetos do problema. Atualmente, diversos modelosde predição fazem uso desses algoritmos, já clássicos, ou adaptados para inúmeras áreas do dia-a-dia.

Bijalwan *et al.* (2014) aplicaram o algoritmo KNN (*K-Nearest Neighbors*) para fazer a mineração de textos e documentos, para classificá-los de acordo com sua categoria, havendo cinco opções de categorias disponíveis. Os autores verificaram queo KNN obteve uma altíssima taxa de acurácia, com valores acima de 98%. E foi a maioracurácia entre outras técnicas de mineração testadas.

Yuan *et al.* (2010) utilizaram o algoritmo SVM (*Support Vector Machine*) para classificar o tráfico de internet em sete classes, para isso foram extraídas 19 características que discriminam o fluxo na internet. Os autores testaram quatro tipos de funções kernel no SVM, e obtiveram uma taxa de 97,1% na acurácia dessa classificação após a otimização do processo com a seleção das melhores características.

Uma aplicação bastante comum de predição com aprendizagem de máquina éno setor

bancário. Tsai *et al.* (2014) compararam classificadores clássicos e *ensembles*numa predição de falência. Para isso utilizaram árvores de decisão, redes neurais, SVM e classificadores combinados. As bases de dados utilizadas foram extraídas de bases crediárias da Alemanha, Japão e Austrália. Os autores tiveram melhores resultados vindos da classificação por meio do *boosting*, além de esse mesmo modelo ter sido ode menor custo computacional.

2.6 Termografia no diagnóstico do câncer de mama

Em 1956, com uma maior abertura cientifica pós-guerra, Lawson apurou que a temperatura da região do tumor mamário era mais alta que de outro tecidos do mesmocorpo sem o câncer (LAWSON, 1956). Essa revelação iniciou uma série de pesquisas na área termográfica em pessoas, principalmente ligados ao câncer de mama, que permanece até os dias de hoje (ASSIS, 2015). As pesquisas focadas na detecção precoce do câncer de mama, através da classificação de imagens termográficas, vêm desenvolvendo diferentes abordagens e técnicas diversas para um mais rápido, simples,generalista e assertivo processo de diagnóstico .

A metodologia do estudo proposto por Rodrigues *et al.* (2018) teve como objetivo avaliar a classificação, com e sem selecionadores bioinspirados de atributos, em quatroclasses de imagens termográficas. A base de dados utilizada continha 339 imagens ao todo, sendo subdivida em 66 sem lesão alguma, 73 com cistos, 121 com lesões benignas e 76 com lesões malignas. Foram extraídos 169 atributos com a utilização dos momentos Haralick e Zernike, que foram posteriormente reduzidos a 57 e 59 atributos após o emprego do AG (algoritmo genético) e PSO respectivamente. Os melhores resultados, que foram alcançados utilizando o classificador SVM, mostraramuma pequena redução na acurácia após a seleção dos atributos, de 90,6% a 86% e 85,6%, porém com um ganho de tempo e redução de dimensionalidade consideráveis.

Buscando demonstrar a viabilidade do uso das imagens termográficas na detecção do câncer de mama, Acharya *et al.* (2012) estudaram uma base de 50 imagens termográficas, coletadas do Departamento de Radiologia do Hospital Geral de Singapura, onde 25 delas eram de pacientes saudáveis e 25 de pacientes com o tumor maligno. Os 16 atributos extraídos foram características de textura, que utilizam a inten- sidade dos pixels e níveis de cinza para o cálculo. Entretanto, apenas 4 atributos foramusados na classificação final após a seleção pelo *o-value* menor que 0,001. Como resultado, utilizando o algoritmo SVM, o estudo obteve 85,7%, 90,4%, 81,0% e 88,1% de sensibilidade, especificidade, acurácia e AUC (*area under curve*) respectivamente.

Milosevic *et al.* (2014) utilizaram 40 imagens termográficas (sendo essas de26 pacientes saudáveis e 16 com tumor), e contou com a extração de 20 atributos de textura, como: entropia, contraste e dissimilaridade, para cada termograma. A classificação foi conduzida utilizando os algoritmos SVM, KNN e *Naive Bayes,* e pôde confirmar, após excelentes resultados, que os atributos extraídos são parâmetros úteis no diagnóstico via termografia. O algoritmo de classificação que se saiu melhor foi o KNN com 92,5% de acurácia, seguido do SVM com 85% e do KNN com 80%.

Baffa *et al.* (2016) desenvolveram uma técnica de segmentação de imagens termográficas que utiliza limiarização com refinamento adaptativo, o que torna possível a segmentação das pregas inframamárias. A base de dados usada para o estudo foi obtida de uma base pública que contém 283 imagens termográficas. Para alcançar osmelhores resultados foram introduzidos três parâmetros avaliados experimentalmente, tendo em vista que as pregas inframamárias identificadas podem variar dependendo das mamas consideradas. Após obter os valores para os três parâmetros que maximizam os resultados, e a utilização de atributos estatísticos na classificação, os autores obtiveram, em classificações binárias, 96%, 98% e 95% para acurácia, sensibilidade e especificidade respectivamente.

Araújo (2014) propôs em sua pesquisa um método para classificar imagens termográficas mamárias em 4 classes (cisto, normal, tumor maligno e benigno). No processo de segmentação da região de interesse, ele utilizou o próprio método desegmentação manual por elipse com auxílio da plataforma *Matlab*. Como diferencial, ele empregou um processo morfológico em cada imagem segmentada, e uma extraçãode características contínuas obtidas por meio de medidas intervalares. Para o processo, foi usada uma base de 50 pacientes e a classificação feita por meio da discriminação linear, distância mínima e janela de Parzen, com a técnica *leave-one-out*. O autorobteve como melhor resultado 84% para acurácia; 85,7% para sensibilidade e 86,5%para especificidade da Classe Maligno.

Madhu *et al.* (2016) analisaram 265 imagens termográficas (78 com câncer e 187 sem câncer) com intuito de prover um método que tivesse uma alta especificidade.Foi utilizada uma técnica para selecionar *hot spots* e *warm spots*, regiões da ROI que foram consideradas como mais importante para a extração de características. Emcada uma dessas regiões foram extraídas características como temperatura relativa, comparação de temperaturas de mamas contralaterais, quantidade e área das regiões anormais termicamente. O trabalho obteve resultados expressivos utilizando o algoritmo *Random Forest* com 100% de sensibilidade e 98,9% de especificidade.

3 FUNDAMENTAÇÃO TEÓRICA

Para o entendimento deste trabalho é necessário o conhecimento de conceitos relacionados ao câncer de mama, a métodos utilizados para detecção da doença, à termografia infravermelho e a técnicas computacionais. Estes conceitos foram compilados neste capítulo.

3.1 O câncer de mama

O câncer, como um todo, é definido como sendo um grupo doenças que com- partilha a característica de multiplicação celular exagerada e fora de controle. Esse processo patológico de proliferação anormal e descontrolada, conhecido por neoplasia,pode ocorrer com qualquer tipo de célula, e com isso originar um câncer (NOVA, 2017).

Além da particularidade de multiplicação intensa, o câncer também é identificado pelo acúmulo de mutações no genoma da célula. Essas alterações transformam uma célula normal em uma célula cancerígena, que não mais responderá os sinais que ditam o controle natural da comunidade celular (LELES *et al.*, 2015).

Dentre as diversas confirmações que existem sobre o câncer, a de maior impacto é que o mesmo é causado majoritariamente advindo de fatores externos que se relacionam com a mutação no DNA na célula (PRADO, 2014).

Entretanto, não há uma forma definitiva de extinguir as chances de se ter câncer, pois a etiologia do câncer ainda não é totalmente conhecida. Entretanto, segundo Jurberg *et al.* (2006), os fatores de risco já confirmados estão no dia-a-dia de quase todos os indivíduos, entre eles estão: álcool, tabaco, dieta rica em gordura e com poucas frutas e vegetais, sedentarismo e fatores genéticos (NOVA, 2017).

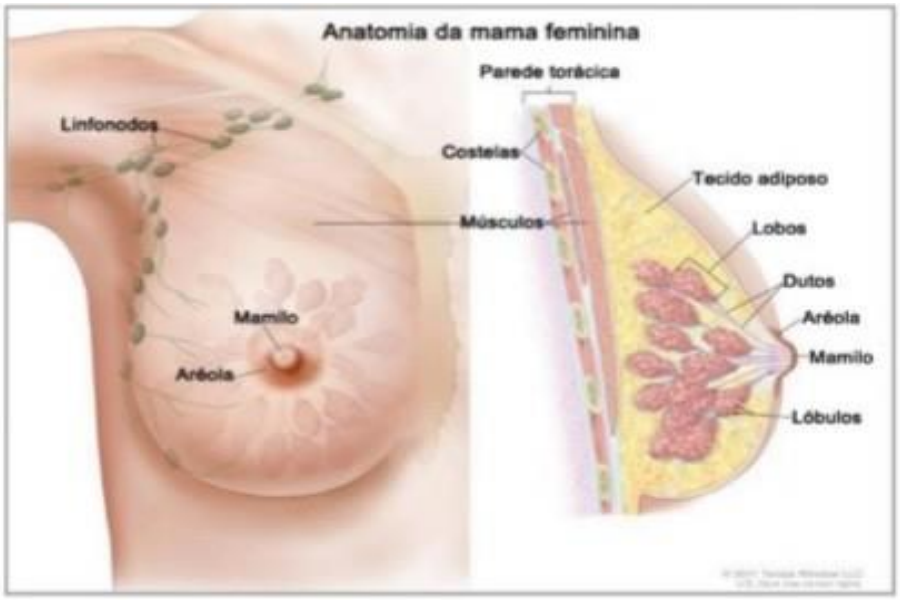
No caso específico do câncer de mama, o mesmo ocorre quando a região afetada pelo tumor é a estrutura mamária (BAFFA; LATTARI, 2018). Essa estrutura é formada pela glândula secretora de leite e o tecido que recobre a estrutura de fato, umórgão glandular cutâneo, situado sobre o tórax em número par, que cobre o músculo peitoral, e tem na extremidade do órgão a papila mamária, ou mamilo (ARAÚJO, 2009).

A mama é formada pelos tecidos adiposo, glandular e conjuntivo, e é compostapor lobos, dutos e pela estroma (SOUZA *et al.*, 2018). Há vasos sanguíneos e linfáticos,e os dutos que vão interligar os lobos, lóbulos e bulbos (BORCHARTT, 2013). Segue abaixo um resumo da função

de cada parte citada da mama e a ilustração da mesmana Figura 5.

- Lobos: Glândulas onde se produz o leite.
- Dutos: Regiões em formatos de tubo por onde o leite perpassa, tendo cada mama de 15 a 20 lobos, que abrangem pequenos lóbulos.
- Lóbulos: Conjunto de dezenas de bulbos.
- Estroma: Tecido adiposo e conjuntivo que circunda dutos, lobos, vasos sanguíneos e linfáticos, basicamente é o preenchimento final da mama.

Figura 5– Anatomia da Mama



Fonte: ACS (2020)

Além disso, entre 80% a 85% do casos de câncer de mama estão localizados nos dutos, 10% a 15% nos lóbulos e entre 5% a 10% se encontram na região do estroma (BAFFA; LATTARI, 2018). Por uma questão de nomenclatura, muitas pessoasse referem ao câncer com o nome dado ao local que o tumor se iniciou, sendo o carcinoma ductal, tendo inicio nos ductos e o carcinoma lobular, tendo origem nos lobos (BEZERRA *et al.*, 2007).

Existem diversos tipos de câncer de mama, pois cada tumor tem suas próprias individualidades. Entretanto, na maior parte dos casos é possível detectar o problemapor meio de alguns sintomas padrão:

- Nódulo, geralmente duro, irregular e indolor, sendo esse o mais importante sinaldo

câncer, estando presente em 90% dos casos.

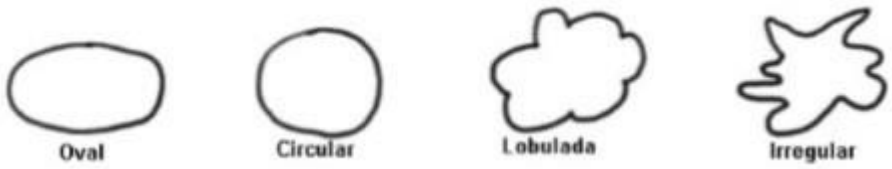
- Alterações no mamilo.
- Pele da mama avermelhada ou retraída.
- Pequenos nódulos nas axilas ou pescoço.

Os nódulos diferem uns dos outros por aspectos físicos, como a sua borda, seu tamanho, isso também é um fator de ajuda para diferir se é um nódulo maligno ou benigno (BEZERRA *et al.*, 2007). Nas Figuras 6 e 7 é possível ver claramente adiferença das massas de acordo com a borda e forma.

Figura 6– Diferenças entre massas de acordo com a borda

Fonte: Nunes *et al.* (2009).

Figura 7– Diferenças entre massas de acordo com a forma



Fonte: Nunes *et al.* (2009).

Logo, é extremamente importante as mulheres (por serem 99% dos casos de câncer de mama) sempre que possível se atentarem para o aparecimento dessas alterações, para que profissionais de saúde possam avaliar a região e diagnosticar precocemente o câncer de mama (INCA, 2020).

3.1.1 Proliferação do Tumor Maligno

É importante destacar as principais diferenças entre células saudáveis e cancerosas, e suas consequências no funcionamento do corpo. As células que sofreram a mutação, não tem controle alguma sobre a divisão celular, isto é, elas não respondemaos estímulos naturais que

fazem ocorrer a mitose, e após diversas divisões conseguemse aglomerar, formando grandes massas de células, os tumores.

Aparentemente pequenos à primeira vista, um tumor já pode conter milhões de células mutadas quando captado pelo médico ou pelo Raio X (PRADO, 2014). Entretanto, os tumores não são sempre iguais e com o mesmo efeito sob o corpohumano, por isso eles são divididos em duas classes: Benignos e Malignos.

O tumor benigno se mantém estacionado no tecido ao qual se desenvolveu,e possui uma taxa de crescimento bastante lenta. Esse tipo de tumor não deve ser classificado como um câncer, contudo, podem e devem ser removidos se afetarem algum órgão primordial (PRADO, 2014). O tumor maligno diferentemente do benigno,não se assemelha com o tecido que o deu origem, possuindo formas diferentes, geralmente com disposições irregulares. Ele tem sua multiplicação celular desenfreada,e por isso, é denominado câncer.

Na Figura 8 é possível observar os critérios que diferenciam tumores benignos e malignos (ARAÚJO, 2009).

Figura 8– Diferenças entre Benigno e Maligno.

CRITÉRIOS	BENIGNOS	MALIGNOS
Cápsula	Presença freqüente	Geralmente ausente
Crescimento	Lento, expansivo e bem delimitado	Rápido, infiltrativo com delimitação imprecisa
Morfologia	Reproduz o aspecto do tecido de origem	Características diferentes do tecido de origem
Mitoses	Raras e típicas	Frequentes e atípicas
Antigenicidade	Ausente	Presente (geralmente fraca)
Metástases	Não ocorrem	Frequentes

Fonte: Ministério da Saúde, Programa de Oncologia (2010)

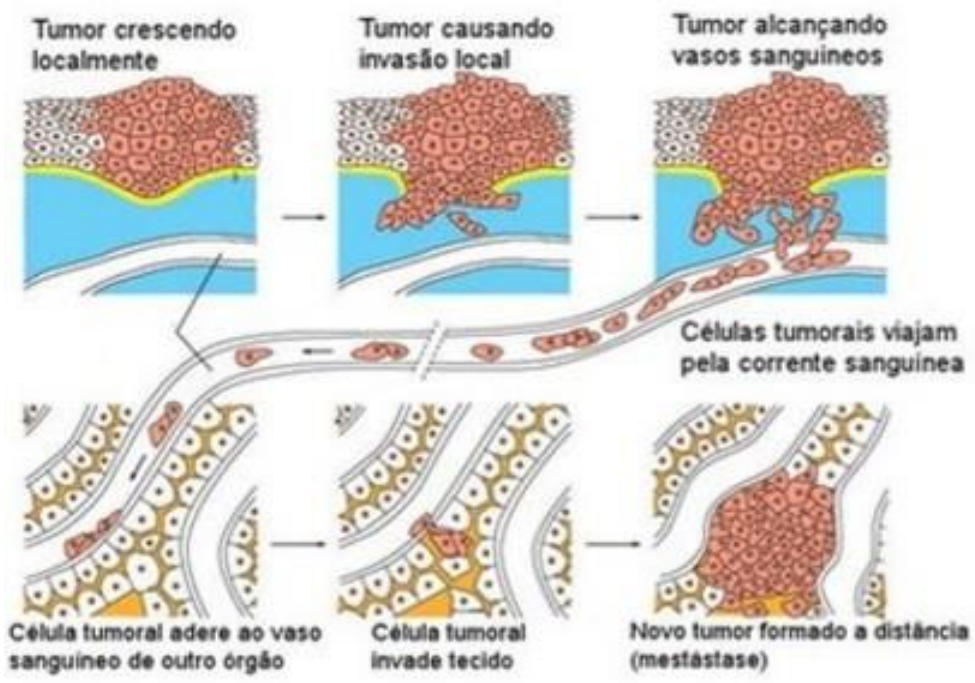
Mitose: A mitose é identificada como sendo o processo de divisão celular. Consequentemente, quanto mais vezes esse processo ocorrer, maior propensão de o tumor se espalhar. Nos tumores benignos, como a taxa de multiplicação é baixíssima,as mitoses não são muito comuns, sua divisão celular é igual a de partes normais do corpo. Nos tumores malignos, as mitoses ocorrem diversas vezes, sendo algumas delas de forma não-convencional (SOUZA *et al.*, 2018).

Outra característica das células cancerosas, e considerada a pior de todas, é a possibilidade que elas tem de se infiltrar em outros tecidos do corpo, atacandoas vizinhanças da região inicial. Essa característica é denominada de metástase. A metátese pode ser definida como "crescimento tumoral maligno secundário à umadisseminação de um foco tumoral primário, situada a distância" FILHO (1976). Existe algumas etapas para que as células infectadas consigam se espalhar para outros locais.Pelo fato de não apresentar encapsulação, esse tipo de célula pode se soltar do foco tumoral de origem. Com a ajuda da secreção de enzimas de digestão elas se estendem,fazendo com que as regiões próximas sejam destruídas, facilitando o caminho até um vaso sanguíneo.

Mesmo que o caminho até um vaso seja de baixo índice de sobrevivência, estima-se que 1 entre 10 mil células chegam ilesas. Se apenas uma conseguir se adentrar no vaso, ela será capaz de invadir esse novo tecido, e então, por meio desinais químicos haverá o aumento dos vasos sanguíneos, que irá manter aquele tumorfuncionando, e preparando para mais um ciclo de infestação.

A Figura 9 detalha o processo de metástase.

Figura 9– Processo de Metástase.



Fonte: Neto *et al.* (2017)

3.1.2 Diagnóstico precoce

Muito casos de câncer só são descobertos quando já se encontram em fases

avançadas, por vezes com metástase, o que torna as probabilidades de melhora beirando o impossível (PRADO, 2014). Praticamente todos os problemas na ciênciatêm mais chance de serem resolvidos quando descobertos cedo, essa mesma teoria serve para o câncer, quanto mais precoce ele é detectado, maior a chance de cura e de um tratamento mais eficaz.

Essa detecção busca identificar o tumor quando ele ainda está no tecido de origem, ou até mesmo lesões que indicam a futura malignidade que virá a se formar (LELES *et al.*, 2015). Diferente do que pensa grande parte da população, o câncer de mama, que não tem associação com fator hereditário, corresponde a 90% dos casos no mundo todo.

As pesquisas já feitas, apontam que no caso do câncer de mama os grandes responsáveis estão ligados a síntese de esteroides sexuais, isso pode ser notado no corpo nos casos de: menarca precoce, menopausa tardia, gestação e uso de estrógenos exógenos (LELES *et al.*, 2015).

O tumor na mama possui um tempo de duplicação celular de quatro mesesem média, em razão disso, muitas pessoas quando detectam algo diferente no corpo acham pequeno, ou de fraco crescimento. Após observar alguns dias, e quando tem dimensão visível do problema, as células modificadas já se multiplicaram várias vezes(SOUZA *et al.*, 2018). Em vista disso, o autoexame de mama é essencial para que se conheça o próprio corpo, identificando quando ele dá um sinal de alerta nessa região (INCA, 2020).

3.1.3 Mamografia

O Brasil segue a recomendação da OMS e estimula a detecção feita por rastreamento mamográfico. A mamografia é um tipo de radiografia feito na região dasmamas com auxílio de ferramentas de Raios X, que conseguem reconhecer a partirde imagens, suspeitas do câncer antes do aparecimento dos sintomas iniciais (INCA, 2020).

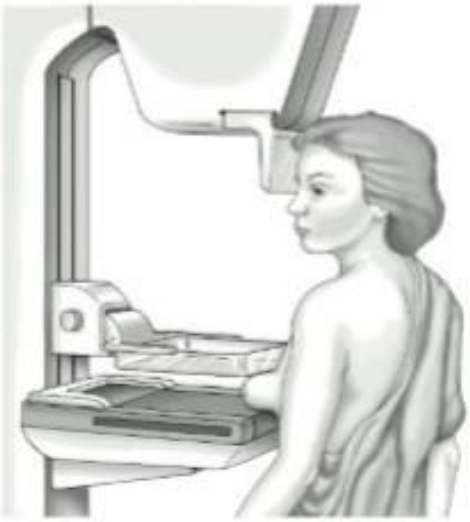
A mamografia tem sido de grande serventia na redução da taxa de mortalidadedo câncer de mama com a descoberta inicial do mesmo, possibilitando o tratamento com antecedência. Entretanto, há também alguns riscos ligados a esse exame, como por exemplo:

- Suspeita de câncer sem confirmação, gerando estresse para a paciente
- Exposição aos Raios X, radiação ionizante, que podem gerar um novo câncer(INCA, 2020).
- Pelo fato de comprimir a mama para adquirir imagens melhores para avaliação existe também um desconforto, mesmo que suportável, nesse movimento (LELES *et al.*, 2015)

Outros fatores que interferem a acurácia da mamografia são a dependência da habilidade de o radiologista de interpretar a imagem, e a densidade das mamas. Dado que mulheres jovens tem a mama mais densa e 35% delas mantêm a mama densa pós menopausa. Essa densidade é formada pelo tecido glandular, que torna alguns nódulos não identificáveis no exame mamográfico, sendo necessário outros examesadjuntos para confirmação (LELES *et al.*, 2015).

Nas Figuras 10 e 11 estão ilustrados respectivamente o procedimento da mamografia e uma imagem mamográfica.

Figura 10– Exame de mamografia



Fonte: ACS (2020).

Figura 11– Imagem mamográfica

Fonte: Sampaio *et al.* (2011).

3.2 Termografia

3.2.1 Radiação Térmica

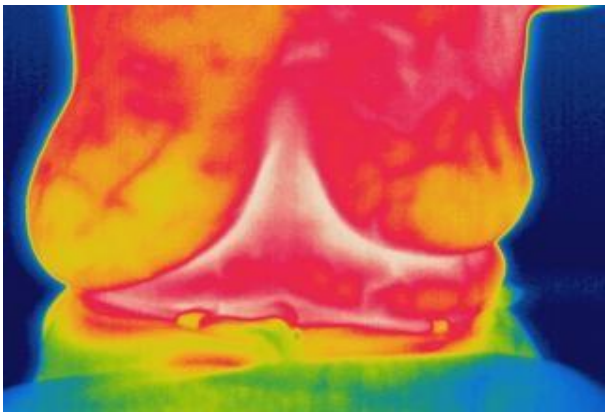
Um corpo negro pode ser definido como um objeto que irá absorver toda equalquer radiação que incide sobre ele. Por essa razão ele é chamado de absorvedor e radiador perfeito, e é usado em comparação com outros corpos. O corpo humano, tal como os outros tipos de corpos, irá emitir apenas uma parte da energia que o corpo negro emitiria em condições idênticas. Com o uso do corpo negro como ponto de partida para construir essa relação, algumas variáveis são utilizadas para medir bem a influência na TIR, entre elas estão: emissividade, transmissividade e refletividade. Todas do próprio corpo analisado, como também de outros corpos no mesmo ambiente e da atmosfera (CASTANEDO, 2005).

3.2.2 Termografia

A utilização da termografia infravermelha vem tendo destaque nos últimos anosem diferentes áreas. Esse método é baseado na visualização do calor que é irradiado de qualquer corpo que esteja acima de 0 K (-273 °C), calor esse que não é visívelpor estar em um espectro eletromagnético que a visão humana não tem capacidade de observação. O uso da técnica é visto em indústrias, construção civil, aviação e na área médica. O uso em humanos é bastante importante por trazer respostas que tem ligação com síndromes, deficiências vasculares, neurológicas e auxílio a diagnósticos relacionados ao câncer (MARINS *et al.*, 2015).

O uso da termografia na área do câncer de mama é feito com o auxílio de umacâmera térmica sensível que irá captar a radiação térmica e quantifica-la em valores digitais. O software para a visualização dos dados obtidos das imagens pode ser apresentado em escala de cinza ou em pseudo cores (VASCONCELOS *et al.*, 2017). Quanto maior é a diferença de temperaturas captadas, o resultado se tornará mais relevante para estudo (GONÇALVES *et al.*, 2017). Porquanto, todo o procedimento é feito sabendo que áreas afetadas pelo tumor cancerígeno e áreas adjacentes terão uma maior temperatura, devido a uma maior vascularização sanguínea que uma regiãonormal (GOLESTANI *et al.*, 2014). A variação térmica entre os dois lados de um corpo normal não deve ser maior que 0,5°C, e entre as duas mamas maior que 0,14°C (RESMINI, 2016). A Figura 12 mostra a diferença entre uma mama saudável e outra com câncer.

Figura 12– Exemplo termográfico de mama saudável e mama com câncer



Fonte: Próprio Autor

Os grandes diferenciais da TIR em relação a outros exames que visam detectar o câncer de mama, principalmente a mamografia, é o fato de ele ser não inva-sivo, não haver contato algum com o paciente e não emitir radiação (RASTGHALAM; POURGHASSEM, 2016).

Dependendo do tamanho do tumor, alguns autores afirmam que a termografia pode detectar o câncer de mama cerca de 10 anos antes que a mamografia (GOLESTANI *et al.*, 2014). Isso se torna possível devido à neo-angiogênese causada pelo tumor na região, ocasionando um aumento de temperatura com potencial de ser reconhecidopela técnica da TIR, que pode verificar diferenças de até 0,025 °C (MAHMOUDZADEH *et al.*, 2015).

Atualmente, devido ao grande número de casos de câncer de mama, o investimento em tecnologia nessa área vem aumentando, e as câmeras que antes eram de baixa resolução, hoje podem chegar a ter alta sensibilidades (ETEHADTAVAKOL *etal.*, 2013). Como exemplo, os sensores de nível médio tem sensibilidade na ordem de0,1°C, com os mais avançados chegando a 0,025°C (SANCHES, 2009). Uma câmera infravermelha está ilustrada na Figura 13.

3.2.3 Parâmetros de normalidade do corpo humano

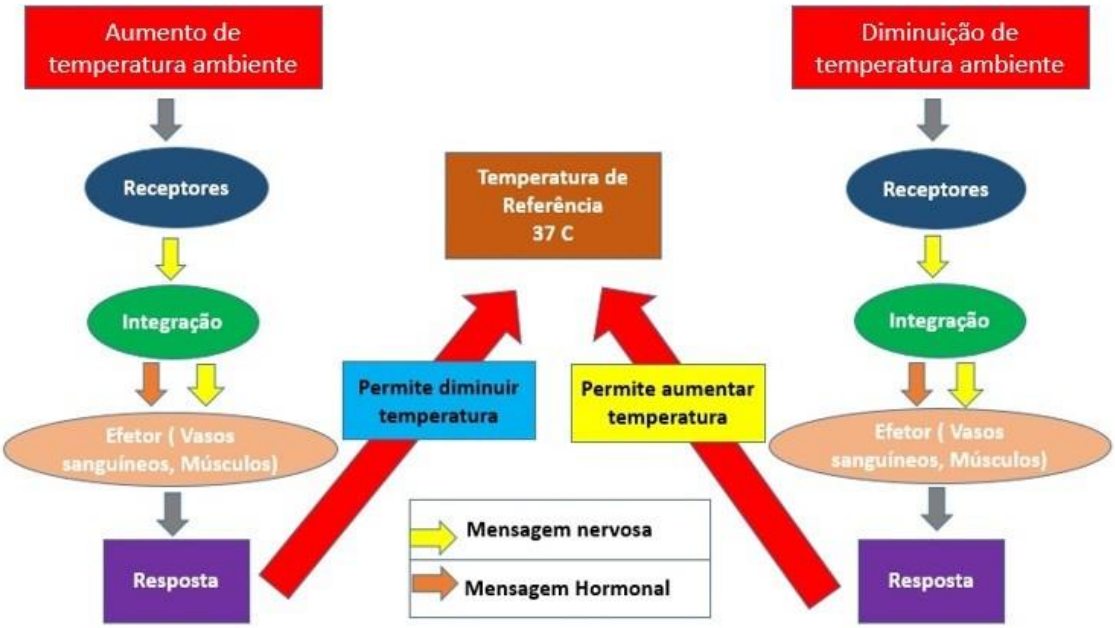
O corpo humano necessita que sua temperatura se mantenha quase que cons-tante para que todos os processos aconteçam de forma normal. A permutação de calorcom o ambiente externo é a forma que o mesmo utiliza para permanecer com sua temperatura ideal, esse recurso é conhecido por termorregulação (BRAZ, 2005). Essa temperatura média natural é de 37°C, ao passo que a temperatura da superfície do corpo se mantém na faixa dos 33°C, com homens e mulheres apresentando o mesmo controle de temperatura corporal, algo que muda em pessoas doentes e mais idosas (CHUDECKA; LUBKOWSKA, 2012). A Figura 14 mostra como é o

funcionamento de regulação de temperatura do corpo humano.

Figura 13– Câmera termográfica

Fonte: FLIR (2020)

Figura 14– Mecanismo de auto regulação do corpo humano



Fonte: Adaptado de Assis (2015).

O mecanismo de termorregulação mantém o corpo humano, a parte interna,com uma faixa de apenas 0,7°C para mais ou menos, tendo como referencia os 37°C.Conforme existam estímulos que possam alterar essa temperatura, como doenças ou mudanças climáticas externas, esse processo irá moderar o fluxo sanguíneo de microcirculação cutânea para balancear as causas fora da normalidade (LELES *et al.*,2015).

A temperatura da superfície, diferentemente da interna, tem uma maior variação ao longo do corpo, se assemelhando a um mapa de isotermas. Essa parte do corpo sofre um maior efeito, tanto de fatores externos, como de fatores internos. Desse modo, uma parte mais externa tende a ter uma temperatura média menor que as partes internas. (CHUDECKA; LUBKOWSKA, 2015)

Há mais de um efeito que causa essas diferenças entre temperaturas internas eexternas. Fatores como as propriedades térmicas dos tecidos que estão separando os órgãos internos da superfícies acima do órgão, fluxo sanguíneo, umidade da pele no momento, além da temperatura do meio ambiente (CHUDECKA; LUBKOWSKA, 2012).

3.2.4 Protocolo de captura de imagens termográficas

O protocolo adequado para captura das imagens termográficas permitirá garantir a exatidão das medidas e a repetibilidade do experimento. A radiação infravermelha emitida por um corpo está vinculada com as condições externas no momento, como por exemplo: umidade do local, fluxo de ar e temperatura do ambiente. Qualquer pequenavariação desses elementos irá alterar os valores reais do resultado (LAHIRI *et al.*, 2012).

A sala do exame deve ser espaçosa para que haja espaço suficiente para opaciente, o técnico, os equipamentos e eles possam se movimentar nela. A sala não deve ter portas e janelas abertas. Havendo alguma, essa deve ser coberta.

Durante o tempo do exame o local não deve ter mudança na temperatura, e ao mesmo tempo a temperatura deve ser agradável para que o paciente não transpire nem sinta muito frio (COCKBURN, 2020). No geral esse procedimento é feito com as salas estando de 19 a 23 °C e umidade de 40% a 75%. Dependendo do clima do países onde será feito o exame, essa temperatura ideal da sala pode ficar entre 25 a 28°C (ARAÚJO, 2009).

Recomendações técnicas como: afastamento de equipamentos de computaçãoda área de leitura, luzes desligadas na hora da captura das imagens e cobrir as superfície refletoras devem ser seguidas para uma melhor confiabilidade (ARAÚJO,2009).

Além das condições do ambiente, existem alguns protocolos que o paciente deve seguir para que o corpo do mesmo esteja funcionando em condições padrão no momento, dado que a superfície da pele é extremamente afetada pela transpiração. Esses protocolos devem ser passados ao paciente pelo menos 24 horas antes do exame, entre eles estão (COCKBURN, 2020) :

- As áreas que serão analisadas não podem ser depiladas no dia do exame.

- Nada que contenha cafeína ou álcool pode ser ingerido até 24 horas antes do exame.
- Não fazer exercícios físicos até 4 horas antes do exame.
- Evitar roupas muito apertadas.

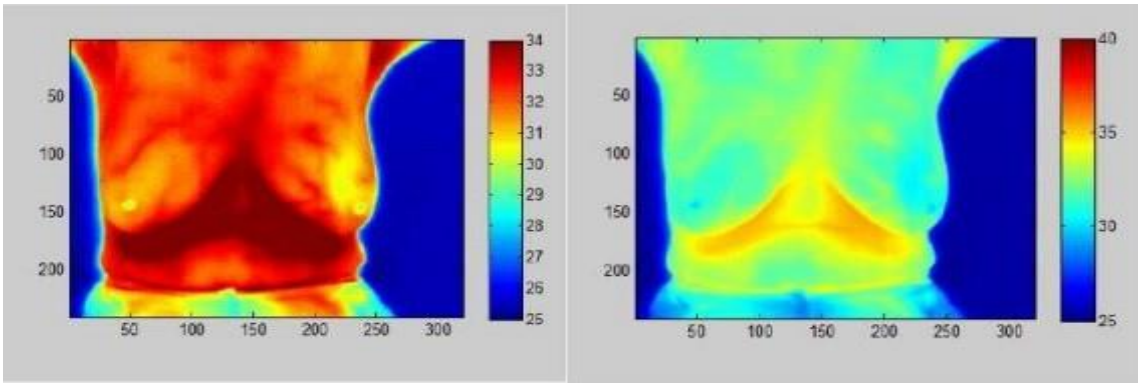
Nos 5 minutos prévios ao inicio do procedimento propriamente dito, a paciente irá pôr as mãos sob a cabeça para favorecer um imageamento mais objetivo para as mamas (BRONZINO, 2006).

3.3 Segmentação

As imagens termográficas podem ser retratadas por uma função bidimensional f(x,y) relacionada a uma matriz de temperatura T(x,y), onde x e y se referem as coordenadas espaciais no plano, e T(x,y) à temperatura de determinado ponto (x,y) (ARAÚJO, 2014).

A imagem termográfica apresentada nas análises, conhecida como *pseudo color*, fará uma associação de cada ponto de temperatura da matriz a uma cor oriundade um mapa de cor. As cores dos *pixels* da imagem serão definidas de acordo com os limites superior e inferior impostos previamente, bem como o tipo de mapa de cor escolhido. Uma vantagem de se trabalhar com imagens digitais é que, embora existamdiferentes mapas de cor e seja possível mudanças na imagem de acordo com a escalaescolhida, a matriz de temperaturas permanecerá constante (QUEIROZ *et al.*, 2016). Na Figura 15 é possível observar a mesma imagem termográfica com duas diferentes escalas de temperatura.

Figura 15– Imagem termográfica vista com duas diferentes escalas de temperatura. a)25-34 ºC; b) 25-40 ºC



Fonte: Queiroz et al. (2014).

A segmentação propriamente dita tem como objetivo selecionar e separar as regiões de

interesse (ROI) do restante da imagem. Essas ROI são as regiões ondeestão concentradas as informações úteis para análise e que servirão para aplicar osmétodos e reconhecimentos de padrões (DOUKAS; MAGLOGIANNIS, 2007).

Muitas aplicações de segmentação são feitas na área médica para selecionar exclusivamente a região que pode vir a ter uma anomalia. Como exemplo, na busca por padrões de temperatura para o tumor maligno do câncer de mama, o ideal é umaseleção apenas do contorno da mama até um pouco abaixo da clavícula.

Ao longo dos anos, diversos algoritmos foram desenvolvidos para a segmentação de imagens, eles podem ser divididos em três tipos: manuais, automáticos esemiautomáticos. Os do tipo automático, são mais rápidos e com uma função pré-definida de corte da imagem, sem intervenção humana (GAO *et al.*, 2010). Enquanto que os manuais são indicados para especialistas, pois o mesmo fará a segmentação da ROI baseado na visão e experiência prévia. Porém, esse método ainda assim abre margem para o erro humano. Contudo, apesar das vantagens da opção automática, os métodos semiautomáticos vêm sendo bastante utilizados para suprir os possíveis erros que podem aparecer no modelo automático devido à diferença de corpos entre humanos e à assimetria das mamas de uma mesma pessoa.

3.4 Extração de Características

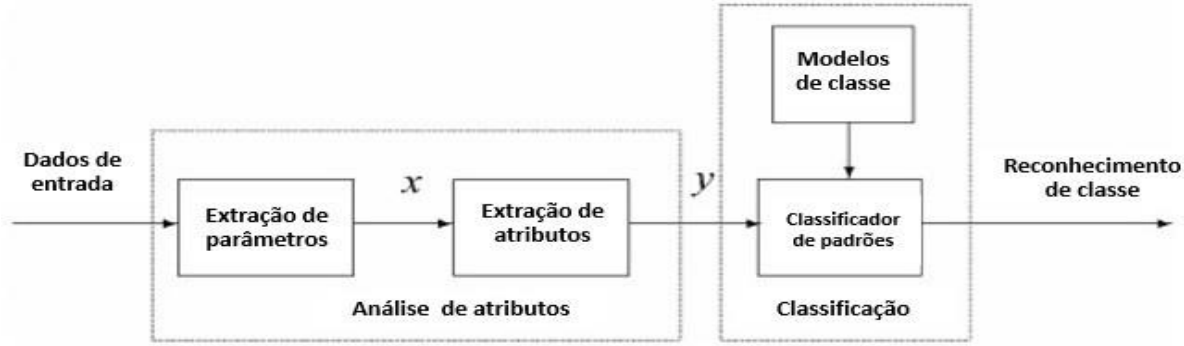
A classificação via aprendizagem de máquina funciona a partir da obtençãode dados sobre o que se deseja classificar. Esses dados são comumente chamadosde características, pois eles que irão representar determinado objeto de estudo. Ascaracterísticas podem ser captadas em dados de sons, imagens, sinais e outras formas,e geralmente são quantificadas utilizando dados estatísticos, histogramas e intervalos (DUDA *et al.*, 2012).

As características devem ser as mais discriminante possíveis, isto é, elas de-vem conter informação que caracterize e diferencie objetos para que no processo de aprendizagem de padrões as classes possam ser facilmente definidas. Deve-se evitar trabalhar com uma alta dimensão, alto número de características, logo, as escolhidas devem contar informação suficiente e relevante para o bom andamento da classificação (WANG; PALIWAL, 2003).

Segundo Dougherty (2009), em geral, as melhores características são aquelas: robustas discriminantes, confiáveis e independentes. Tais características devem ser invariantes sob translação, orientação, escala, iluminação, e os valores de objetos de classes diferentes não devem ser sobrepostos. Valores de objetos de mesma classe devem ser próximos, e as características não devem ser correlacionadas para não haver redundância.

Na Figura 16 é apresentado um esquema simples de reconhecimento de padrões a partir da aquisição de dados e do uso das características extraídas.

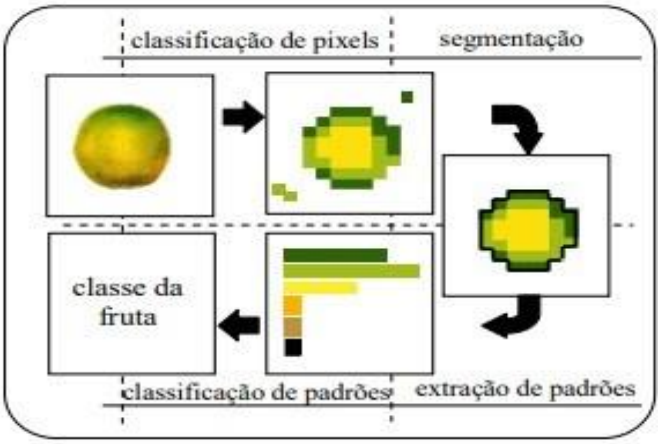
Figura 16– Padrão de reconhecimento convencional



Fonte: Adaptado de Wang e Paliwal (2003).

Na Figura 17 é possível observar um exemplo da extração de padrões de uma imagem digital para classificar o tipo de uma laranja.

Figura 17– Aquisição de características de imagem digital



Fonte: Simões e Costa (2003).

3.5 Redução de Dimensionalidade

Uma das grandes dificuldades na aprendizagem de máquina é sua alta complexidade na criação de regras para predição, a partir das características extraídas dosobjetos das classes em estudo. Além disso, predições com alto grau de dimensionalidade, muitas características, tendem a ter um alto custo computacional, e ser uma "caixa preta"para interpretação, dificultando otimizações futuras (COVÕES, 2010).

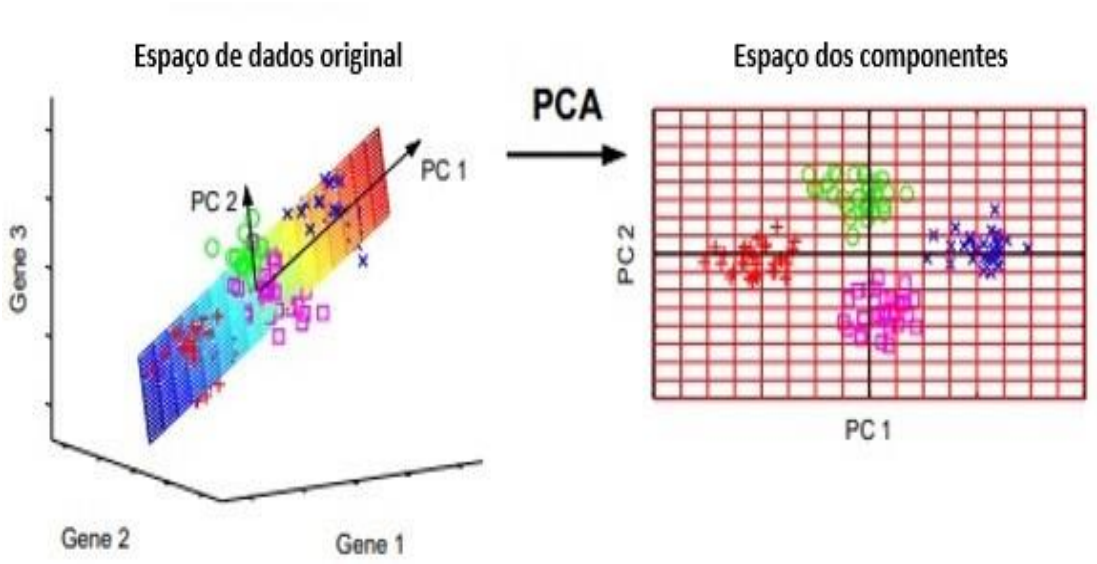
Por esses motivos, é costumeiro aplicar métodos de redução de dimensionalidade para

alcançar uma melhora geral nos resultados. Esses métodos podem ser divididos em seleção de características (ou atributos), onde o foco é reduzir o número de atributos já extraídos inicialmente, ou transformação de atributos (TA), onde por meio das características já extraídas são feitas transformações, ortogonais por exemplo, quevisam criar novas características a partir das previamente correlacionadas (ALMEIDA *et al.*, 2018).

3.5.1 PCA

O PCA (*Principal Component Analysis*) é um método matemático bastante conhecido de TA que tem como propósito a retenção da maioria da variância da basede dados no espaço das características. Em suma, a aplicação é feita através da identificação das direções, ao longo das quais a variância é máxima, conhecidas como componentes principais (RINGNÉR, 2008). Na Figura 18 é possível observar um exemplo com 3 dimensões (3 atributos), após a aplicação do PCA as classes dos dados ficaram mais bem separáveis com os dois novos atributos gerados.

Figura 18– Redução de atributos utilizando PCA



Fonte: Adaptado de Scholz (2006).

Um sequência objetiva do passo-a-passo da realização do PCA é listada abaixo:

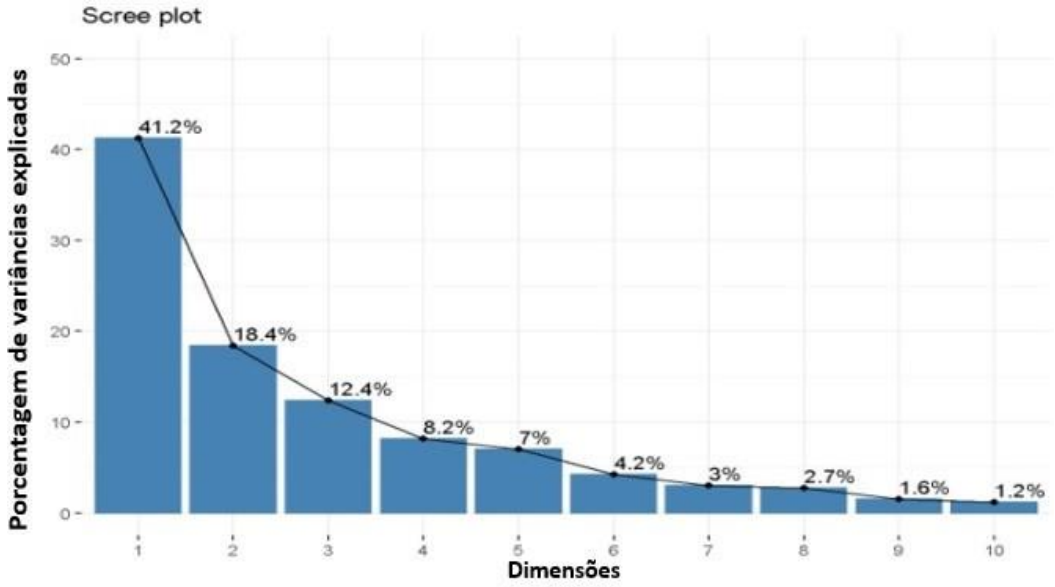
- Passo 1: selecionar os dados da base a ser otimizada;
- Passo 2: subtrair a média de cada valor (para cada dimensão);
- Passo 3: calcular a matriz de covariância;
- Passo 4: calcular os autovalores e autovetores da matriz de covariância;

- Passo 5: escolher os componentes e criação do vetor de característica

No Passo 5, após já haver calculado um autovalor para cada autovetor, deverá ser definido quantos componentes principais serão utilizados para a classificação. Quanto maior o autovalor maior será a importância do componente principal dele para o conjunto (SMITH, 2002).

Uma forma de decidir a quantidade total de componentes principais é pela visu-alização no *scree plot*. Um exemplo pode ser visto na Figura 19. No *scree plot* vemos a contribuição, em percentual, que cada componente principal carrega da variância do problema. Logo, é necessário avaliar se a retirada de algumas componentes principais, para diminuição da dimensionalidade, será mais interessante que a perda de variância atrelada à elas (HOLLAND, 2008).

Figura 19– Exemplo de um scree plot com 10 dimensões



Fonte: Adaptado de Kassambara (2017).

3.5.2 PSO

A PSO (Particle Swarm Optimization) pode ser definida como um algoritmo estocástico de otimização que utiliza a inteligência evolutiva como base para a busca do melhor resultado. Esse algoritmo foi proposto por Kennedy e Eberhart (1995), tendo se desenvolvido após eles observaram o comportamento social de enxames na busca por comida. Os animais presentes nesses enxames aprendem tanto com a própria experiência como com a experiência adquirida pelo enxame como um todo, e com isso conseguem encontrar melhores maneiras e

caminhos para chegar na comida (WANG et al., 2018).

Depois de pesquisas na área de comportamento social de enxames e sistemas de cooperação artificial, Bergh (2001) propôs cinco princípios básicos para nortear o desenvolvimento de algoritmos evolutivos baseados em cooperação.

- Proximidade: O enxame deve ser capaz de realizar tarefas simples de cálculos de espaço e tempo.
- Qualidade: O enxame deve ser capaz de sentir a qualidade de mudança no ambiente e responder a isso.
- Diversidade: O enxame não deve limitar seu caminho para obter recursos em um escopo estreito.
- Estabilidade: O enxame não deve mudar seu modo de comportamento com cada mudança ambiental.
- Adaptatividade: O enxame deve mudar seu modo de comportamento quando essa mudança vale a pena.

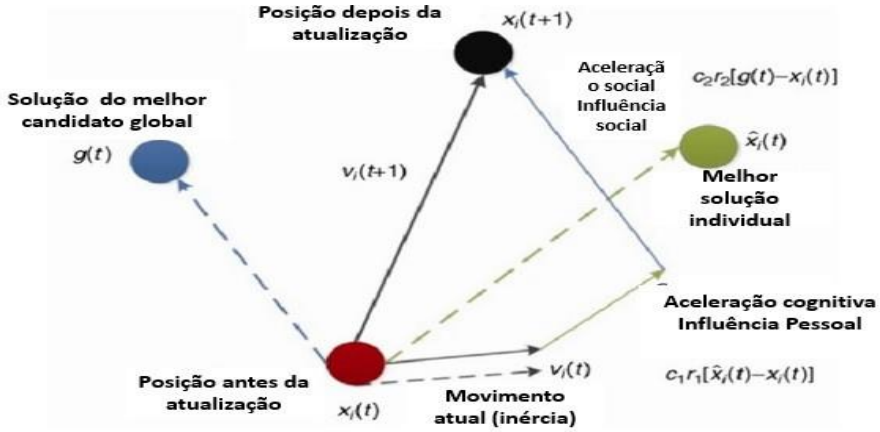
É importante observar que o algoritmo estará sempre atualizando, a cada iteração, os vetores velocidade (v_{id}) e posição (x_{id}) em busca da solução ótima, que éa melhor solução global, representada pelo *gbest* (p_{gd}), e *pbest* (p_{id}), melhor posição individual (CERVANTE *et al.*, 2012).

Na atualização do vetor velocidade (v_{id}) é empregada a Equação (3.1). Nelaaparecem as constantes W, c_1 e c_2 que são respectivamente o peso de inércia e asconstantes de aceleração do indivíduo e global. Grandes valores de W levam a uma busca global e pequenos valores facilitam buscas locais. Os termos $rand_1$ e $rand_2$ sereferem a valores randômicos uniformemente distribuídos entre 0 e 1 (CHAVAN *et al.*,2015).

$$v_{id}(k+1) = Wv_{id}(k) + c_1 rand_1(p_{id} - x_{id}) + c_2 rand_2(p_{id} - x_{id}) \quad (3.1)$$

Na Figura 20 é possível observar com mais facilidade, no esquema vetorial do PSO, como as constantes de aceleração social e individual influenciam na atualização do movimento das partículas a cada atualização.

Figura 20– Esquema vetorial do PSO



Fonte: Adaptado de Medium (2020)

Quando a intenção é utilizar o PSO para seleção de características, são feitaspequenas modificações no algoritmo e ele passa a ser chamado de BPSO. Nessa modificação é incluída uma função objetivo, que contém a taxa de acerto de um classificador selecionado previamente. Essa função será a responsável por fazer abusca da melhor combinação de características. O resultado final das posições das dimensões é dado na forma binária (0 ou 1), Equação (3.2), onde 0 significa que aquela característica foi excluída e 1 que ela deve permanecer. A modificação para um PSO discreto é feita através de uma função sigmoide, Equação (3.3), que transforma ovetor velocidade em um vetor probabilidade. Em seguida é aplicado na equação deatualização da posição, onde δ é um número randômico aleatório entre 0 e 1. (UNLER;MURAT, 2010).

$$x_{ij}^{t} = \begin{cases} 1 \; se \; \delta < s_{ij}^{t} \\ 0 \; caso \; contrário \end{cases} \tag{3.2}$$

$$s_{ij}^{t} = \frac{1}{1+e^{-v_{ij}^{t}}} \tag{3.3}$$

3.5.3 Ranqueamento

A seleção de atributos via ranqueamento é feita com a utilização de métricas pré-estabelecidas que irão qualificar determinado atributo. O subconjunto de atributospode ser escolhido por um número fixo dos melhores avaliados ou um limiar de seleçãopara os valores das métricas. As métricas mais comuns nesse tipo de seleção sãoas que avaliam o grau de associação do atributo com a classe, como por exemplo o *Information Gain* (IG), *Gain Ratio* e *Gini Index* (ALMEIDA *et al.*, 2018).

- *Information Gain*: É uma métrica usada para seleção de atributos que utiliza o conceito da entropia. Ela é fortemente difundida por ser de fácil computação e interpretação. O IG de um atributo X dada uma classe Y é calculada com auxílio da Equação (3.4). Onde H(X), Equação (3.5), é a entropia de dado atributo X, e H(X|Y), Equação (3.6), é a entropia de X após observar Y. Quanto mais informações um atributo X tiver em relação a uma classe Y, menor será a entropia condicionada H(X|Y) (PORKODI, 2014).

$$IG(X,Y) = H(X) - H(X|Y) \tag{3.4}$$
$$H(X) = -\sum_i P(x_i)\, log_2\big(P(x_i)\big) \tag{3.5}$$
$$H(X) = -\sum_i P_{yj} \sum P\big(x_i|y_j\big)\, log_2\left(P\big(x_i|y_j\big)\right) \tag{3.6}$$

- *Gain Ratio*: É uma métrica não-simétrica que foi elaborada para compensar o enviesamento gerado no cálculo do *Information Gain*. Isso ocorre pois, por vezes,o IG superestima a qualidade dos atributos com uma grande quantidade de valores. Logo, o cálculo do *Gain Ratio*, Equação (3.7), normaliza o IG dividindo-opela entropia do atributo estudado. Com valores entre 0 e 1, a métrica no seu valor máximo indica que o conhecimento do atributo X prediz completamente a classe Y. E com o valor mínimo, indica que não existe relação entre o atributo X e a classe Y (PORKODI, 2014).

$$GR = \frac{IG}{H(X)} \tag{3.7}$$

- *Gini Index* : Pode ser definida como uma medida estatística de impureza, e geralmente é aplicada em ordenamentos, sistemas binários e valores numéricos contínuos. Essa métrica tem como ideia principal valorar os atributos de acordo com as partições das instâncias. Quanto mais próximo de 0 (valor mínimo) é calculado o índice de determinado atributo, Equação (3.7), melhor será a sua classificação. Isso ocorre pelo fato de todas as instâncias pertencerem à mesma classe e a obtenção de informações úteis ser máxima. No caso contrário, quandoas instâncias estão igualmente distribuídas pela quantidade de classes, a obtenção de informações úteis passa a ser mínima, e o índice tem seu valor máximo. Na Equação (3.8), P_i é a probabilidade de qualquer instância pertencer à classe C_i ($1 \leqslant i \leqslant m$) (SHANG *et al.*, 2007).

$$Gini(S) = 1 - \sum_{i=1}^{m} P_i^2 \tag{3.8}$$

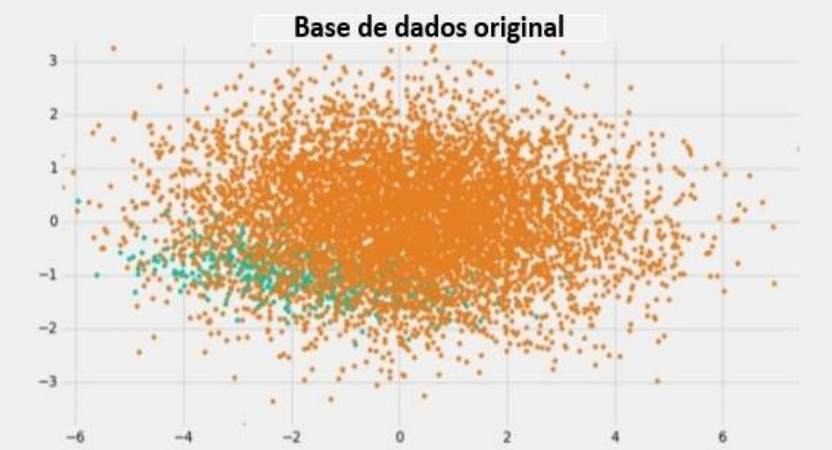
3.6 Desbalanceamento e Técnicas de reparo

O problema do desbalanceamento de amostras a serem classificadas é considerado recente, e teve seu início após grandes aplicações em negócios, indústrias e medicina (PRATI, 2006). A classificação nas bases de dados desses tipos de aplicaçõestêm em comum que a classe rara, a de menor ocorrência, é a mais importante parao problema. Como exemplo para essas aplicações estão a detecção de fraudes emcartões de crédito e diagnóstico médico (MORE, 2016).

As bases de dados podem ser designadas como desbalanceadas quando ao menos uma classe possui um número muito menor de instâncias comparadas a uma outra classe, comumente chamada de majoritária (BHOWAN *et al.*, 2012). Devido ao fato de os algoritmos aprendizagem serem muito sensíveis, o desbalanceamento causaum enviesamento na hora da classificação, ocasionando na tendência de valorizar mais a(s) classe(s) de maior predominância (PRATI, 2006).

Na prática, o desbalanceamento irá afetar diretamente a acurácia. Pois a mesma poderá ter sido calculada com um aparente alto valor. Porém com uma alta taxa de falsos negativos, que por estarem em número absoluto menor, não diminuirão a taxa de acurácia como um todo. O grande problema disso é que a alta taxa de falsos negativosadvém da classe rara, que geralmente é a de maior interesse no estudo (BHOWAN*et al.*, 2012). Logo, é necessário utilizar outras métricas de avaliação quando há uma desproporção de objetos de classes diferentes. Na Figura 21 fica mais perceptível, aovisualizar no espaço amostral, o quão difícil é classificar classes desbalanceadas.

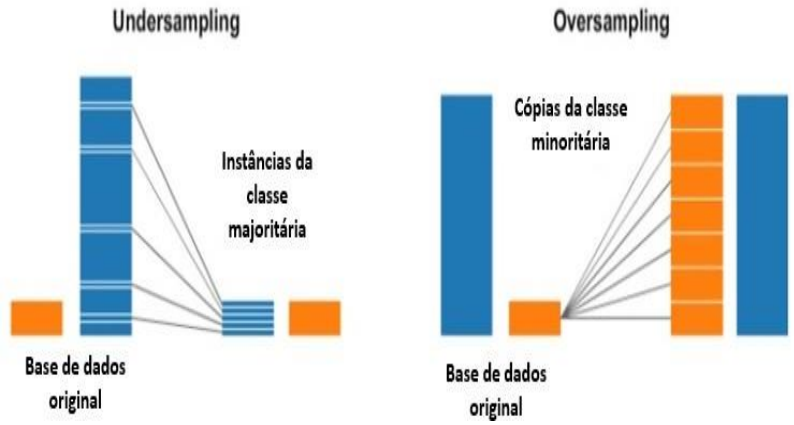
Figura 21– Espaço amostral de duas classes desbalanceadas



Fonte: Adaptado de More (2016).

Para conter esse comum problema do desbalanceamento existem diversas técnicas que vêm sendo desenvolvidas e testadas ao longo do anos. Entretanto, se analisadas de perto, todas elas se resumem a duas categorias: sobreamostragem (*over- sampling*) e subamostragem (*undersampling*). Esses procedimentos visam aumentar o número de objetos da classe rara ou diminuir o da classe majoritária, entretanto *undersampling* e *oversampling* aleatórios podem levar, respectivamente, a perda de informação e *overfitting*, que representa na estatística quando um modelo se ajusta muito bem aos dados observados mas se mostra ineficaz para novas classificações (BARELLA, 2016). Na Figura 22 é possível ver, através de um esquema visual, como funciona esses dois procedimentos na base de dados original.

Figura 22– Undersampling e oversampling



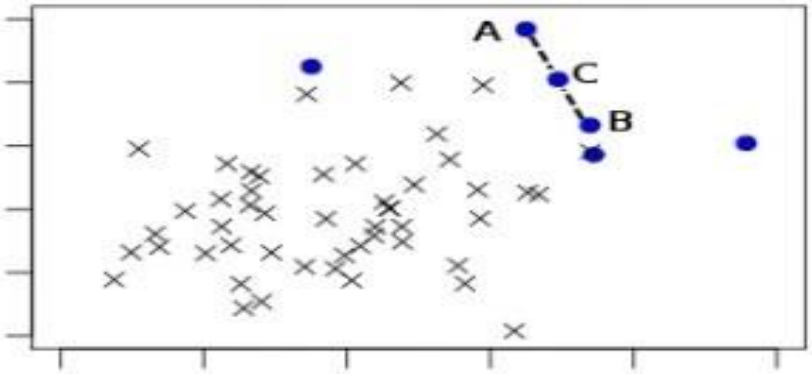
Fonte: Adaptado de Medium (2019)

Quando se trata de undersampling e oversampling informativos, isto é, aqueles que utilizam técnicas de fato para redução e aumento dos objetos, dois métodos, SMOTE (Synthetic Minority Oversampling Technique) e CNN (Condesed Nearest Neigh- bor Rule) modificada, são os mais comumente aplicados.

O SMOTE é um método de *oversampling* desenvolvido por Chawla *et al.* (2002) que gera vetores sintéticos da classe minoritária. Esses novos vetores são gerados a partir da Equação (3.9), onde o novo vetor pertencerá à semi-reta (devido ao β que deverá ser entre 0 e 1) que liga o vetor escolhido como base gerador naquela iteraçãoao seu vizinho da mesma classe mais próximo (BARELLA, 2016). A Figura 23 ilustra acriação de um novo elemento com o método SMOTE.

$$Obj_{novo} = Obj_{esc} + (Obj_{viz} - Obj_{esc}) * \beta \quad (3.9)$$

Figura 23– Elemento C criado através do SMOTE



Fonte: Barella (2016)

A CNN modificada, técnica de *undersampling* apresentada por Hart (1968) e atualizada por Kubat *et al.* (1997), utiliza o algoritmo KNN com um vizinho (k=1) pararetirar os objetos de maior redundância da classe majoritária. A técnica funciona seleci-onando aleatoriamente um objeto da classe majoritária, e apenas utilizando o objeto selecionado tenta classificar todos os outros, lembrando de definir o k=1. Os classificados corretamente (aqueles de mesma classe e mais próximos ao objeto escolhido como referência) são definidos como redundantes, e os incorretos, juntamente com oreferência, são definidos como representativos (os que serão usados na classificação real) (PRATI, 2006).

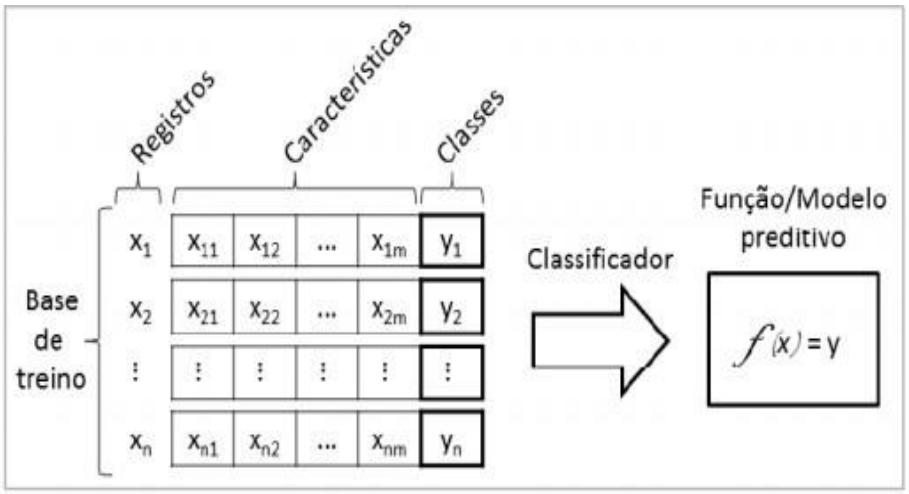
3.7 Aprendizagem de Máquina Supervisionada

A Aprendizagem Supervisionada é realizada com a utilização de instâncias, ou objetos, com classes já definidas e conhecidas previamente pelas pessoas responsáveis pelo estudo. É necessário que haja essas instâncias já rotuladas, devido ao fato que uma parte delas será usada no grupo de treino. Com tal quadro, o algoritmo classificador "aprenderá"sobre os padrões das classes para aplicar as regras de classificaçãoem casos não-rotulados (KOTSIANTIS *et al.*, 2007).

A Figura 24 representa, de forma resumida, o esquema de aprendizagem demáquina supervisionado, onde há uma base previamente rotulada utilizada para treino, características extraídas e o classificador que gera um modelo preditivo.

Com relação à base de treino é possível utilizá-la simplesmente dividindo o conjunto inicial em teste, para a verificação posterior dos acertos, e treino. Geralmenteessa divisão é feita na proporção de 30/70% para teste e treino, respectivamente. Uma outra forma de divisão é conhecida como validação cruzada, que pode ser utilizada naforma de *k-fold* ou *leave-one-out*.

Figura 24– Esquema de aprendizagem de máquina supervisionada

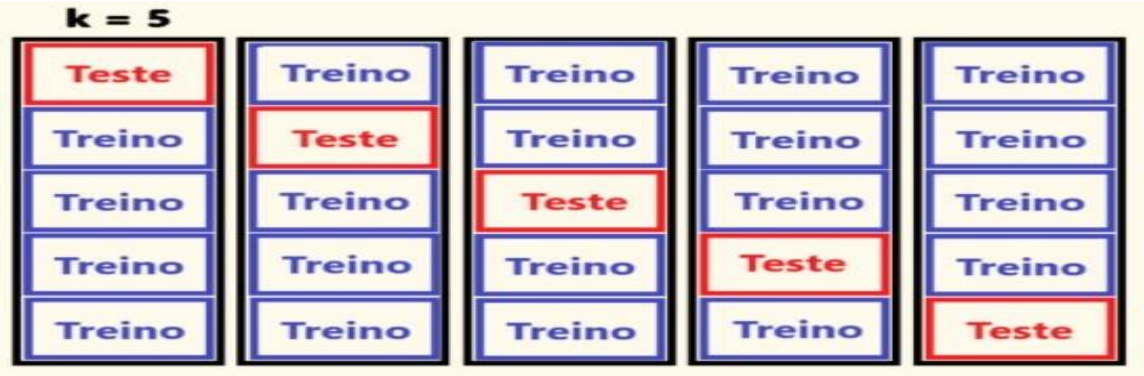


Fonte: Borchartt (2013)

Na validação cruzada com *k-fold* o conjunto inicial é subdividido em k grupos independentes. Durante o processo um grupo é empregado como sendo conjunto de teste enquanto os outros juntos são empregados como conjunto de treino. Isso é feito até que todos os grupos tenham sido utilizados como conjunto de teste uma vez (OLIVERA, 2016).

O *leave-one-out*, que é uma caso específico do *k-fold*, utiliza N-1 instâncias para o treinamento do modelo (admitindo N a quantidade total de instâncias) e apenas umainstância para o teste. O processo é repetido N vezes, até que todos os indivíduostenham sido usados para teste. Devido ao fato de ser um processo com muito maisiterações, e com um custo computacional muito grande, esse método é utilizado em circunstâncias específicas (OLIVERA, 2016). A Figura 25 facilita a visualização do funcionamento da validação cruzada.

Figura 25– Validação cruzada k-fold com k=5



Fonte: Didatica (2020a)

3.8 Classificadores

Diversos tipos de algoritmos de classificação podem ser aplicados na aprendizagem de máquina. Alguns irão apresentar melhores resultados com mais características outros com menos. As dificuldades de alguns deles podem ser em decorrência do enviesamento ou da variabilidade. Sempre haverá vantagens e desvantagens com a utilização de algum algoritmo, logo, é de extrema importância avaliar todas as condições do problema estudado para tomar a decisão de quais algoritmos levar em consideração na classificação.

Abaixo há uma breve explicação do funcionamento de cinco algoritmos clássicos bastante utilizados em problemas de classificação supervisionada.

3.8.1 KNN - K-Nearest Neighbors

O KNN é um método de larga utilização em problemas de classificação pelo fato de ser um dos mais simples e que, muita das vezes, corresponde com uma boa performance (DUDA *et al.*, 1973). Para a classificação das instâncias não rotuladas,o KNN utiliza a classe majoritária dos K vizinhos mais próximos de uma determinada instância. A métrica para calcular a distância dos vizinhos é escolhida previamente, sendo a Euclidiana, dada pela Equação (3.10), a utilizada na maioria dos casos (SUN;HUANG, 2010). Onde na equação um vetor x_i (ou x_j) pode ter n dimensões, o a_r seria um exemplo da r^a dimensão, e o w_r um possível peso (diferente de um) atribuído a uma dimensão

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} w_r \left(a_r(x_i) - a_r(x_j)\right)^2} \tag{3.10}$$

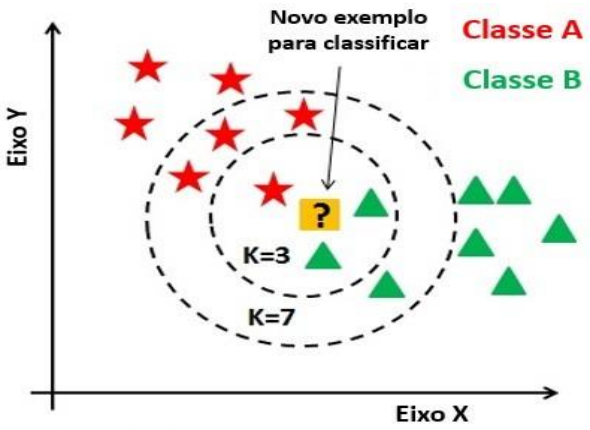
O número de vizinhos também é definido previamente, e tem importante influência na classificação final. Nesse caso, é comum utilizar valores próximos do resultado da Equação (3.11) para o K, sendo o N o número total de instâncias. Quanto maior foro número de instâncias no estudo, é interessante que o K também seja grande para evitar problemas com *outliers* (instâncias posicionadas fora da região onde a maioria das outras instâncias de sua classe se encontram).

$$K = \sqrt{N} \tag{3.11}$$

Apesar de sofrer com altas dimensões e enviesamento, é o método mais indicado

quando se pede agilidade e simplicidade na interpretação dos resultados. Um exemplo ilustrativo pode ser visto na Figura 26, onde a mudança de K=3 para K=7 alteraria a classificação final.

Figura 26– Influência na escolha do K na classificação final



Fonte: Adaptado de Datacamp (2020)

3.8.2 Regressão Logística

A Regressão Logística (RL) é um dos mais importantes algoritmos usados na estatística e na classificação de dados. Dentre suas vantagens é possível citara facilidade para lidar com variáveis independentes categóricas, o fornecimento dos resultados em termos probabilísticos, o baixo número de suposições necessárias e a facilidade para aplicação de um alto número de técnicas de otimização (MAALOUF,2011).

A RL classifica as instâncias baseado nas probabilidades de ocorrência de evento para cada classe. Esse cálculo é feito com auxílio da função logística, que podeser vista na Equação (3.12).

$$P(Y = 1) = \frac{1}{1+e^{-g(x)}} \tag{3.12}$$

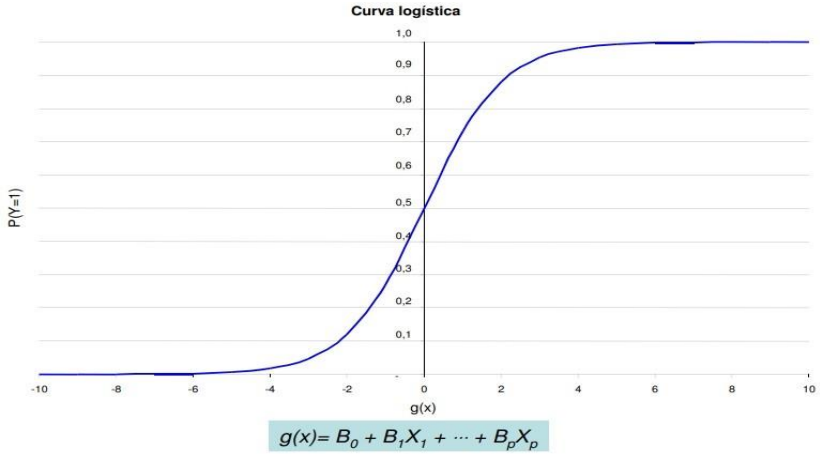
$$g(x) = B_0 + B_1X_1 + \cdots + B_jX_j \tag{3.13}$$

Na Equação (3.12) foi admitido que o Y poderia assumir dois valores (0 ou 1). Na Equação (3.13) os X_j (j=1,p) são as características daquela determinada instância, e os coeficientes $\boldsymbol{B_j}$ são estimados com uso do método da máxima verossimilhança, que busca encontrar a combinação de coeficientes que irá maximizar a probabilidade daquele evento (de

ser daquela classe).

Um comportamento característico observado nas curvas logísticas é o seu formato em forma de S. Na Figura 27 é possível ver com mais clareza de que forma ascurvas logísticas tendem a ser (HOSMER; LEMESHOW, 1989).

Figura 27– Curva de uma regressão logística



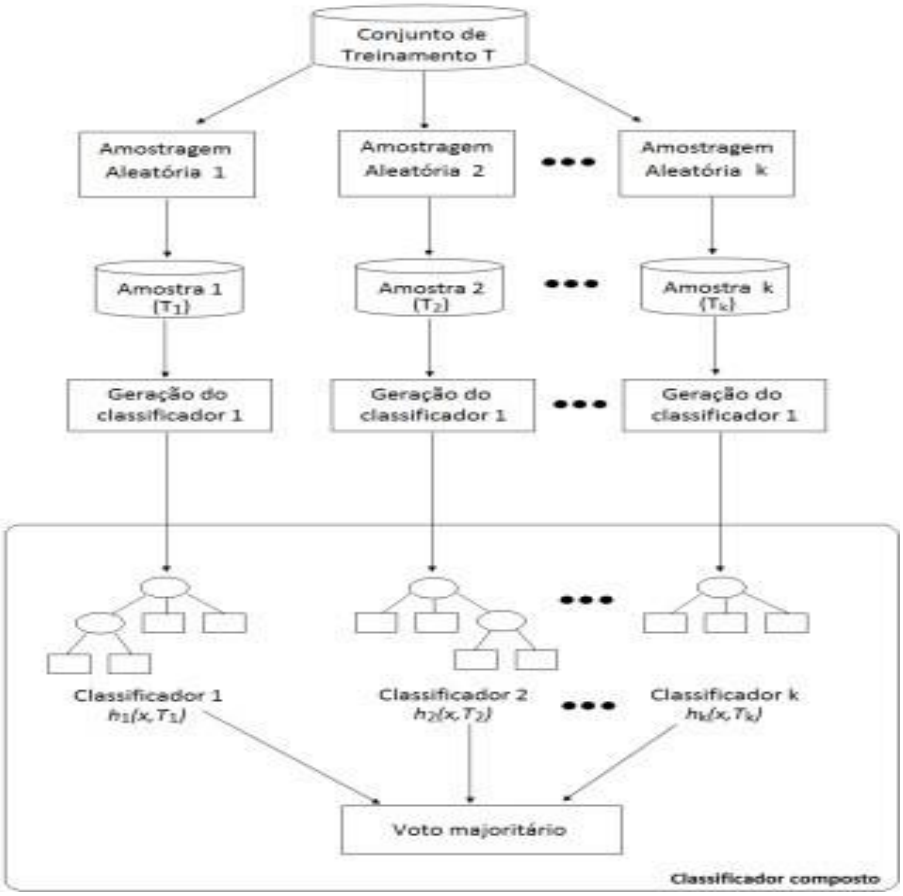
Fonte: Autor

Para a classificação em si, num caso binário por exemplo, a RL iria calcular a probabilidade de uma instância ser da classe Y=1. Sendo essa probabilidade maiorque 0,5 (50%), a instância seria atribuída à classe 1, sendo menor que 0,5 ela seria atribuída à classe 0.

3.8.3 Random Forest

A *Random Forest* é um algoritmo de classificação categorizado como *ensemble*.Os classificadores pertencentes a essa categoria optam, na maioria das vezes, por definir o resultado da classificação final deles através do voto majoritário. Essa votação é feita com o comparativo da quantidade de vezes que o mesmo resultado de classificação foi alcançado pelos diverso ramos do *ensemble*. Na Figura 28 o esquema resumido de classificação por meio do voto majoritário com a *random forest* é ilustrado (OSHIRO, 2013).

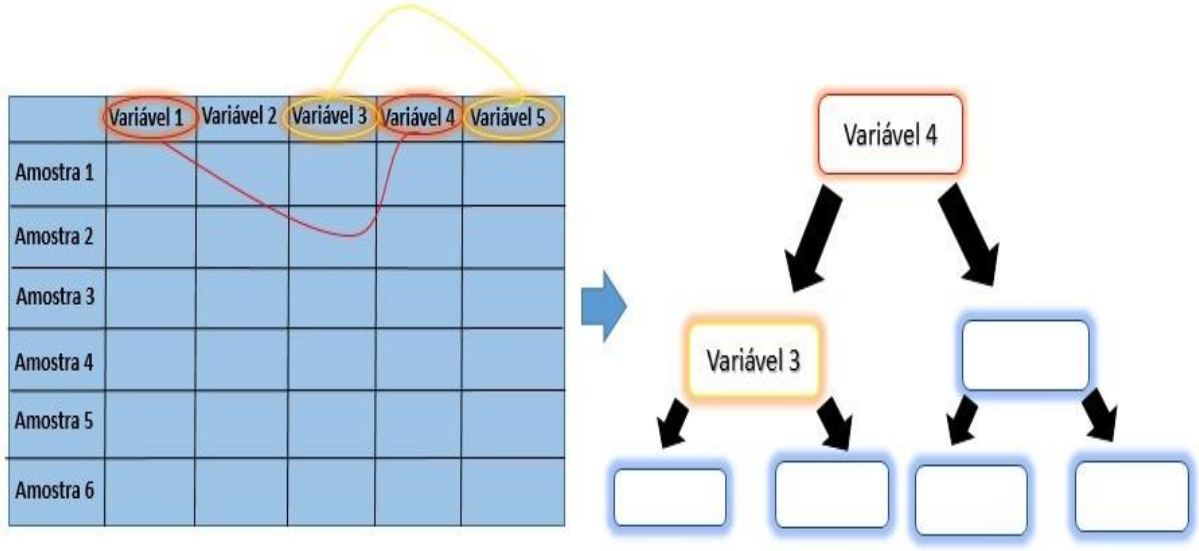
Figura 28– Classificação final por meio do voto majoritário



Fonte: Oshiro (2013)

No caso específico da *Random Forest*, a partir do conjunto de treino são selecionadas aleatoriamente algumas amostras que formarão a primeira árvore do conjunto.O primeiro nó dessa árvore é criado com uma seleção aleatória de duas variáveis. Emseguida essas duas variáveis são comparadas pelos algoritmos da entropia (Equação 3.5) ou do índice de Gini (Equação 3.8), para identificar qual é a mais relevante. A criação do nós seguintes é feita da mesma forma, porém as variáveis previamente sorteadas não podem ser selecionadas novamente (STROBL *et al.*, 2008). Na Figura 29 é possível visualizar o processo de criação dos nós da árvore.

Figura 29– Criação de um nó numa árvore de Random Forest



Fonte: Adaptado de Didatica (2020b)

Na Figura 29, é possível perceber que o primeiro nó foi construído com a variável 4 e um dos segundos nós com a variável 3. E como dito anteriormente, elas foram escolhidos por serem mais relevantes (comparação via algoritmo de entropia) que as variáveis 1 e 5 respectivamente, que também foram escolhidas aleatoriamente, para aconstrução dos nós.

Como a seleção das amostras é feita com reamostragem, essas variáveis podem ser selecionadas mais uma vez na construção da árvore seguinte, que seguirá o mesmo esquema de elaboração acima. Essas escolhas de *bootstrap* (seleção com reamostragem) e seleção aleatória de duas variáveis, para comparação em cada nó, embora pareçam facilitar a criação de árvores individualmente "ruins", fortalecem o conjunto da *Random Forest* por evitar o *overfitting*.

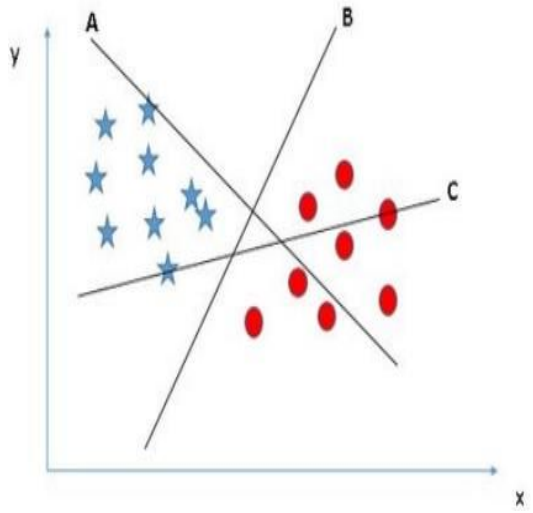
3.8.4 Support Vector Machine

A SVM (Support Vector Machine - Máquina de Vetores de Suporte), desenvolvida por Vapnik (1995) pode ser considerado um dos classificadores mais recentes e populares aprendizagem de máquina. Baseado na teoria da aprendizagem estatística e fundamentado na Minimização do Risco Estrutural (*Structural Risk Minimization* - *SRM*), esse algoritmo é amplamente utilizado para resolver problemas não-lineares, com alta dimensionalidade e encontrar mínimos locais (DERIS *et al.*, 2011)

Na classificação, o algoritmo executa uma plotagem dos vetores (valor das instâncias) em um espaço n-dimensional, onde n se refere ao número de características (ou

atributos) utilizados naquele problema. Em seguida, é criado um hiperplano de separação das classes estudadas. Os chamados vetores de suporte são os vetores quese localizam nas fronteiras de separação entre as classes. É a partir desses possíveis vetores que o algoritmo testará diversos hiperplanos até encontrar o que melhor separe as classes (ANDREOLA, 2009). Na Figura 30 está ilustrado um exemplo onde há três possíveis hiperplanos para o problema de classificação com duas classes.

Figura 30 – Exemplo de três possíveis hiperplanos para separação linear com SVM

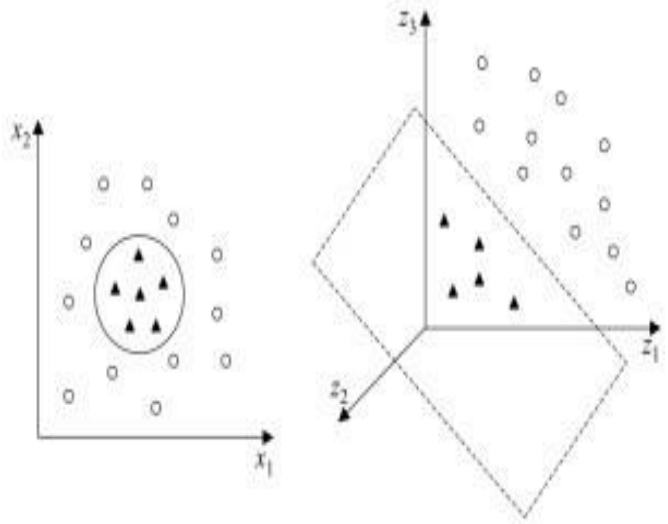


Fonte: Próprio Autor

O SVM define o melhor hiperplano para o problema identificando a separação mais precisa entre as classes. Mesmo quando isso ocorre é preciso selecionar aquele que tem a distância maximizada para os vetores de suporte de ambas as classes (ANDREOLA, 2009).

Nos casos em que não há possibilidade de separar as classes linearmente, o SVM mapeia o conjunto de dados de treinamento para utilizá-los como entrada na transformação de um espaço de maior dimensionalidade. As funções de transformação não precisam ser adicionadas manualmente. Na prática são comumente utilizadas as funções *kernel* mais conhecidas, como a Polinomial, Gaussiana e a Sigmoidal.Em cada uma dessas funções há parâmetro(s) distinto que devem ser modificados de acordo com o problema (LORENA; CARVALHO, 2007). Na Figura 31 há uma clara transformação de um espaço bidimensional para um tridimensional, facilitando aseparação das classes.

Figura 31– Aplicação da transformação de dimensão para SVM não-linear



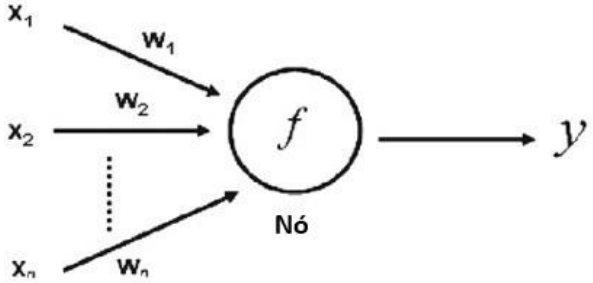
Fonte: Lorena e Carvalho (2007)

3.8.5 Redes Neurais

O algoritmo Redes Neurais, de aprendizagem de máquina, também conhecido como ANN (*Artificial Neural Network*), provém da ideia de simular um cérebro humano. Embora os computadores de hoje operem em altíssima velocidade, eles ainda nãopodem ser comparados ao sistema de processamento de um ser humano. Foi a partir dessa percepção, que o desenvolvimento de algoritmos que tentam replicar a ideia cerebral nasceu. Mais precisamente, esse algoritmo replica a arquitetura dos neurôniosbiológicos e seu entorno, como: axônios, sinapses e dendritos (ZOU *et al.*, 2008)

Partindo desse princípio biológico, o algoritmo irá receber *inputs* com seus respectivos pesos e aplicá-los numa função de transferência. A partir do resultado dessa função é que o sinal será ativado ou não, gerando o *output*. Um esquema do funcionamento de um modelo básico de rede neural está ilustrado na Figura 32.

Figura 32 – Modelo básico de um rede neural com um nó



Fonte: Adaptado de Zou *et al.* (2008)

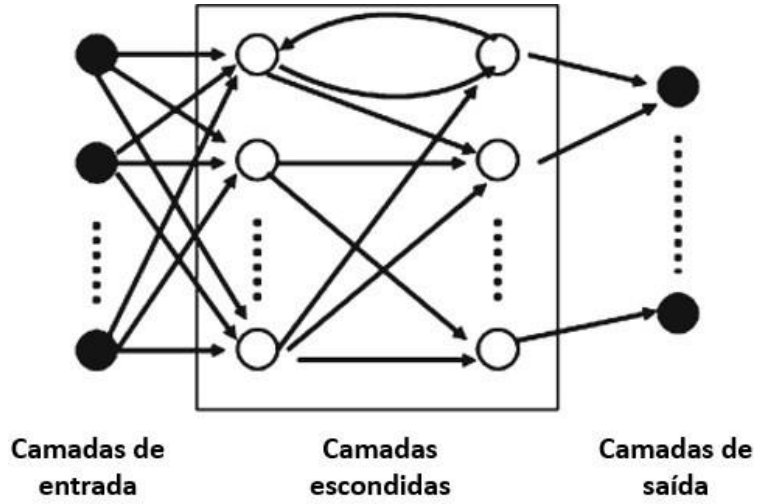
A forma mais simples do funcionamento de uma ANN é a função de limiar, vistana Equação (3.14). Onde a funçã]o objetivo é simplesmente ser maior que o T, valor limite, para poder ser ativada (HAYKIN, 2007). Entretanto, diversos tipos de funções podem ser testados e outras mais complexas, por se adequarem melhor ao problema,são comumente utilizadas na literatura, como: sigmoide, tangente e hiperbólica.

$$y = f(\sum_{i=0}^{n} w_i x_i - T) \tag{3.14}$$

Além da escolha das funções, outros parâmetros como a quantidade de neurô nios escondidos (ou intermediários) são essenciais na construção da arquitetura da rede neural. É preciso mensurar o tamanho que a rede deve ter, pois embora ela ganheem processamento e aprendizado com mais neurônios, o seu custo computacional também irá crescer.

O grande diferencial das redes neurais é a possibilidade de utilização da re-aprendizagem, que na forma mais complexa é feita com a técnica *feedback*. Nessecaso, a rede é considerada dinâmica, com os *outputs* de algumas camadas sendo os *inputs* de outros. Os pesos são atualizados a cada iteração até que o sistema convirja(taxa de erro muito baixa) ou o número de iterações pré-determinado chegue ao fim. O parâmetro taxa de aprendizagem é crucial na velocidade e alcance da convergência ótima. Na Figura 33 há um esquema simples de reaprendizagem pelo método *feedback*

Figura 33– Esquema de reaprendizagem de uma rede neural com feedback



Fonte: Adaptado de Zou *et al.* (2008)

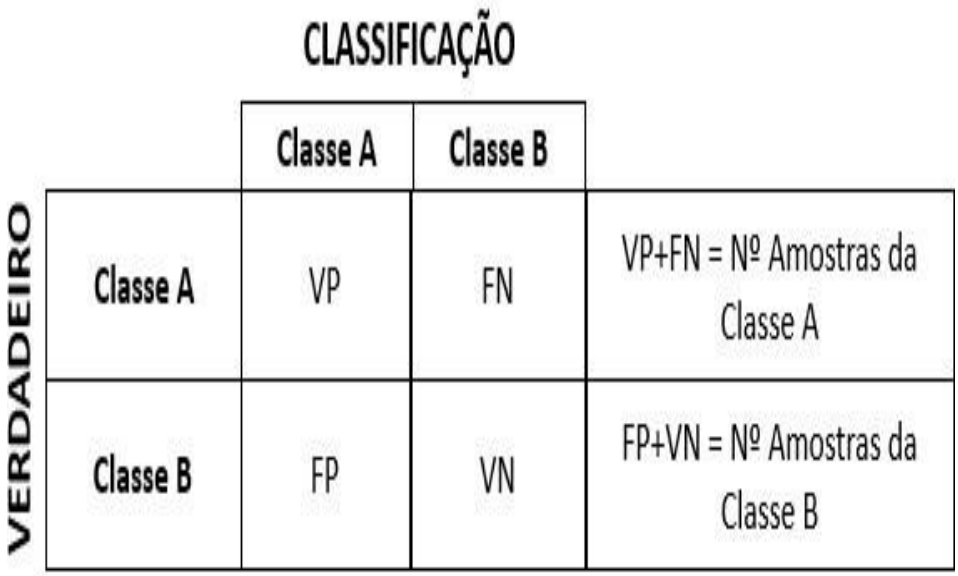
3.9 Métricas de Avaliação

Após a aprendizagem e classificação por parte dos algoritmos escolhidos, é ne-cessária

uma avaliação da performance, afinal só é possível qualificar uma classificação depois de uma quantificação dos resultados. Existem diferentes métricas de avaliação usadas em aprendizagem de máquina, no entanto todas utilizam os VP (Verdadeiro Positivo), VN (Verdadeiro Negativo), FP (Falso Positivo) e FN (Falso Negativo) comobase para o cálculo.

Uma tabela conhecida como matriz de confusão é o local onde ficam organizados os VN, VP, FN e FP de determinada classificação. No eixo vertical estão as classes classificadas pelo algoritmo, no eixo horizontal as classes verdadeiras, e na diagonal principal estão indicado os acertos de classificação. Na Figura 34 há um modelo de matriz de confusão para classificação binária.

Figura 34– Modelo de matriz de confusão



Fonte: Adaptado de VASCONCELOS *et al.* (2017)

As métricas mais difundidas na classificação supervisionada são: Acurácia, sensibilidade, especificidade e precisão. A acurácia, Equação (3.15), também conhecidacomo taxa de acerto, traz uma performance mais geral da classificação, medindo a proporção de VP e VN para todos os casos analisados. A sensibilidade, Equação (3.16), assim como a especifidade, Equação (3.17), mede a proporção de acertos para uma determinada classe sob a quantidade real de instâncias naquela classe. E a precisão, Equação (3.18), diferentemente da sensibilidade, mede o percentual de acerto de umaclasse sob todas as instâncias classificadas naquela classe.

Entretanto em estudos binários (como diagnósticos médicos), a sensibilidade é atribuída à taxa de VP, proporção de pacientes corretamente diagnosticados como doentes, e a

especificidade à taxa de VN, proporção de pacientes corretamente diagnosticado como saudáveis. (BARATLOO *et al.*, 2015).

$$Acurcia = \frac{VP+FP}{VP+FP+VN+FN} \quad (3.15)$$

$$Sensibilidade = \frac{VP}{VP+FN} \quad (3.16)$$

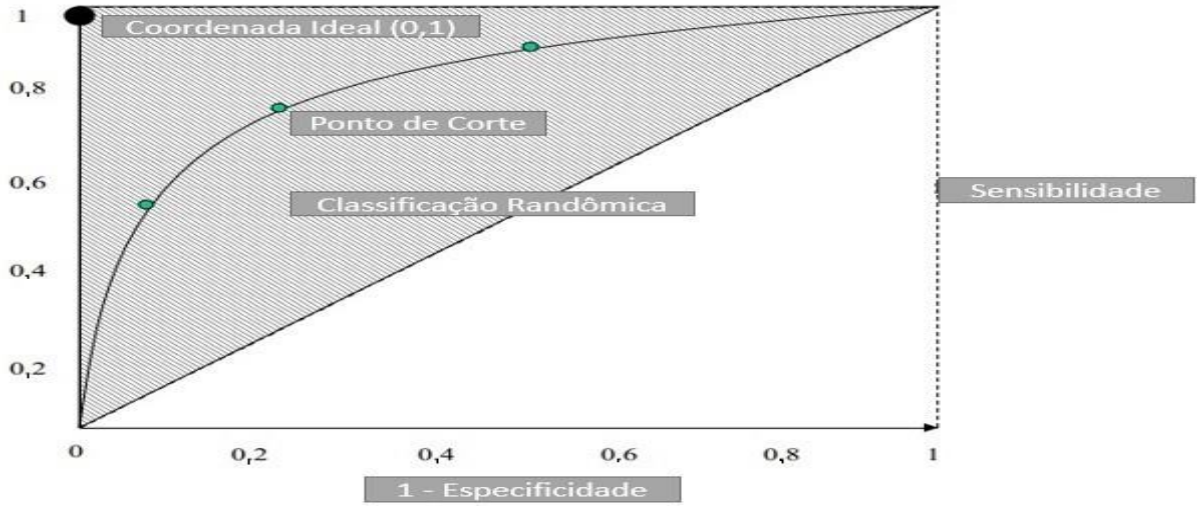
$$Especificidade = \frac{VN}{VN+FP} \quad (3.17)$$

$$Preciso = \frac{VP}{VP+FP} \quad (3.18)$$

Contudo, em classes desbalanceadas essas três métricas podem ficar com valores fora da realidade, e gerar erros na interpretação. Por exemplo em uma análise binária de 100 objetos onde há apenas 2 de uma determinada classe, e tudo é classificado como sendo da classe majoritária, a acurácia e a especificidade serão altíssimas, porém a sensibilidade seria de 0%. Por esse motivo alguns pesquisadores utilizam outras métricas, menos usuais, mas que não sofrem tanto com o efeito do desbalanceamento.

Uma métrica que não sofre efeito do desbalanceamento é a análise ROC (*Receiver Operating Characteristics*), que é feita com a plotagem de um espaço quetem como eixos as proporções de VP (sensibilidade) e FP (1- especificidade). Quanto mais próximo do ponto (1,0) os pontos da proporção real do problema estiverem, melhor será a qualidade da classificação. A quantificação dessa métrica se dá com o cálculo da área sob a curva formada pelos pontos de proporção, valores acima de 80% sãoconsiderados bons (ZHU *et al.*, 2010). Na Figura 35 é possível observar o espaço ROC com mais detalhes.

Figura 35– Espaço ROC



Fonte: Zhu *et al.* (2010)

O *F1-Score*, Equação (3.19), é uma outra alternativa ao uso das métricas padrão, ao passo que, também como a ROC, tem seu impacto minimizado pelo desbalance amento. O desenvolvimento dessa métrica foi feito utilizando a média harmônica da precisão e sensibilidade. Logo, caso umas dessas duas métricas esteja com grandediferença das outras no valor apresentado, o resultado da F1-*Score* será afetado diretamente. Entretanto, o mais indicado pela literatura é a análise de todas essas métricasem conjunto, para uma certeira verificação dos pontos fortes e fracos do classificador.

$$F1 - Score = \frac{2 * Preciso * Sensibilidade}{Preciso + Sensibilidade} \tag{3.19}$$

3.10 Orange Canvas

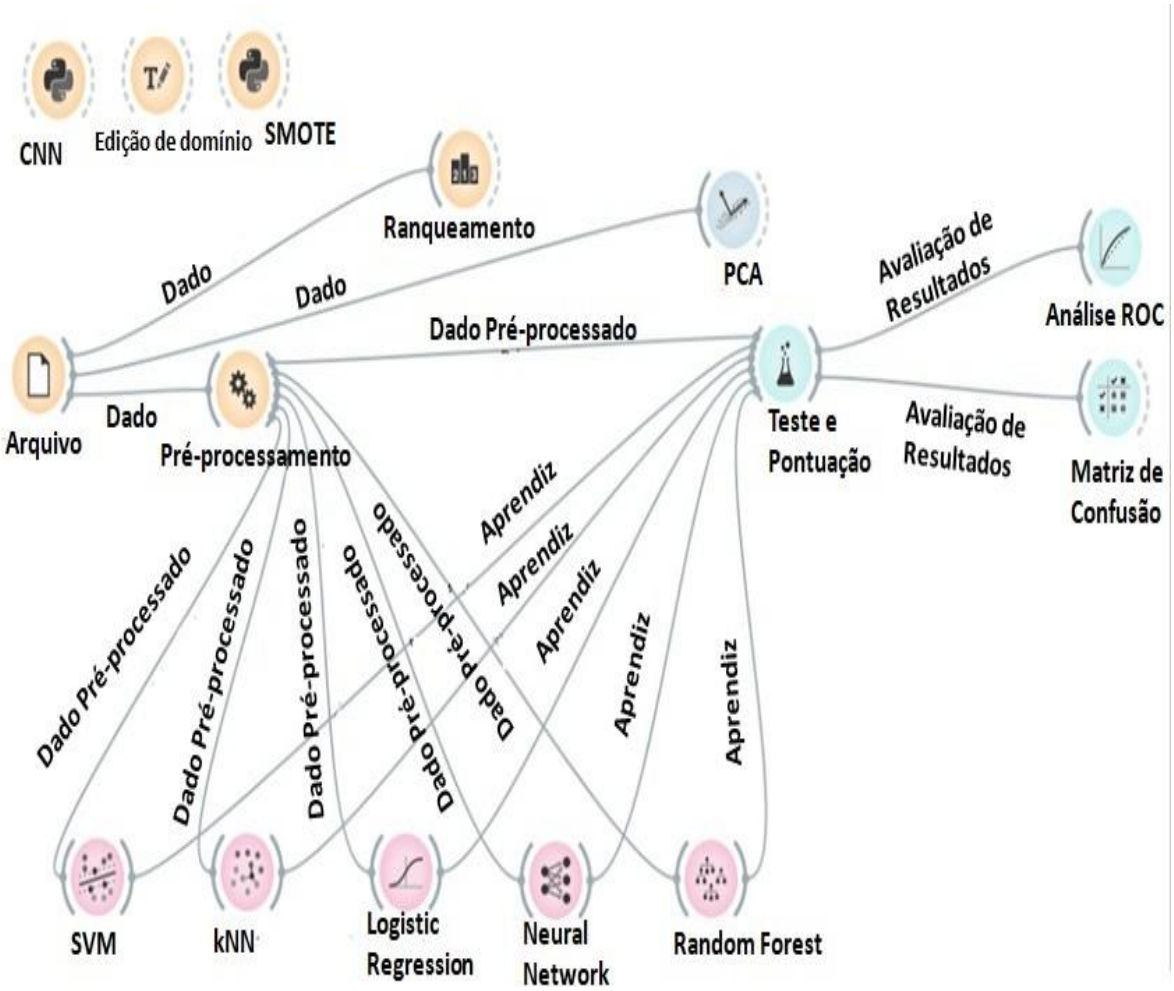
Orange Canvas é um *software* de código aberto, escrito na linguaguem Python e desenvolvido pela *University of Ljubljana*. O *software* consiste em uma interface visual pela qual o usuário insere *widgets* e cria um fluxo de trabalho.

Esses *widgets* oferecem funcionalidades, como ler os dados, visualizar os dadosde diferentes formas, pré-processamento, algoritmos de aprendizagem de máquina, etc. O usuário tem a liberdade de criar seu próprio fluxo de trabalho a partir de da união dos *widgets*, além de poder criar sua própria funcionalidade em um *widget* por meio daescrita de código em Python. O Orange vem crescendo cada vez mais no âmbito da bioengenharia por trazer facilidade de trabalho na exploração dos dados à pessoas deoutras áreas de estudo (DEMŠAR *et al.*, 2004).

Na Fig 36 é possível observar um típico fluxograma de trabalho no Orange.Com

destaque para os algoritmos de aprendizagem, na parte inferior, as otimizações do processo como: SMOTE, CNN, Ranqueamento, mais acima, e a parte de teste e avaliação de resultados, no canto direito da imagem.

Figura 36– Área completa de trabalho do Orange Canvas para classificação



Fonte: Próprio autor

4 METODOLOGIA

O presente capítulo tem como objetivo apresentar a metodologia utilizada nessa dissertação e detalhar as ferramentas empregadas em cada fase do processo de classificação. A pesquisa teve como foco buscar a simplificação e refinamento do procedimento de classificação binária e multiclasses (Maligno, Benigno, Cisto e Normal) de imagens termográficas.

Com finalidade de se obter os melhores resultados na classificação final, foramaplicadas diversas otimizações durante o processo. A plataforma computacional *Orange Canvas* foi largamente explorada por ser de simples utilização, de código aberto ede incorporar boa parte dos passos do método proposto. Como complementação ao *Orange Canvas*, também foram utilizados os sistemas CAD (*Computer-Aided Design*) desenvolvidos por Araújo (2014) e Dourado (2014).

Os cinco algoritmos de classificação mais conhecidos (KNN, SVM, ANN, RF e LR) foram utilizados, bem como a técnica *leave-one-out*, possibilitando uma comparação com os trabalhos de Queiroz *et al.* (2016) e NOVA (2017), que utilizaram as duas bases de dados utilizadas aqui analisadas.

O fluxograma da Fig 37 ilustra as etapas da metodologia adotada.

Figura 37– Fluxograma da metodologia proposta

Fonte: Próprio autor

4.1 Base de Dados

As duas amostras utilizadas nessa pesquisa são de imagens termográficas obtidas do HC/UFPE entre os anos de 2005 e 2014. Essas amostras foram adquiridas seguindo os protocolos citados anteriormente no capítulo 3.

A primeira amostra, chamada de Base 1, foi utilizada por NOVA (2017), e é composta por 62 imagens previamente divididas em 4 classes. Desse total, 17 são benignos, 17 malignos, 14 com cisto e 14 não apresentam anormalidade. A segunda amostra, chamada de Base 2, utilizada por Queiroz *et al.* (2016) e de maior tamanho,possui 229 imagens e também foi dividida em 4 classes. A segunda amostra é compostapor 77 casos benignos, 44 malignos, 41 com cisto e 67 sem anormalidades.

A Tabela 1 mostra a distribuição da quantidade de pacientes nas quatro classes para a primeira e segunda amostra.

Tabela 1– Distribuição por classe da quantidade de instâncias para as duas amostras

Diagnóstico	**Base 1**	**Base 2**
Tumor Maligno	17	44
Tumor Benigno	17	77
Cisto	14	41
Normal	14	67

Visando o estabelecimento de processos adequados à triagem médica, foi utilizada a classificação binária (câncer e não-câncer). Nesse tipo de classificação foram considerados não-câncer os pacientes diagnosticados com o tipo benigno, comcisto e sem anormalidades. Com isso, a primeira e a segunda amostras passaram a ter17 câncer e 45 não-câncer, e 77 câncer e 152 não-câncer respectivamente.

A Tabela 2 mostra a distribuição da quantidade de pacientes nas duas classes para a primeira e segunda amostras.

Tabela 2– Distribuição por classe da quantidade de instâncias para as duas amostras

Diagnóstico	**Base 1**	**Base 2**
Câncer	17	77
Não-Câncer	45	152

O desbalanceamento, principalmente na segunda amostra, enviesa os resultados dos classificadores à Classe Benigno e pode tornar o resultado final prejudicado. Logo, foi feito balanceamento de classes com a técnica de criação de vetores sintéticos nas duas amostras, e a técnica da retirada de objetos não representativos apenas paraa segunda amostra. Para os estudos com triagem, o enviesamento ocorre para a classe não-câncer de forma bastante acentuada. Portanto, a técnica de criação de vetores sintéticos foi usada nessa situação.

4.2 Segmentação

A metodologia utilizada para a segmentação das regiões de interesse (ROI) das imagens termográficas foi feita com duas frentes: técnicas de forma automática, elaborada por Dourado (2014), e de forma manual, elaborada por Araújo (2014).

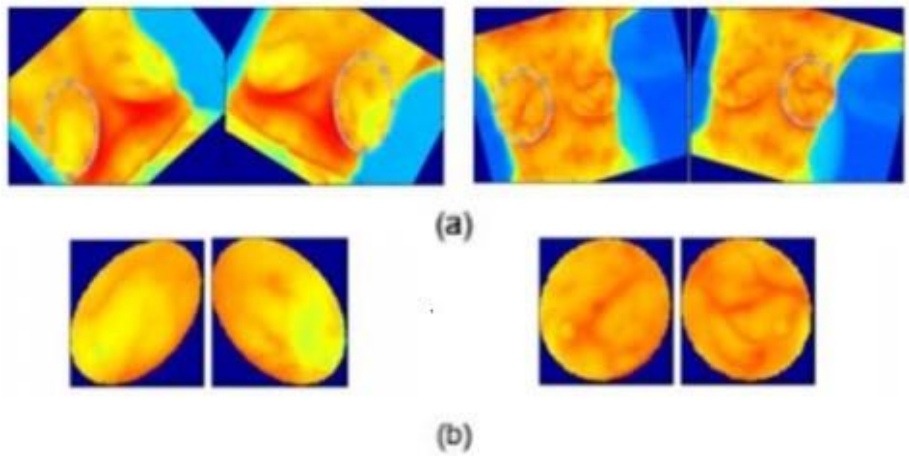
Nos dois casos os arquivos das imagens digitais, já no formato .csv, foram importados no *Matlab*. Na plataforma, a imagem foi reconstruída utilizando cores atribuídas aos diferentes valores de temperatura.

No processo automático, a região de interesse é alcançada com a separação docorpo da paciente do fundo da imagem comparando os níveis de cinza. As outras partes do corpo, como

pescoço, axilas e abdômen, são retiradas utilizando limites inferiores e superiores pré determinados. Todo o processo automático é feito sem intervenção humana. De forma diferente, no processo manual o usuário necessita selecionar aregião das mamas desejada, com auxilio de elipses geradas pela plataforma.

Os dois métodos de segmentação utilizados são considerados efetivos para o objetivo proposto, com validações em outras pesquisas da área. Uma comparação dos dois métodos de segmentação, para uma mesma linha de metodologia, traz à tona asvantagens e desvantagens entre eles a partir dos resultados alcançados. Um exemploda segmentação manual utilizada pode ser visualizado na Figura 38.

Figura 38– (a) Imagens rotacionadas e com elipses delimitando as regiões de interesse. (b) Imagens finais obtidas para as regiões de cada uma das mamas.



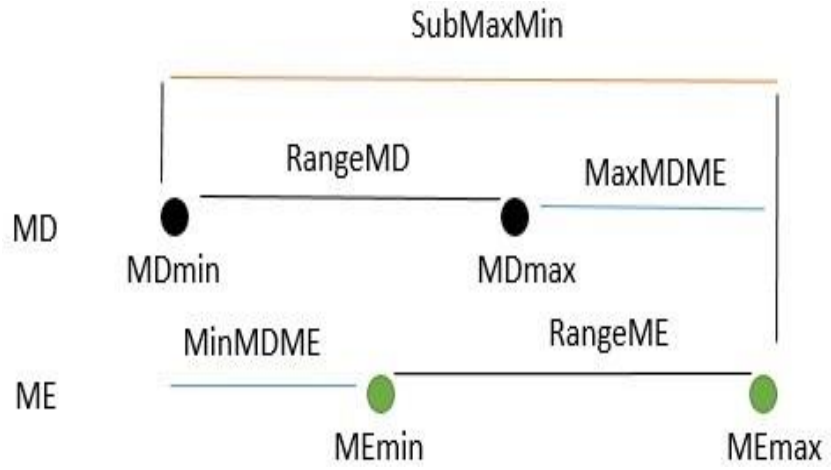
Fonte: Araújo (2014).

4.3 Extração de características

Após a conclusão dos dois tipos de segmentações, e obtenção das regiões de interesse de cada imagem termográfica, iniciou-se o processo de extração dascaracterísticas da mesma. Em todos os casos o *Matlab* foi utilizado como ferramenta para fazer essa extração. No primeiro tipo de características (estatísticas), a extração foi feita em conjunto com a segmentação, isto é, usando no mesmo código. No segundo tipo (fractais), um código à parte foi desenvolvido partindo da imagem já segmentada.

Além das características estatísticas mais comuns, foram extraídas, tambémnesse primeiro grupo, faixas de variações (variáveis intervalares) que utilizam as temperaturas máximas e mínimas de cada mama do mesma imagem termográfica(ARAÚJO, 2014). A Figura 39 traz de forma visual a construção desse tipo de variável.

Figura 39 – Demonstração visual das faixas de variações que utilizam temperaturas máximas e mínimas



Fonte: Autor.

As Equações (4.1), (4.2), (4.3), (4.4) e (4.5) são, respectivamente, as variações máximas em cada mama, a variação máxima entre as duas mamas, a variação entre os máximos de cada mama e a variação entre os mínimos de cada mama.

$$RangeMD: \left|Max_{MD}(P_{ij}) - Min_{MD}(P_{ij})\right| \quad (4.1)$$

$$RangeME: \left|Max_{Me}(P_{ij}) - Min_{Me}(P_{ij})\right| \quad (4.2)$$

$$SubMaxMin: \left|Max_{max}(P_{ij}) - Min_{min}(P_{ij})\right| \quad (4.3)$$

$$MaxMDME: \left|Max_{MD}(P_{ij}) - Max_{Me}(P_{ij})\right| \quad (4.4)$$

$$MinMDME: \left|Min_{MD}(P_{ij}) - Min_{Me}(P_{ij})\right| \quad (4.5)$$

Além das variáveis associadas aos intervalos de temperatura nas mamas (Eq.4.1 a 4.5), foram extraídas também as seguintes variáveis estatísticas: Média (Eq. 4.6), Desvio Padrão (Eq. 4.7), Assimetria (Eq. 4.8) e Curtose (Eq. 4.9) para cada mama de uma mesma imagem. A Média fornece o valor da temperatura média em cada mama, o Desvio Padrão informa o quão uniforme é um conjunto de dados, a Assimetria mede a assimetria das caudas da distribuição de probabilidade e a Curtose caracteriza o achatamento da curva da função de distribuição de probabilidade.

$$Media: \mu = \frac{1}{M.N}\sum_{i=1}^{N}\sum_{j=1}^{M} P_{i,j} \quad (4.6)$$

$$Desvio\ Padrão: \sigma = \sqrt{\frac{1}{M.N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(P_{i,j} - \mu\right)^2} \quad (4.7)$$

$$Assimetria: \gamma_1 = \sigma^{-3}\left[\frac{1}{M.N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(P_{i,j} - \mu\right)^3\right] \quad (4.8)$$

$$Curtose: \sigma^{-4}\left[\frac{1}{M.N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(P_{i,j} - \mu\right)^4\right] \quad (4.9)$$

As características fractais fornecem ótimas explicações sobre superfícies e fenômenos da natureza. Por esse motivo, elas vêm sendo cada vez mais utilizadas na caracterização de rugosidade e de superfícies. Por aplicar critérios que facilitam no rastreamento de padrões entre *pixels* com valores próximos, esse tipo de característica também é muito utilizado no processamento de imagens médicas (MANDELBROT; MANDELBROT, 1982).

Os critérios para a criação de padrões são dados através de descritores de textura, as características fractais. No presente trabalho as características fractais foram extraídas com os métodos da Lacunaridade, do coeficiente de *Hurst* e *Box-Counting*.

O coeficiente de Hurst está relacionado com o quanto a imagem estudada ocupa o espaço que a contém, dessa forma se faz necessário decidir o tamanho da janela móvel para que a imagem seja dividida em quadrantes, para este trabalho foi utilizado Hurst com janela de 5 e de 7. De forma diferente, a Lacunaridade busca caracterizar aimagem quantificando o total de vazios nela (SERRANO *et al.*, 2010).

Por fim, no método box-counting a dimensão fractal é medida com o a divisão da imagens em quadrados e posteriormente a contagem dessa quantidade de quadrados. Uma diminuição do tamanho dos quadrados é realizada, e com isso o aumento da quantidade deles, para que seja possível calcular o coeficiente angular da reta obtida pelo Log (N_1) x Log (N_2), onde N_1 e N_2 são respectivamente a quantidade de quadradosem cada imagem (ANTONIAZZI, 2007).

4.4 Redução de desbalanceamento

Após todas as características extraídas, foi realizado o primeiro tipo de otimização nas bases de dados, o balanceamento. As duas bases possuem algum grau desbalanceamento, algo

que afeta os resultados, por fazer o classificador ser tendencioso para as classes de maior quantidade de elementos. Logo, técnicas de *oversampling* e *undersampling* foram testadas a fim de comparar os resultados finais com e sem a presença desses artifícios.

No caso do *oversampling*, o SMOTE foi a técnica escolhida e aplicada em todos os tipos de classificação (binário e quatro classes) por se tratar de um aumentode instâncias. Entretanto, no caso do *undersampling*, a técnica CNN, que removeinstâncias, foi aplicada somente nas classificações binárias, pois as quantidades deinstâncias por classe nesse caso era bem maior, possibilitando tal remoção. Essasduas técnicas foram aplicadas por códigos de programação diretamente no *framework* do Orange, plataforma computacional utilizada.

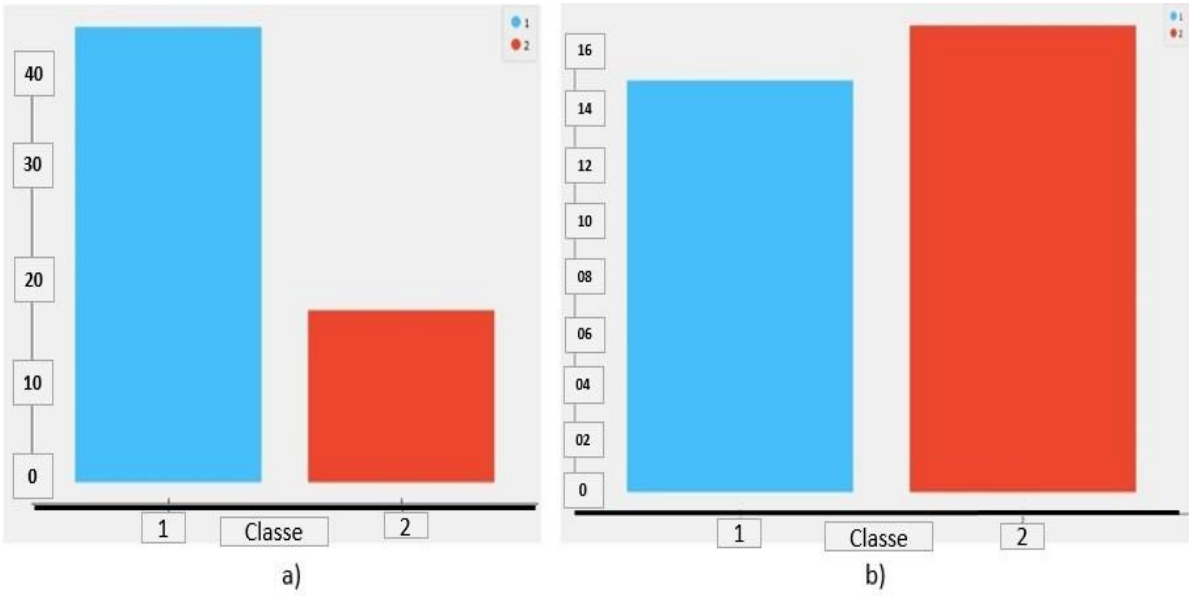
Nas Figuras 40 e 41 é possível observar o impacto final na quantidade de instâncias das bases após a aplicação das técnicas. Mais especificamente, na Figura 40 foi utilizada a técnica CNN, com a perceptível queda no número de instâncias da Classe 1, caindo de 40 para 16. E na Figura 41, após aplicação do SMOTE, as Classes3 e 4 igualaram em quantidade as Classes 1 e 2.

A Tabela 3 traz uma legenda para melhor entendimento da cor/numeração da classe correspondente, vale ressaltar que essa legenda também será a válida no capítulo 5 (Resultados).

Tabela 3– Legenda de cores/numeração - classes

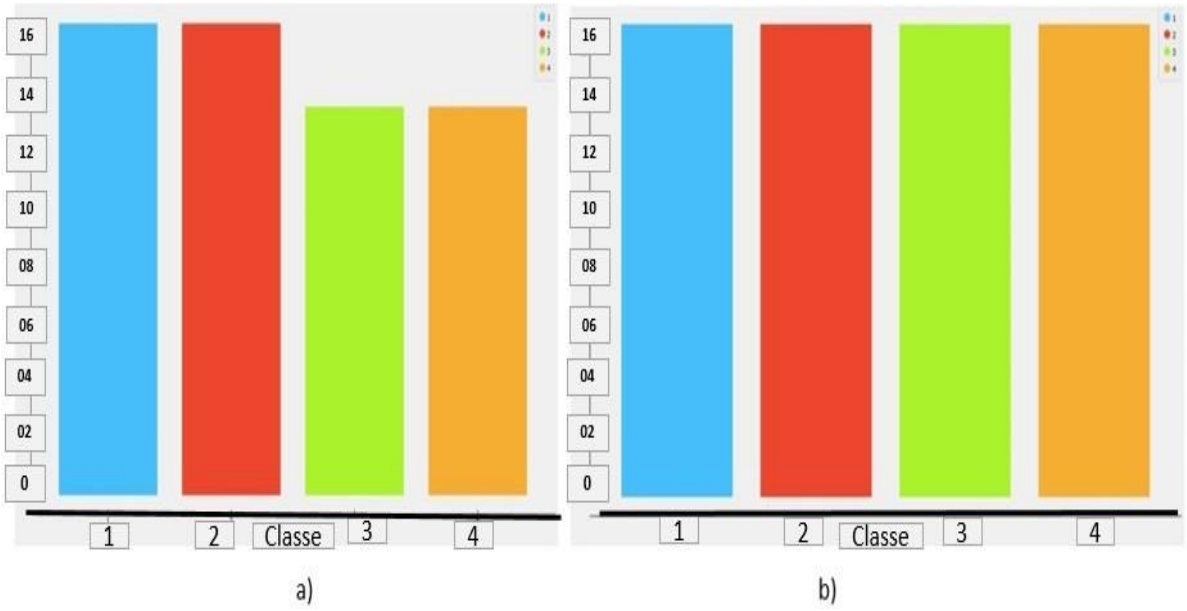
Classe	Cor	Numeração
Benigno		1
Maligno		2
Cisto		3
Normal		4

Figura 40– (a) Distribuição da quantidade de elementos da Base 1 sem CNN (b) Distribuição da quantidade de elementos da Base 1 com CNN.



Fonte: Próprio autor.

Figura 41– (a) Distribuição da quantidade de elementos da Base 1 sem SMOTE (b)Distribuição da quantidade de elementos da Base 1 com SMOTE.



Fonte: Próprio autor.

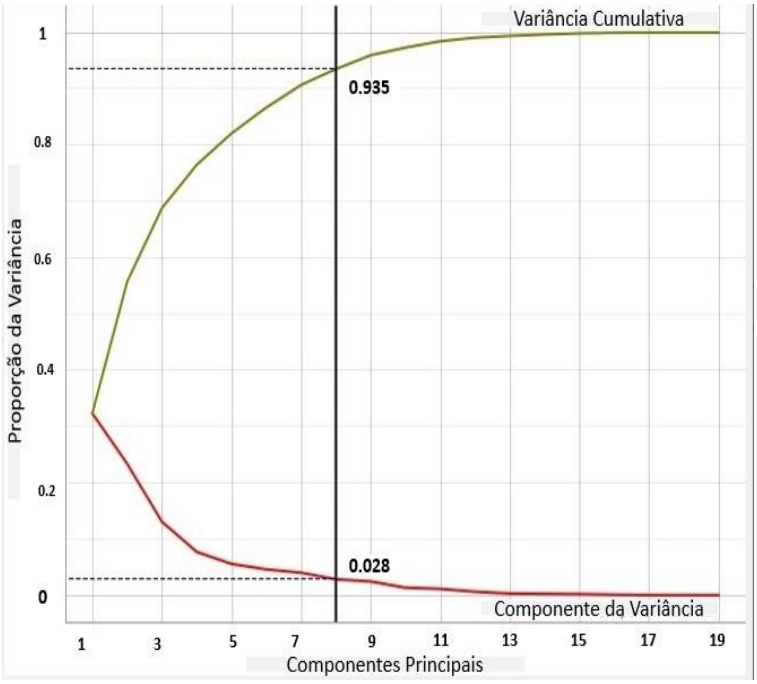
4.5 Redução da dimensionalidade

O segundo tipo de otimização utilizado nas bases foi a redução da dimensio-nalidade. Essa otimização é indicada quando a quantidade de características é alta,algo que gera uma maior complexidade do problema, dificultando a classificação eaumentando o tempo de processamento.

Assim como a redução de desbalanceamento essa otimização não é obrigatória. Entretanto, no presente trabalho ela foi usada através de três diferentes técnicas, paracomparação do resultado final, e sempre com a base já balanceada pela primeira técnica de otimização.

A primeira técnica usada foi o PCA que, diferentemente das outras duas técnicas, não reduz o número de características por exclusão das mesmas, mas sim por transformação. Após diversos testes, com os melhores resultados alcançados, a quantidade de características transformadas com o PCA foi padronizada como sendo de oito para todos os casos. Entretanto, a variância cumulativa para esse quantidade de variáveis se alterava para cada caso, mas com uma média de 89%, o que aindaindicava um bom número de variabilidade que se manteve. Na Figura 42, oriunda de uma tela de saída do *Orange Canvas*, é possível visualizar a variância cumulativa de 93,5% para 8 características, bem como a alteração da variância conforme é alterado onúmero de características selecionadas com o PCA.

Figura 42– Alteração da variância cumulativa de acordo com o número de características selecionados



Fonte: Próprio autor.

A segunda técnica, o Ranqueamento, é mais objetiva por se tratar apenas de selecionar as características que permanecerão na classificação utilizando um *rank*. Asatribuições de "pontuação"desse rank são feitas com a qualificação das caracteríticas por meio das métricas *Information Gain*, *Gain Ratio* e *Gini Index*. Assim como o PCA, após testes apontarem os melhores resultados com uma determinada quantidade de características, foi pré-determinado,

para fins de comparação, como sendo de quatro características a quantidade a ser padronizada para esse método de redução. A Figura 43, também retirada da tela de saída do *Orange Canvas*, traz o esquema detalhado, destacando as quatro características selecionadas via ranqueamento, em um dos testes feitos no presente trabalho.

Figura 43– Ranqueamento dos atributos de acordo com as métricas escolhidas

	#	Info. gain	Gain ratio	Gini
MedME		0.098	0.049	0.051
DevPadME		0.071	0.036	0.034
RangedosMin		0.068	0.034	0.039
MedMD		0.068	0.034	0.039
DevPadFDD		0.051	0.026	0.027
RangeMD		0.045	0.022	0.026
RangedosMax		0.043	0.022	0.024
DevPadMD		0.043	0.022	0.024
DevPadFDE		0.036	0.018	0.018
CurtoseMD		0.035	0.017	0.018
MedFDE		0.035	0.017	0.019
CurtoseME		0.034	0.017	0.019
RangeME		0.024	0.012	0.013
HurstDW7		0.017	0.009	0.009
HurstEW7		0.013	0.007	0.007
HurstEW3		0.013	0.007	0.007
RangeMaxMin		0.010	0.005	0.005
HurstDW3		0.009	0.004	0.005
LacuD		0.004	0.002	0.002

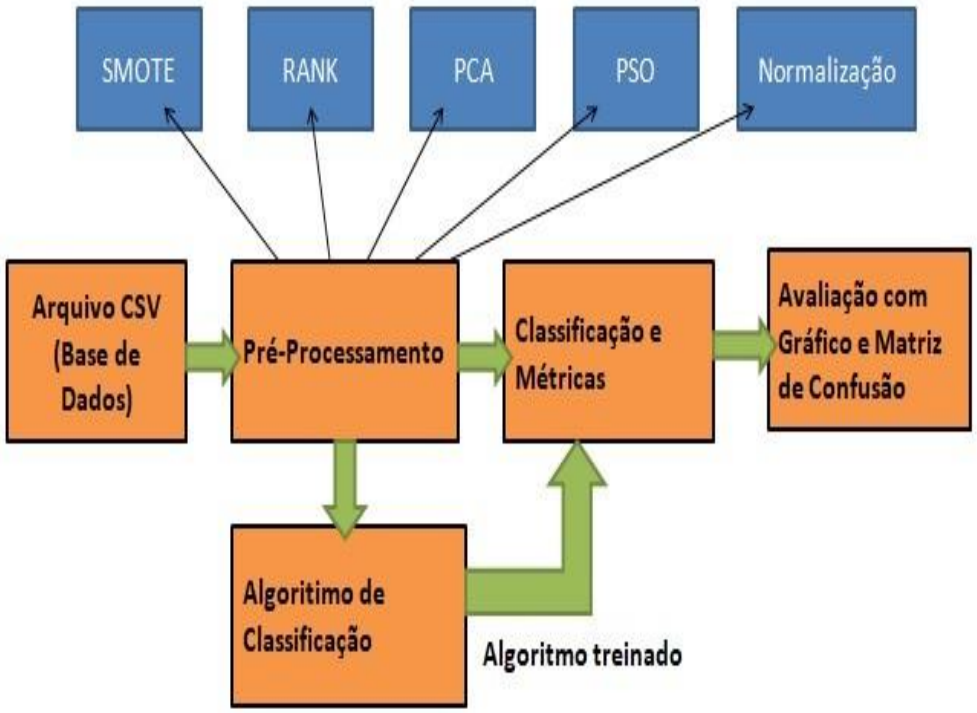
Fonte: Próprio autor.

A terceira técnica usada foi o PSO. Embora, assim como o Rank, tenha como produto final a exclusão de algumas características originais, o PSO tem processo bemmais complexo para chegar no seu resultado. No presente trabalho, a implementaçãodo código do PSO foi realizada utilizando linguagem Python e auxílio da biblioteca "pyswarms". A regressão logística, por ser o algoritmo mais indicado na literatura, foiusado como classificador de varredura da função objetivo.

4.6 Técnicas de classificação

Depois de extrair todas as características e aplicar as otimizações mencionadas anteriormente, a fase da classificação propriamente dita foi iniciada. Nessa etapa os cinco tipos de classificadores escolhidos foram acoplados ao ambiente pré-criado com as otimizações anteriores na plataforma *Orange Canvas*. A Imagem 44 retrata, no formato de diagrama de blocos, como é desenhado o ambiente de pré-processamentoe classificação com os algoritmos escalados.

Figura 44– Diagrama de blocos de funcionamento do Orange Canvas



Fonte: Próprio autor.

É importante recordar que os classificadores escolhidos foram *Logistic Regression, Random Forest, KNN, SVM* e ANN. Como cada um deles tem um processo declassificação bastante diferente dos outros, é interessante observar como as distribuições das bases de dados, bem como as características extraídas se adaptaram a cadaclassificador.

Em todos os casos, a técnica de validação cruzada, *leave-one-out*, foi utilizadapara gerar uma maior segurança quanto aos resultados, porque a quantidade de dadosnas duas bases, a nível de aprendizagem de máquina, é bastante pequeno. Para permitir futuras possibilidades de replicações e melhoras, os parâmetros utilizados emcada classificador estão disponíveis no Apêndice A.

Com o objetivo de testar em diferentes vertentes de classificação e melhorar os resultados dos trabalhos que usaram as duas bases de dados aqui analisadas, opresente trabalho seguiu dois caminhos. No primeiro, as duas bases foram classificadas em quatro classes pelos cinco classificadores com as opções de: sem otimização, utilizando as otimizações SMOTE e CNN (só para a Base 2), e com uma das duas opções da primeira otimização aliada à uma das três opções do segundo tipo de otimização. Dessa forma foi possível observar diretamente nos resultados o impacto das otimizações, e suas melhores

incorporações aos classificadores. Além disso, ao fim das classificações foi feita uma comparação dos resultados entre os casos em quea segmentação das imagens foi realizada de forma automática e de forma manual. Para Base 1, além da comparação entre as resultados oriundos de diferentes segmentações,uma outra comparação, entre casos em que a segmentação manual da mesma base foi feita por dois autores diferentes, foi realizada. Para a Metodologia 2, procurandootimizar resultados na análise para triagem médica, a sequência de comparações se manteve, tendo única diferença a classificação dos casos, que foi feita de forma binária.

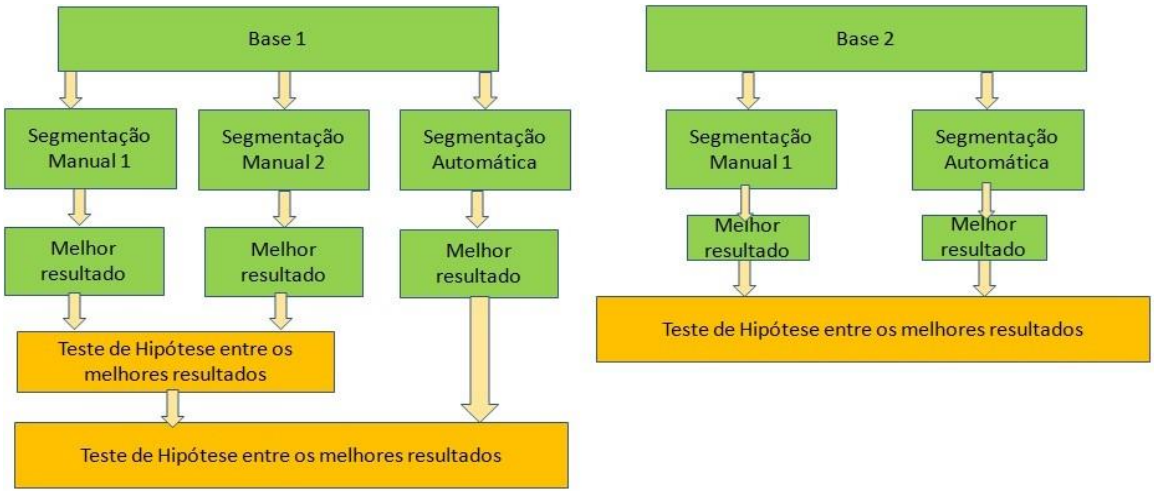
A eficiência das classificações, a fim de comparações entre os resultados, foi mensurada com auxílio de métricas de avaliação como Acurácia, Sensibilidade paraClasse Maligna, Precisão, *F1-Score* e ROC.

4.7 Análise estatística

Uma análise pós-teste é importante em todos os trabalhos que envolvam comparação numérica de resultados. Logo, após alcançar os resultados finais nas diferentesbases com as diversas formas de otimização, foram realizados testes de hipótese e testes de Wilcoxon pareado (devido ao fato de não poder afirmar a distribuição normaldos dados das amostras avaliadas) . Esses testes visam comparar se os valores encon-trados nos resultados, para os diferentes casos, possuem diferenças estatisticamente significativas.

Com isso, foi realizado um teste de hipótese entre o melhores resultados da Base 1 com a segmentação manual feita por dois usuários diferentes (tendo como métrica a acurácia), e entre os melhores resultados de cada tipo de segmentação na mesma base (Base 1 e Base 2). A Figura 45 exibe, de forma mais clara, o esquema detestes de hipótese entre as formas de segmentação e as bases.

Figura 45– Fluxograma da aplicação do teste de hipótese



Fonte: Próprio autor.

Os testes foram feitos utilizando a tabela *t-student* pelo fato de a captação de dados ter sido feito com apenas 10 amostras (valor menor que 30) de cada casoselecionado para o teste. A hipótese nula foi escolhida como sendo que os resultadosnão possuem diferenças estaticamente significativas, uma vez que se quer provar o contrário, e como segunda hipótese, que o resultado do Caso A é maior (ou menor) que resultado do Caso B.

A Equação (4.10) foi utilizada para esse tipo de teste de hipótese, o valor resultante dela foi comparado com o valor designado na tabela *t-student* de acordocom o número de dimensões subtraído de dois, nesse caso específico foi o valor foi18 por se utilizarem (10+10) amostras. O nível de significância para todos os testes foi de 5%. Na Equação (4.10) os X_{m1}, X_{m2}, Sp^2, n_1 e n_2 se referem, respectivamenteàs médias de métrica de comparação (acurácia), a variância amostral combinada e a quantidade de amostras.

A Equação (4.11) exibe como é feito o cálculo da variância amostral combinada,sendo σ_1 e σ_2 os desvios padrões de cada caso.

$$t_0 = \frac{(X_{m1} - X_{m2})}{\sqrt{Sp^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{4.10}$$

$$Sp^2 = \frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2} \tag{4.11}$$

A mesma lógica se seguiu para os testes de Wilcoxon, com a comparação de dados das mesmas 10 amostras dos melhores resultados de segmentações diferentesnas respectivas bases, mesma hipótese nula e nível de significância.

Entretanto, para o teste de Wilcoxon a equação utilizada é diferente. No caso de pequenas amostras (quantidade de amostras menor que 20) o cálculo do *p-value* éfeito com o auxílio da Equação (4.12) aplicando cálculo binomial ao teste dos sinais.

x - número de vezes que o sinal menos frequente ocorre.

n - tamanho da amostra descontando os empates.

$$p = P(X \leq x) = C_n^x . (\frac{1}{2})^n + C_n^{x-1} . (\frac{1}{2})^n + C_n^{x-2} . (\frac{1}{2})^n + \cdots + C_n^0 . (\frac{1}{2})^n \quad (4.12)$$

5 RESULTADOS

Neste capítulo são mostrados os resultados obtidos com a aplicação das metodologias desenvolvidas no capítulo anterior. Esses diversos resultados são realizados para que se possa ter clareza e confirmação da técnica e algoritmo de classificaçãoque melhor se comporta classificando as imagens termográficas das bases de dados consideradas.

Inicialmente, na Metodologia de classificação multiclasses (Metodologia 1), foram extraídos os resultados das duas bases de dados utilizando uma classificação em quatro classes (Benigno, Maligno, Cisto e Normal) com e sem aplicações de técnicas adicionais como vetores sintéticos, redução de instâncias, PCA, Rank e PSO. Além disso, cada base de dados teve uma extração de características via segmentação manual, desenvolvida por Araújo (2014), e via segmentação automática, desenvolvidapor Dourado (2014), o que gerou a comparação de resultados obtidos entre as duas técnicas de segmentação.

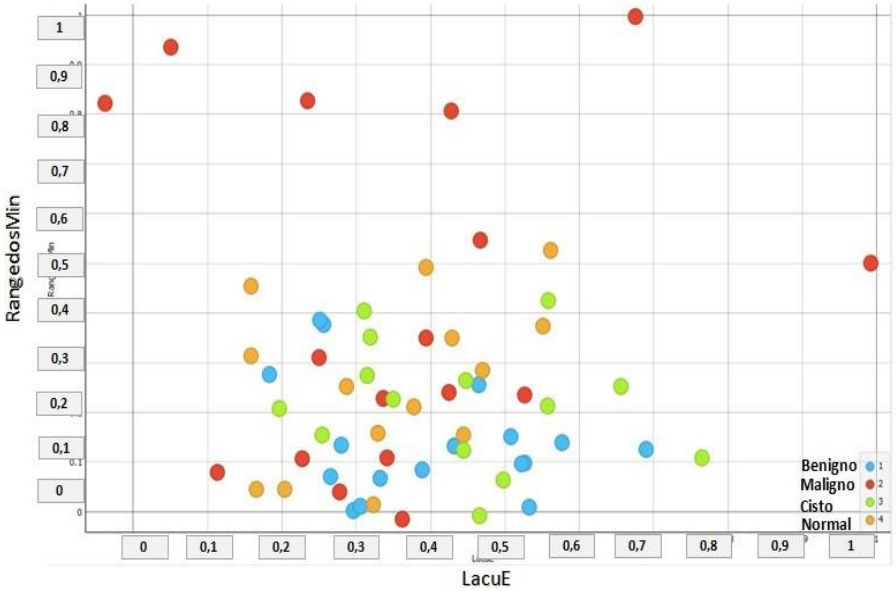
Posteriormente, na Metodologia de Classificação Binária (Metodologia 2), utilizando classificação binária (câncer e não-câncer), o mesmo procedimento foi realizadonas duas bases de dados. Assumiu-se como "não-câncer"os pacientes que estavam previamente classificados como "cisto", "benigno"e "normal".

5.1 Resultados obtidos através da Metodologia de Classificação Multiclasses

Base 1

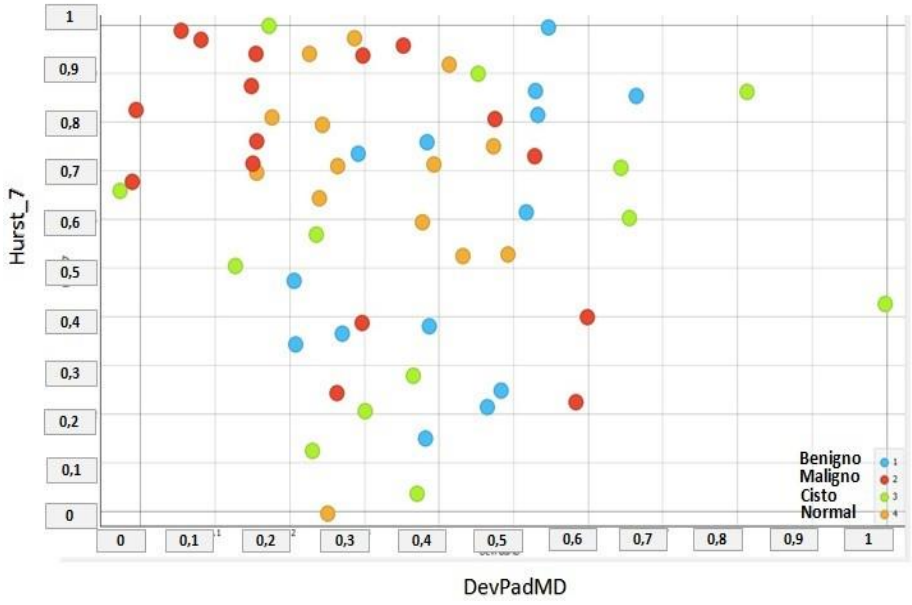
Na Figuras 46 e 47 é possível observar a distribuição das instâncias, com suas respectivas classes, em dois espaços bidimensionais formados pelas melhores combinações de duas características para cada caso, conforme descrito na metodologia.

Figura 46– Espaço de características, extraídas via segmentação manual, para a Base 1



Fonte: Próprio autor.

Figura 47– Espaço de características, extraídas via segmentação automática, para a Base 1



Fonte: Próprio autor

Os resultados postos a seguir foram construídos visando comparar, em uma única tabela para cada tipo de segmentação, os melhores valores alcançados em cadamétrica, dentre os resultados dos cinco algoritmos para cada técnica. É importante reforçar que, como já mencionando anteriormente, para a Base 1 com 4 classes, atécnica da CNN não foi utilizada pelo fato de a quantidade de instâncias por classeser pequena. Na Base 1 o algoritmo ANN

obteve a melhor performance, e por issoseus resultados foram os utilizados para comparação entre as técnicas e avaliação da classificação.

Nas Tabelas 4 e 5, com extrações oriundas de duas diferentes segmentações manuais, os melhores resultados foram alcançados com a técnica do **SMOTE+PCA** de**oito** características. Embora a Tabela 5 aparente resultados um pouco melhores, se faz necessário um teste estatístico para confirmação dessas diferenças, para entãoter mais certeza da influência do usuário na segmentação manual. Na Tabela 6, os melhores resultados vieram com a utilização da técnica do **SMOTE+Ranqueamento** para as **quatro** melhores características, e foram bastante similares aos da Tabela 4.

Tabela 4– Resultado da melhor performance de cada técnica via segmentação manual feita pelo próprio autor

Técnica	**Acurácia**	**Sensibilidade Maligno**	**Precisão**	**AUC**	**F1**
Normal	35,3%	35,5%	35,5%	63,7%	35,3%
SMOTE	50,0%	53,8%	50,8%	71,3%	50,1%
SMOTE + PCA 8	**52,9%**	**47,1%**	**53,8%**	**74,5%**	**53,0%**
SMOTE + Rank 4	44,1%	35,3%	45,0%	72,7%	44,3%
SMOTE + PSO	48,5%	41,2%	49,2%	69,2%	48,0%

Tabela 5– Resultado da melhor performance de cada técnica via segmentação manual feita por NOVA (2017)

Técnica	**Acurácia**	**Sensibilidade Maligno**	**Precisão**	**AUC**	**F1**
Normal	42,6%	37,5%	42,6%	67,1%	42,1%
SMOTE	51,5%	52,9%	51,0%	73,6%	51,0%
SMOTE + PCA 8	**54,4%**	**58,8%**	**54,3%**	**76,6%**	**54,3%**
SMOTE + Rank 4	38,2%	47,1%	37,6%	67,4%	37,9%
SMOTE + PSO	48,5%	41,2%	48,3%	71,9%	47,9%

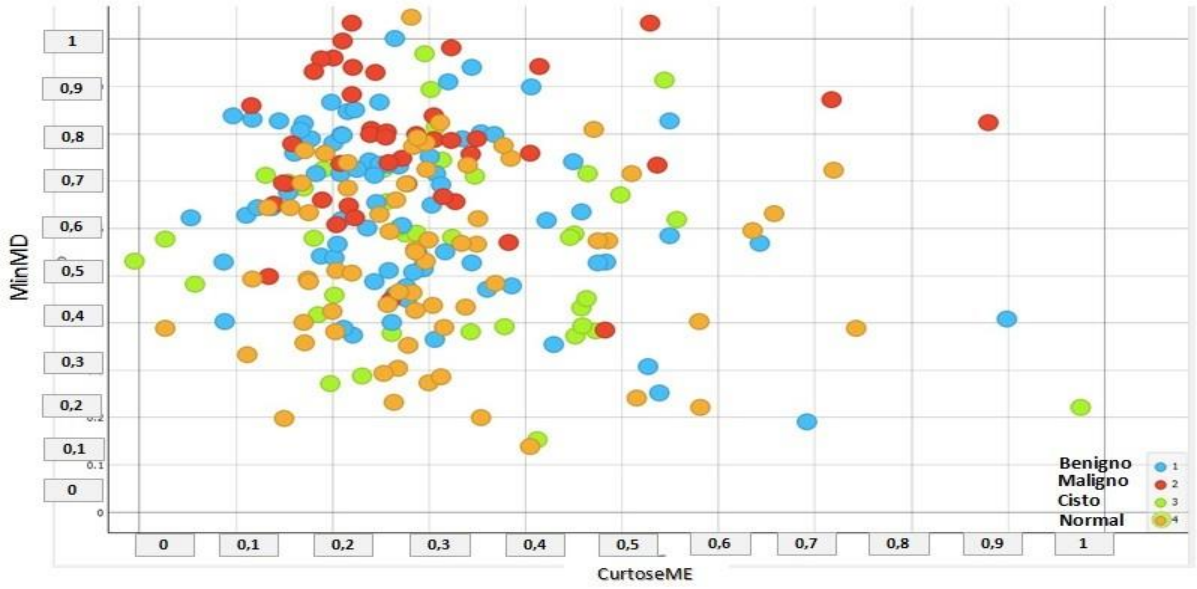
Tabela 6– Resultado da melhor performance de cada técnica via segmentação automática

Técnica	**Acurácia**	**Sensibilidade Maligno**	**Precisão**	**AUC**	**F1**
Normal	33,3%	37,5%	33,2%	56,4%	32,9%
SMOTE	48,4%	43,8%	49,4%	65,9%	48,4%
SMOTE + PCA 8	51,6%	31,2%	57,5%	57,8%	50,5%
SMOTE + Rank 4	**53,1%**	**43,8%**	**56,0%**	**71,9%**	**53,1%**
SMOTE + PSO	50,0%	37,5%	51,1%	70,9%	49,2%

5.1.2 Base 2

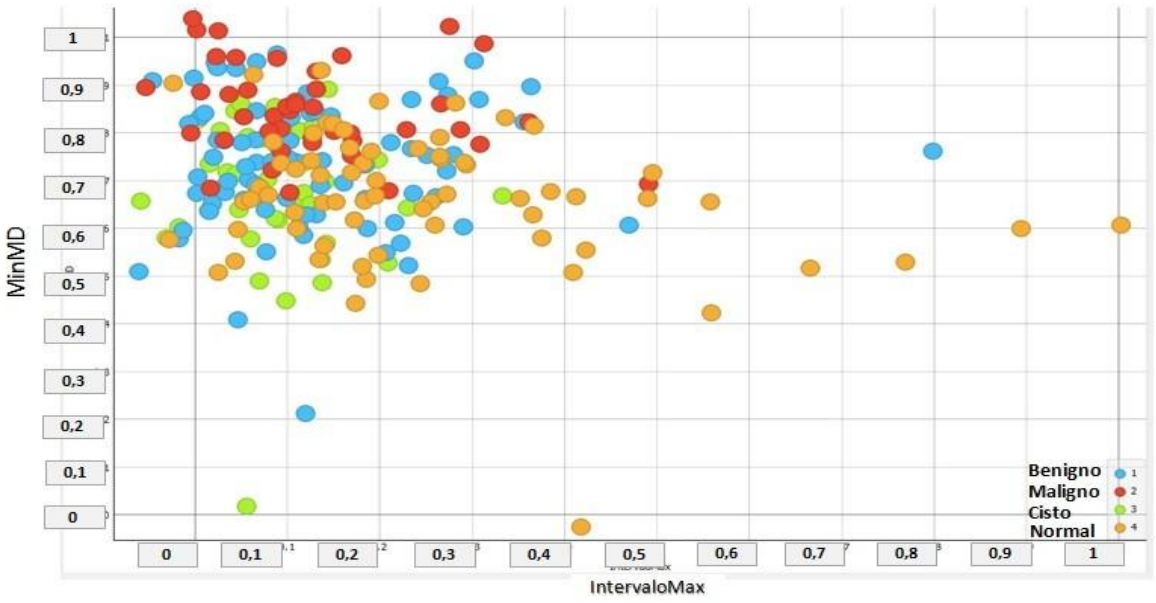
Assim como na Base 1, as Figuras 48 e 49 mostram as instâncias das quatroclasses para a Base 2 em espaços bidimensionais de características. Um ponto a ser observado nessas duas imagens é que, embora a quantidade de instâncias sejabem maior e haja uma grande sobreposição, as classes nesse caso estão mais bem agrupadas que na Base 1, o que facilita a classificação pelos algoritmos.

Figura 48– Espaço de características, via segmentação manual, para a Base 2



Fonte: Próprio autor.

Figura 49 – Espaço de características, via segmentação automática, para a Base 2



Fonte: Próprio autor.

Para a Base 2, o algoritmo de comparação, aquele que trouxe os melhoresresultados, foi o KNN com número de vizinhos igual a 4. Nas Tabelas 6 e 7 pode-se observar que a técnica **SMOTE** foi a de maior destaque, e os melhores resultados foram bastante similares. É importante frisar que, como a técnica CNN trouxe resultados muito ruins, ela não foi utilizada em adição com as outras.

Como previamente analisado nas Figuras 48 e 49, a performance da classificação, com as mesmas métricas, para a Base 2 foi maior que com a Base 1. Levando em consideração que são quatro classes, uma acurácia de 65% aliada a uma sensibilidadeao maligno de 92,2% se mostra satisfatória, pois o mais importante nesse tipo de classificação é que haja um baixo número de falsos negativos à Classe Maligna. Dado que, caso o paciente esteja com o tumor maligno, ele seja diagnosticado corretamentee possa iniciar o tratamento precoce.

Tabela 7– Resultado da melhor performance de cada técnica via segmentação manual

Técnica	**Acurácia**	**Sensibilidade Maligno**	**Precisão**	**AUC**	**F1**
Normal	48,0%	61,4%	46,9%	69,4%	69,4%
SMOTE	**65,6%**	**92,2%**	**62,0%**	**80,9%**	**63,6%**
CNN	26,2%	33,3%	58,8%	54,2%	26,2%
SMOTE + PCA 8	60,7%	85,7%	58,8%	81,2%	59,1%
SMOTE + Rank 4	49,4%	72,7%	47,9%	75,9%	48,4%
SMOTE + PSO	61,0%	85,7%	60,3%	80,3%	59,7%

Tabela 8– Resultado da melhor performance de cada técnica via segmentação automática

Técnica	**Acurácia**	**Sensibilidade Maligno**	**Precisão**	**AUC**	**F1**
Normal	40,3%	44,2%	40,9%	64,9%	40,6%
SMOTE	**63,1%**	**91,0%**	**61,8%**	**81,6%**	**59,6%**
CNN	30,1%	30,8%	31,7%	60,0%	30,1%
SMOTE + PCA 8	63,1%	87,2%	61,2%	82,0%	60,1%
SMOTE + Rank 4	51,6%	76,9%	49,7%	73,7%	50,3%
SMOTE + PSO	62,8%	88,5%	61,0%	81,1%	60,5%

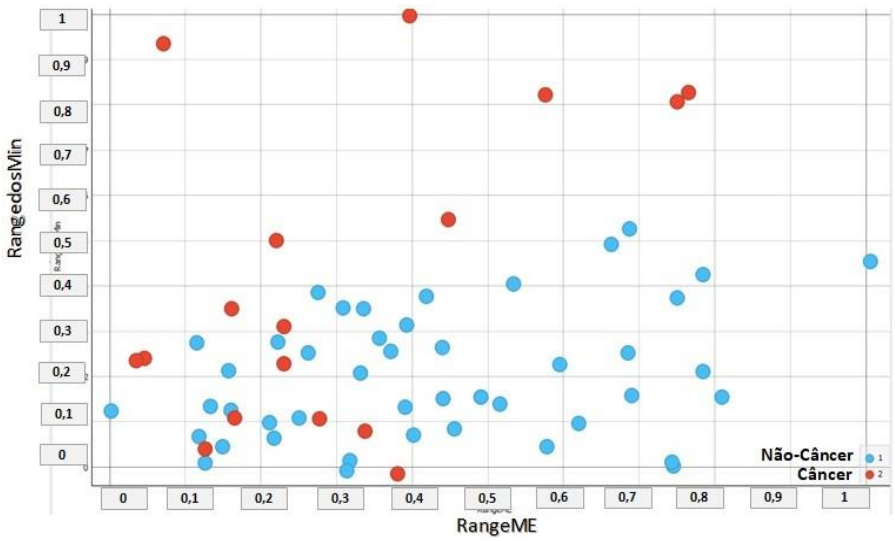
5.2 Resultados obtidos através da Metodologia de Classificação Binária

5.2.1 Base 1

Na Metodologia 2, que utilizou apenas duas classes, já era esperada uma classificação mais favorável em todas as métricas. Essa afirmação pode ser observadanas Figuras 50 e 51, que são os espaços de características para classificação binária. Aclasse em azul, se refere ao "não-câncer"e está bem mais agrupada. Diferentemente da Metodologia 1, onde havia quatro classes

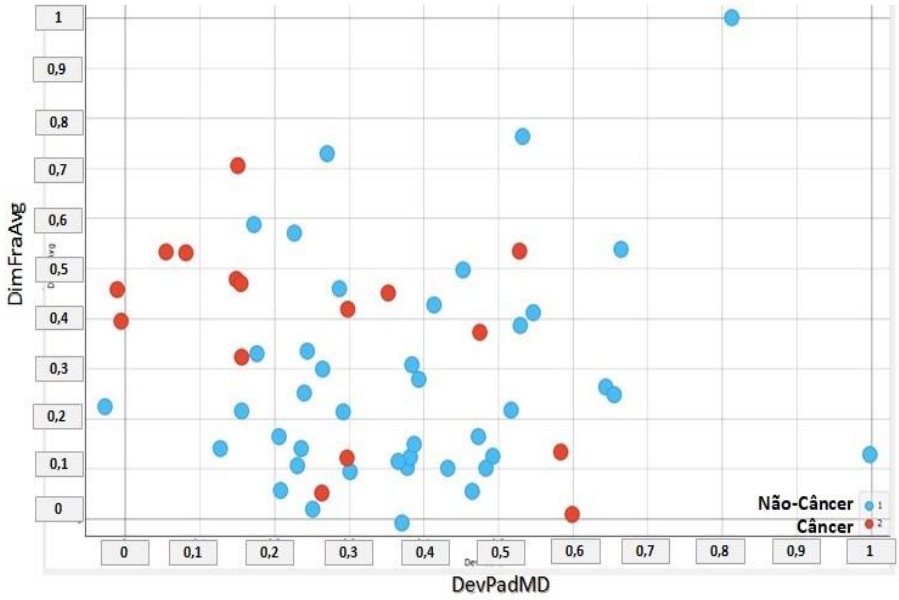
com quantidades parecidas, nessa situaçãoé perceptível que a classe "câncer"em vermelho está com bem menos instâncias, algoque indicava que a técnica de balanceamento teria grande importância na melhoria daperformance.

Figura 50– Espaço de características na classificação binária, via segmentação manual, para a Base 1



Fonte: Próprio autor.

Figura 51– Espaço de características na classificação binária, via segmentação automática para a Base 1



Fonte: Próprio autor.

Nas Tabelas 9 e 10, assim como na Metodologia 1, é possível observar que a performance vinda da segmentação manual feita por NOVA (2017) se sobrepõe sobre a feita

pelo autor do presente trabalho. Nessa situação a diferença nas métricas é sensivelmente maior que na metodologia anterior, além de ter alcançado uma classificação melhor que via segmentação automática, na Tabela 11.

Para esses três casos, o algoritmo de classificação KNN com 4 vizinhos foio que obteve os melhores resultados e utilizado para comparação. Como previsto anteriormente, após a utilização da técnica **SMOTE** houve um grande salto na qualidade da classificação, com as acurácias e sensibilidade à classe maligna aumentando cercade 15% e 40% respectivamente. De forma diferente, com conceito oposto ao SMOTE,a técnica CNN mais uma vez trouxe resultados não satisfatórios. Analisando os valoresem si, é notável perceber a grande capacidade de acerto na classificação binária, coma segmentação manual (feita por NOVA (2017)), na Tabela 9, tendo alcançado valoresem todas as métricas acima de 90%.

Um ponto de explanação sobre os resultados com a segmentação automática serem inferiores, do que com a utilização da segmentação manual, é pelo fato da automática não captar 100% da região de interesse para mamas de diferentes anatomias mamárias do padrão estabelecido no construção do algoritmo da segmentação automática. Com isso, algumas informações de características podem ser perdidas,ocasionando numa pior caracterização da mama e gerando um classificação inferior.

Tabela 9- Resultado da melhor performance de cada técnica via segmentação manual

Técnica	**Acurácia**	**Sensibilidade Maligno**	**Precisão**	**AUC**	**F1**
Normal	71,0%	52,9%	72,1%	72,9%	71,4%
SMOTE	84,4%	97,8%	87,1%	88,3%	84,2%
CNN	61,8%	64,7%	61,8%	65,4%	61,7%
SMOTE + PCA 8	**85,6%**	**95,6%**	**87,1%**	**90,4%**	**85,4%**
SMOTE + Rank 4	81,1%	91,1%	82,4%	89,2%	80,9%
SMOTE + PSO	84,4%	93,3%	85,6%	88,7%	84,3%

Tabela 10– Resultado da melhor performance de cada técnica via segmentação manual feita por NOVA (2017)

Técnica	**Acurácia**	**Sensibilidade Maligno**	**Precisão**	**AUC**	**F1**
Normal	77,0%	56,2%	72,1%	77,0%	81,2%
SMOTE	**92,2%**	**97,8%**	**91,0%**	**95,4%**	**92,0%**
CNN	72,7%	62,5%	73,4%	71,7%	72,4%
SMOTE + PCA 8	83,3%	93,3%	84,7%	87,7%	83,2%
SMOTE + Rank 4	80,0%	84,4%	80,2%	90,0%	80,0%
SMOTE + PSO	92,2%	97,8%	92,7%	95,4%	92,2%

Tabela 11– Resultado da melhor performance de cada técnica via segmentação automática

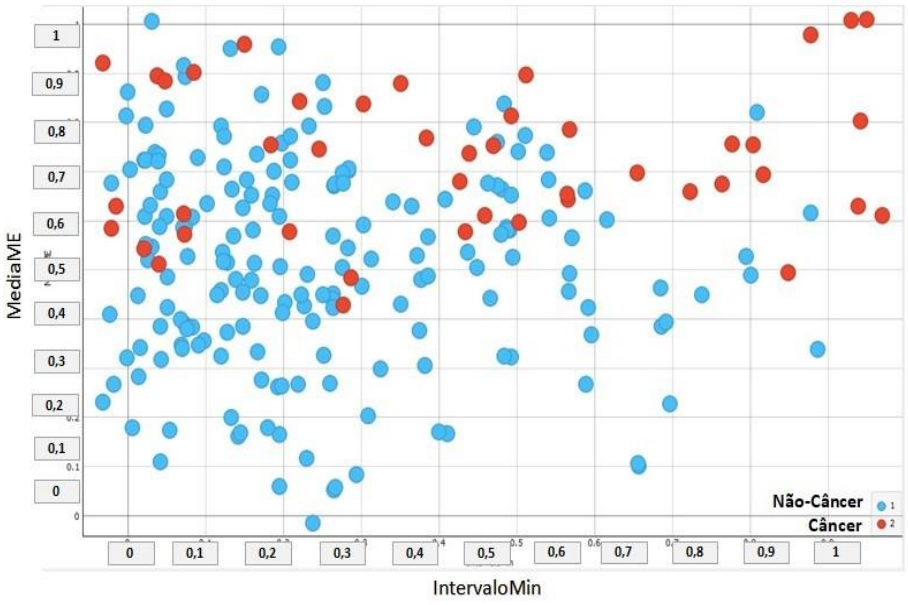
Técnica	Acurácia	Sensibilidade Maligno	Precisão	AUC	F1
Normal	77,2%	43,8%	75,7%	66,0%	75,7%
SMOTE	84,1%	87,8%	84,3%	86,3%	84,1%
CNN	63,3%	56,2%	64,4%	70,1%	63,2%
SMOTE + PCA 8	**84,1%**	**92,7%**	**85,2%**	**93,4%**	**84,0%**
SMOTE + Rank 4	82,9%	92,6%	83,6%	90,0%	82,8%
SMOTE + PSO	82,9%	92,7%	84,2%	90,7%	82,8%

5.2.2 Base 2

Assim como nas situações passadas, as Figuras 52 e 53 mostram o espaço de características para os dois tipos de segmentação. Um fato curioso de ser observado éque os valores das características extraídos para os dois casos foram tão parecidos que seu espaço de características, com a melhor combinação de duas delas, ficou igual.

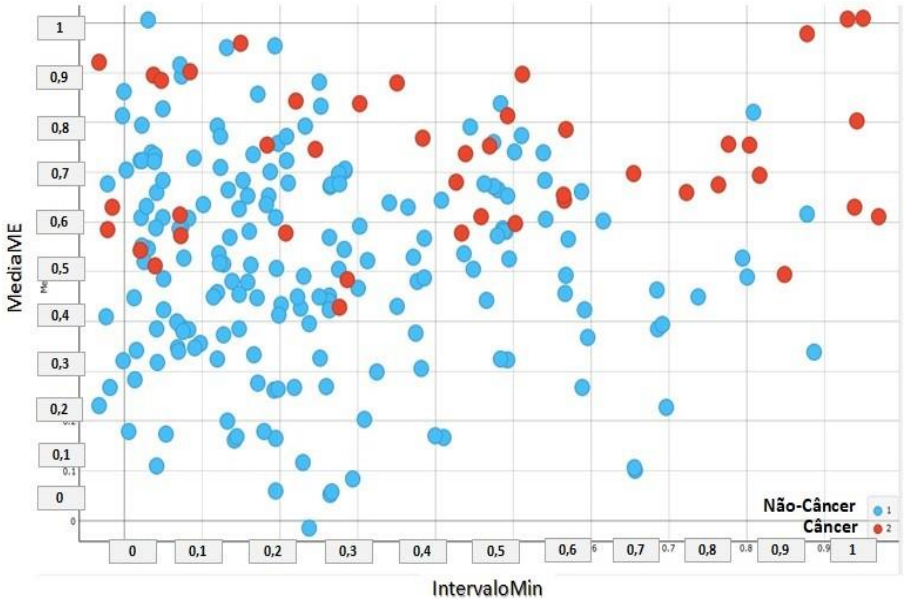
De forma similar à Base 1, há uma maior facilidade de classificação pelo fato deum maior agrupamento da classe "não-câncer", em azul. Entretanto, se torna ainda mais latente a necessidade de utilização de técnicas de balanceamento, pois nessecaso com um maior número de instâncias no geral, a desigualdade numérica entre asclasses se torna ainda mais danosa à classificação, principalmente à sensibilidade àclasse maligna.

Figura 52– Espaço de características na classificação binária, via segmentação manual, para a Base 2



Fonte: Próprio autor.

Figura 53– Espaço de características na classificação binária, via segmentação automática, para a Base 2



Fonte: Próprio autor.

Finalmente, nas Tabelas 12 e 13 estão expostos os melhores resultados para a Metodologia 2 na Base 2. Nesse caso, assim como no anterior, os melhores resultados foram alcançados com o algoritmo KNN com 4 vizinhos.

De forma geral, é possível notar os excelentes resultados em todas as métricas nos dois tipos de segmentação, com destaque para a sensibilidade à classe maligna que alcançou mais de 99% em ambas classificações. Além disso, a utilização das técnicas de otimização mais uma vez proporcionou um aumento nas taxas de acerto das métricas, com a sensibilidade à classe maligna sendo o maior exemplo desseacréscimo.

Tabela 12– Resultado da melhor performance de cada técnica via segmentação manual

Técnica	**Acurácia**	**Sensibilidade Maligno**	**Precisão**	**AUC**	**F1**
Normal	86,0%	54,5%	85,2%	83,4%	85,5%
SMOTE	**90,8%**	**100%**	**92,2%**	**94,7%**	**90,7%**
CNN	58,3%	65,8%	59,1%	59,4%	58,3%
SMOTE + PCA 8	89,2%	98,9%	90,7%	96,4%	89,1%
SMOTE + Rank 4	82,7%	90,3%	83,5%	90,9%	82,6%
SMOTE + PSO	90,8%	99,0%	92,2%	94,0%	90,5%

Tabela 13– Resultado da melhor performance de cada técnica via segmentação automática

Técnica	Acurácia	Sensibilidade Maligno	Precisão	AUC	F1
Normal	86,0%	54,5%	85,2%	83,4%	85,5%
SMOTE	90,8%	98,9%	92,1%	96,3%	90,7%
CNN	64,4%	61,4%	60,4%	64,4%	60,2%
SMOTE + PCA 8	**96,2%**	**99,5%**	**92,9%**	**91,3%**	**91,8%**
SMOTE + Rank 4	90,2%	91,4%	85,5%	90,2%	84,8%
SMOTE + PSO	90,3%	98,9%	91,5%	96,2%	90,2%

5.3 Análise estatística

Para finalizar a análise da presente dissertação, foram realizados testes de hipótese bilateral e de Wilcoxon pareado entre os melhores resultados de cada metodologia/base. Esses testes tiveram como objetivo avaliar os resultados entre as segmentações manual e automática e entre os usuários diferentes na segmentação manual.

Foi admitido, na hipótese nula, que as acurácias dos casos a serem testados não possuíam diferença estatisticamente significativa. Essa escolha foi feita com o pensamento de querer provar o contrário. Além disso, o nível de significância escolhidofoi de 5% e a tabela *t-student* foi utilizada no teste de hipótese por ser mais adequadaem situações onde quantidade de amostras é menor que 30. No presente trabalho, foram calculadas 10 amostras para cada bloco de melhor resultado.

Ao fazer a comparação entre dois blocos de resultados com 10 amostras cada,tem-se que o número de graus de liberdade na tabela *t-student* é dado por 18 (10+10-2). Observando a Tabela 14 é possível notar em destaque o valor (2,101) retirado damesma para as comparações no teste de hipótese.

Tabela 14– Tabela t-student para diferentes graus de liberdade e níveis de significância

Bicaudal	50%	60%	70%	80%	90%	95%	98%	99%
1	1,000	1,376	1,963	3,078	6,314	12,71	31,82	63,66
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878

Fonte: Próprio autor.

Como todos os resultados foram calculados utilizando-se a validação cruzada com leave-one-out, os valores encontrados se mantiveram praticamente iguais em cada uma das 10 amostras, com alteração do valor apenas na quinta casa decimal. Isso implicou que o desvio padrão (σ) em todas as situações esteve baixíssimo, levando a Equação (4.11) tender a "0", e a Equação (4.10) tender ao "infinito".

Na Tabela 15, tem-se como exemplo, os valores calculados para o caso da comparação entre os dois tipos de segmentação com a Metodologia 1 e Base 2. E, como previsto, o valor a ser comparado com o encontrado na t-student foi extremamente alto (6154,07).

Com o valor de t_0 sendo muito grande, ele facilmente supera o valor de 2,101 retirado da tabela *t-student*. Logo, rejeitou-se a hipótese nula em todos os casos, e todas as alegações anteriores, acerca da influência do tipo da segmentação e do usuário na segmentação manual, puderam ser confirmados nesse trabalho.

O mesmo ocorre para o teste de Wilcoxon, pois com os resultados de comparação se mantendo com os mesmos valores (diferença quinta casa decimal), faz com queo teste de sinais mantenha em 0 o valor de vezes que o sinal menos frequente ocorre.Com isso, o cálculo do binomial (e do *p-value*) tende a 0, sendo 0,000976 em um doscasos. Consequentemente, a

hipótese nula com 5% de significância é rejeitada.

A segmentação manual obteve o melhor resultado em 75% dos casos se comparada com a segmentação automática, e a segmentação manual feita por NOVA (2017) em 100% dos casos trouxe melhores resultados.

Tabela 15– Valores calculados para teste de hipótese entre os dois tipos de segmentação utilizando Metodologia 1 e Base 2.

$\sigma Manual$	$\sigma Auto$	Sp^2	X_{Manual}	X_{Auto}	t_0
$9 \cdot 10^{-4}$	$8 \cdot 10^{-4}$	$8,2 \cdot 10^{-7}$	65,605	61,101	6154,07

6 CONCLUSÃO E TRABALHOS FUTUROS

Levando em consideração os resultados alcançados, pode-se afirmar que a termografia infravermelha pode auxiliar tanto no diagnóstico de anomalias mamárias como para um rastreamento inicial das mesmas. A abordagem desenvolvida neste trabalho teve como ponto de partida a forma da segmentação das imagens termográficaspara o melhor recorte da região de interesse, podendo essa segmentação ser feita naforma manual e na automática. Em seguida, a extração de características estatísticas e fractais foi estudada para que uma melhor caracterização de padrões dos diversos tipos de anomalias pudessem ser identificados.

Embora as duas bases de dados utilizadas apresentassem graus de desbalanceamento dos dados distintos, a partir da abordagem e da base utilizada, a técnica deotimização como criação de vetores sintéticos (SMOTE) foi incluída para equilibrar a quantidade de instâncias por classe e trouxe uma grade melhoria na performance dosresultados. Dentre as métricas que melhoraram com o balanceamento, a sensibilidadeà classe maligna deve ser destacada por ela ser de vital importância para o início de umtratamento precoce. Outras técnicas de otimização visando redução e transformação de características também foram incluídas, tendo um impacto positivo na melhoria dosresultados.

Dentre os cinco classificadores utilizados, o KNN e a ANN tiveram grande destaque para ambas as metodologias e bases, estando sempre como a melhor ousegunda melhor performance nas métricas usadas. Na metodologia que empregou a classificação binária (Câncer/Não-Câncer) os melhores resultados alcançaram 96,2% e 99,5% na acurácia e sensibilidade à classe maligna, respectivamente. Para a clas- sificação envolvendo quatro classes (Maligno, Benigno, Cisto e Normal) os resultado mais expressivos trouxeram 65,6% e 92,2% na acurácia e sensibilidade à classe ma- ligna, respectivamente. Essa grande diferença nos resultados já era esperada pelo fatode haver mais tipos de classes a serem classificados e algumas instâncias estarem se sobrepondo no espaço de características. Entretanto, se comparado com outrostrabalhos de classificação mamária envolvendo 4 classes, como por exemplo em VASCONCELOS *et al.* (2017) que obteve 63,46% na acurácia e 86,54% na sensibilidade àclasse maligna, o resultado do trabalho atual se mostrou bastante pertinente.

No desenvolvimento do presente trabalho, foi usada a plataforma computacional aberta Orange Canvas. Grande parte do trabalho foi desenvolvida na mesma, incluindo a classificação, a ferramenta se mostrou de grande valia para que futuros usuários possam fazer análises que auxiliem ao diagnóstico médico.

Por fim, foram realizados testes de hipótese para verificar a significância estatística dos resultados, e com isso poder ratificar algumas observações vistas nos resultados. Tendo em vista que as hipóteses nulas foram rejeitadas, é possível afirmarque em 75% dos casos a segmentação manual trouxe melhores resultados que a automática, e que nas duas metodologias a segmentação manual feita especificamente por NOVA (2017) obteve melhores resultados se comparada com a segmentação manual feita pelo presente autor.

Para continuidade desta linha de investigação, pode-se sugerir:

- Extrair e testar na classificação outros tipos de características, como as morfológicas ou geoestatísticas.
- Aperfeiçoar os estudos nos hiperparâmetros dos classificadores KNN e ANN pelo fato deles trazerem os melhores resultados para o tipo de classificação analisada.
- Utilizar diferentes bases de dados para obter uma maior confirmação da influênciado tipo de segmentação na classificação.
- Integrar o processo de extração de características na plataforma Orange Canvas a fim de centralizar o processo.

REFERÊNCIAS

ACHARYA, U. R.; NG, E. Y.-K.; TAN, J.-H.; SREE, S. V. Thermography based breast cancer detection using texture features and support vector machine. *Journal of medical systems*, Springer, v. 36, n. 3, p. 1503–1510, 2012.

ACS. *American Cancer Society*. 2020. Disponível em: <https://www.cancer.org/cancer/ breast-cancer.html>. Acesso em: 03 Mar. 2020.

ALMEIDA, T. B. *et al. Seleção de atributos usando abordagem Wrapper para classi- ficação hierárquica multirrótulo*. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2018.

ANDREOLA, R. Support vector machines na classificação de imagens hiperespectrais. 2009.

ANTONIAZZI, R. L. Aplicação do método box counting para a estimativa da dimensão fractal de figuras planas digitalizadas. Universidade Federal de Santa Maria, 2007.

ARAÚJO, M. C. de; QUEIROZ, K. F. F. da C.; SOUZA, R. M. C. R. de; LIMA, R. d. C. F. de. Applications of the use of infrared breast images: Segmentation and classificationof breast abnormalities. In: *Biomedical Computing for Breast Cancer Detection and Diagnosis*. [S.l.]: IGI Global, 2021. p. 211–229.

ARAÚJO, M. Costa de. *Utilização de câmera por infravermelho para avaliação de diferentes patologias em clima tropical e uso conjunto de sistemas de banco de dadospara detecção de câncer de mama*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2009.

ARAÚJO, M. Costa de. *Uso de imagens termográficas para classificação de anorma- lidades de mama baseado em variáveis simbólicas intervalares*. Tese (Doutorado) —Universidade Federal de Pernambuco, 2014.

ARAÚJO, N. D. de; FARIAS, R. P. de; PEREIRA, P. B.; FIGUEIRÊDO, F. M. de; MORAIS, A. M. B. de; SALDANHA, L. C.; GABRIEL, J. E. A era da bioinformática: seu potencial esuas implicações para as ciências da saúde. *Estudos de biologia*, v. 30, n. 70/72, 2008.

ASSIS, J. V. d. *Estudo do grau de risco de câncer de mama utilizando a dimensão fractal em imagens infravermelhas*. Dissertação (B.S. thesis) — Universidade Tecnológica Federal do Paraná, 2015.

BAFFA, M.; CHELONI, D.; LATTARI, L. Segmentação automática de imagens térmicas das mamas utilizando limiarização com refinamento adaptativo. In: *Anais Principais do XVI Workshop de Informática Médica*. Porto Alegre, RS, Brasil: SBC, 2016. p. 39–48. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/9896>.

BAFFA, M. d. F. O.; LATTARI, L. G. Convolutional neural networks for static and dynamicbreast infrared imaging classification. In: IEEE. *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.], 2018. p. 174–181.

BAGAVATHIAPPAN, S.; PHILIP, J.; JAYAKUMAR, T.; RAJ, B.; RAO, P. N. S.; VARALAKSHMI, M.; MOHAN, V. Correlation between plantar foot temperature and diabetic neuropathy: a case study by using an infrared thermal imaging technique. *Journal of diabetes*

science and technology, SAGE Publications, v. 4, n. 6, p. 1386–1392, 2010.

BARATLOO, A.; HOSSEINI, M.; NEGIDA, A.; ASHAL, G. E. Part 1: simple definition and calculation of accuracy, sensitivity and specificity. ARCHIVES OF ACADEMIC EMERGENCY MEDICINE (EMERGENCY), 2015.

BARELLA, V. H. *Técnicas para o problema de dados desbalanceados em classificação hierárquica*. Tese (Doutorado) — Universidade de São Paulo, 2016.

BENKO, I.; KOTELES, G.; NEMETH, G. Thermal imaging of the effects of betairradiationon human body surfaces. In: *Proceeding of the Conference on Quantitative Infrared Thermography (QIRT'96), Eurotherm Series*. [S.l.: s.n.], 1996. v. 50, p. 354–359.

BERGH, F. Van den. An analysis of particle swarm optimizers [ph. d. thesis] pretoria. *South Africa: University of Pretoria*, 2001.

BEZERRA, L. A. *et al. Uso de imagens termográficas em tumores mamários para validação de simulação computacional*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2007.

BHOWAN, U.; JOHNSTON, M.; ZHANG, M.; YAO, X. Evolving diverse ensembles using genetic programming for classification with unbalanced data. *IEEE Transactions on Evolutionary Computation*, IEEE, v. 17, n. 3, p. 368–386, 2012.

BIJALWAN, V.; KUMAR, V.; KUMARI, P.; PASCUAL, J. Knn based machine learning approach for text and document mining. *International Journal of Database Theory and Application*, v. 7, n. 1, p. 61–70, 2014.

BORCHARTT, T. Análise de imagens termográficas para a classificação de alterações na mama. *Master's thesis, Instituto de Computação, Universidade Federal Fluminense,Niterói, RJ, Brasil*, 2013.

BOUZIDA, N.; BENDADA, A.; MALDAGUE, X. P. Visualization of body thermoregulation by infrared imaging. *Journal of Thermal Biology*, Elsevier, v. 34, n. 3, p. 120–126, 2009.

BRAZ, J. R. C. Fisiologia da termorregulação normal. *Revista Neurociências*, v. 13, p. 12–17, 2005.

BRONZINO, J. D. *Medical devices and systems*. [S.l.]: CRC Press, 2006.

CASTANEDO, C. I. Quantitative subsurface defect evaluation by pulsed phase thermo-graphy: depth retrieval with the phase. 2005.

CATTANEO, C.; GIANCAMILLO, A. D.; CAMPARI, O.; ORTHMANN, N.; MARTRILLE,L.; DOMENEGHINI, C.; JOUINEAU, C.; BACCINO, E. Infrared tympanic thermography as a substitute for a probe in the evaluation of ear temperature for post-mortem interval determination: a pilot study. *Journal of Forensic and Legal medicine*, Elsevier, v. 16, n. 4, p. 215–217, 2009.

CERVANTE, L.; XUE, B.; ZHANG, M.; SHANG, L. Binary particle swarm optimisation for feature selection: A filter based approach. In: IEEE. *2012 IEEE Congress on Evolutionary Computation*. [S.l.], 2012. p. 1–8.

CHAVAN, S.; ADGOKAR, N. P. *et al.* An overview on particle swarm optimization: basic

concepts and modified variants. *International Journal of Science and Research*, v. 4, n. 5, p. 255–260, 2015.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002.

CHUDECKA, M.; LUBKOWSKA, A. The use of thermal imaging to evaluate body temperature changes of athletes during training and a study on the impact of physiologicaland morphological factors on skin temperature. *Human movement*, Versita, v. 13, n. 1,p. 33–39, 2012.

CHUDECKA, M.; LUBKOWSKA, A. Thermal maps of young women and men. *Infrared Physics & Technology*, Elsevier, v. 69, p. 81–87, 2015.

COCKBURN, M. W. *The Truth About Breast Thermography*. 2020. Disponível em: <https: //www.healingwell.com/articles/post/the-truth-about-breast-thermography>. Acesso em:09 Mar. 2020.

COVÕES, T. F. *Seleção de atributos via agrupamento*. Tese (Doutorado) — Universidadede São Paulo, 2010.

DARA, S.; BANKA, H. A binary pso feature selection algorithm for gene expression data. In: IEEE. *2014 International Conference on Advances in Communication and Computing Technologies (ICACACT 2014)*. [S.l.], 2014. p. 1–6.

DATACAMP. *KNN Classification using Scikit-learn*. 2020. Disponível em: <https://www. datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>. Acesso em: 27 Out. 2020.

DEMŠAR, J.; ZUPAN, B.; LEBAN, G.; CURK, T. Orange: From experimental machine learning to interactive data mining. In: SPRINGER. *European conference on principlesof data mining and knowledge discovery*. [S.l.], 2004. p. 537–539.

DERIS, A. M.; ZAIN, A. M.; SALLEHUDDIN, R. Overview of support vector machine in modeling machining performances. *Procedia Engineering*, Elsevier, v. 24, p. 308–312,2011.

DIDATICA. *O pacote Caret – linguagem R*. 2020. Disponível em: <https://didatica.tech/ o-pacote-caret-linguagem-r/>. Acesso em: 26 Out. 2020.

DIDATICA. *O que é e como funciona o algoritmo RandomForest*. 2020. Disponível em: <https://didatica.tech/o-que-e-e-como-funciona-o-algoritmo-randomforest/>. Acesso em: 29 Out. 2020.

DOUGHERTY, G. *Digital image processing for medical applications*. [S.l.]: Cambridge University Press, 2009.

DOUKAS, C.; MAGLOGIANNIS, I. Region of interest coding techniques for medical image compression. *IEEE Engineering in medicine and Biology Magazine*, IEEE, v. 26,n. 5, p. 29–35, 2007.

DOURADO, H. d. M. *Segmentação e análise automáticas de termogramas: um métodoauxiliar na detecção do câncer de mama*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2014.

DUDA, R. O.; HART, P. E. *et al. Pattern classification and scene analysis*. [S.l.]: WileyNew York, 1973. v. 3.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. [S.l.]: John Wiley & Sons, 2012.

EITRICH, T.; KLESS, A.; DRUSKA, C.; MEYER, W.; GROTENDORST, J.Classification of highly unbalanced cyp450 data of drugs using cost sensitive machine learning techniques. *Journal of chemical information and modeling*, ACS Publications, v. 47, n. 1,p. 92–103, 2007.

ETEHADTAVAKOL, M.; NG, E.; CHANDRAN, V.; RABBANI, H. Separable and non-separable discrete wavelet transform based texture features and image classification ofbreast thermograms. *Infrared Physics & Technology*, Elsevier, v. 61, p. 274–286, 2013.

FILHO, A. S. Mensagem aos médicos: Temas de oncologia–metástase. *Ministério da Saúde, Secretaria Nacional de Saúde, Divisão Nacional de Câncer, Serviço de Programação e Orientação Técnica*, 1976.

FLIR. *Caméra infrarouge FLIR ThermaCAM Série P (P620,P640, P660)*. 2020. Disponível em: <https://medium.com/data-hackers/ como-lidar-com-dados-desbalanceados-em-problemas-de-classificaçao-17c4d4357ef9>. Acesso em: 12 Ago. 2020.

FREITAS, T. E. d. S. *Uso conjunto de variáveis intervalares e variáveis clássicas para classificação de termogramas de mama por meio da distância de mahalanobis*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2021.

GAO, X.; WANG, B.; TAO, D.; LI, X. A relay level set method for automatic image segmentation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, IEEE, v. 41, n. 2, p. 518–525, 2010.

GENG, X.; LIU, T.-Y.; QIN, T.; LI, H. Feature selection for ranking. In: *Proceedings ofthe 30th annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.: s.n.], 2007. p. 407–414.

GOLESTANI, N.; ETEHADTAVAKOL, M.; NG, E. Level set method for segmentationof infrared breast thermograms. *EXCLI journal*, Leibniz Research Centre for Working Environment and Human Factors, v. 13, p. 241, 2014.

GONÇALVES, C. B. *et al.* Detecção de câncer de mama utilizando imagens termográfi-cas. Universidade Federal de Uberlândia, 2017.

H. DO CANCER DE BARRETOS. *Saiba quais são os tipos de crâncer mais comuns nobrasil.* 2015. Disponível em: <https://www.hcancerbarretos.com.br/>.

HAN, H.; WANG, W.-Y.; MAO, B.-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: SPRINGER. *International conference on intelligent computing*. [S.l.], 2005. p. 878–887.

HARDY, J. D.; MUSCHENHEIM, C. Radiation of heat from the human body. v. the transmission of infra-red radiation through skin. *The Journal of clinical investigation*, Am Soc Clin Investig, v. 15, n. 1, p. 1–9, 1936.

HART, P. The condensed nearest neighbor rule (corresp.). *IEEE transactions on information*

theory, Citeseer, v. 14, n. 3, p. 515–516, 1968.

HAYKIN, S. *Redes neurais: princípios e prática*. [S.l.]: Bookman Editora, 2007.

HAYWARD, J. Uma visao global do problema. *Alternativas Diagnósticas e Terapêuticasno Câncer de Mama*, p. 3–11, 1987.

HOLLAND, S. M. Principal components analysis (pca). *Department of Geology, University of Georgia, Athens, GA*, p. 30602–2501, 2008.

HOSMER, D.; LEMESHOW, S. Applied logistic regression wiley & sons. *New York*, 1989.

INCA. *Instituto Nacional do Cancer*. 2020. Disponível em: <https://www.inca.gov.br/ tipos-de-cancer/cancer-de-mama>. Acesso em: 03 Mar. 2020.

JURBERG, C.; GOUVEIA, M. E.; BELISÁRIO, C. Na mira do câncer: o papel da mídia brasileira. *Revista Brasileira de Cancerologia*, v. 52, n. 2, p. 139–146, 2006.

KANDLIKAR, S. G.; PEREZ-RAYA, I.; RAGHUPATHI, P. A.; GONZALEZ-HERNANDEZ,J.-L.; DABYDEEN, D.; MEDEIROS, L.; PHATAK, P. Infrared imaging technology for breast cancer detection–current status, protocols and new directions. *International Journal of Heat and Mass Transfer*, Elsevier, v. 108, p. 2303–2320, 2017.

KASPRZYK-KUCEWICZ, T.; CHOLEWKA, A.; BAŁAMUT, K.; KOWNACKI, P.; KAS- ZUBA, N.; KASZUBA, M.; STANEK, A.; SIEROŃ, K.; STRANSKY, J.; PASZ, A. *et al.*The applications of infrared thermography in surgical removal of retained teeth effects assessment. *Journal of Thermal Analysis and Calorimetry*, Springer, p. 1–6, 2020.

KASSAMBARA, A. *Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra*. [S.l.]: STHDA, 2017. v. 2.

KENNEDY, J.; EBERHART, R. Particle swarm optimization. In: IEEE. *Proceedings of ICNN'95-International Conference on Neural Networks*. [S.l.], 1995. v. 4, p. 1942–1948.

KOAY, J.; HERRY, C.; FRIZE, M. Analysis of breast thermography with an artificial neural network. In: IEEE. *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. [S.l.], 2004. v. 1, p. 1159–1162.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning:A review of classification techniques. *Emerging artificial intelligence applications incomputer engineering*, Amsterdam, v. 160, n. 1, p. 3–24, 2007.

KUBAT, M.; MATWIN, S. *et al.* Addressing the curse of imbalanced training sets: one- sided selection. In: CITESEER. *Icml*. [S.l.], 1997. v. 97, p. 179–186.

LAHIRI, B.; BAGAVATHIAPPAN, S.; JAYAKUMAR, T.; PHILIP, J. Medical applications of infrared thermography: a review. *Infrared Physics & Technology*, Elsevier, v. 55, n. 4, p. 221–235, 2012.

LAWSON, R. Implications of surface temperatures in the diagnosis of breast cancer. *Canadian Medical Association Journal*, Canadian Medical Association, v. 75, n. 4, p. 309, 1956.

LELES, A. C. Q. *et al.* Desenvolvimento de procedimento e análise de imagens térmicaspara a identificação do câncer de mama. Universidade Federal de Uberlândia, 2015.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007.

MAALOUF, M. Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, Inderscience Publishers Ltd, v. 3, n. 3, p.281–299, 2011.

MADHU, H.; KAKILETI, S. T.; VENKATARAMANI, K.; JABBIREDDY, S. Extraction of medically interpretable features for classification of malignancy in breast thermography. In: IEEE. *2016 38th Annual International Conference of the IEEE Engineering in Medicineand Biology Society (EMBC)*. [S.l.], 2016. p. 1062–1065.

MAHMOUDZADEH, E.; MONTAZERI, M.; ZEKRI, M.; SADRI, S. Extended hidden markov model for optimized segmentation of breast thermography images. *Infrared Physics & Technology*, Elsevier, v. 72, p. 19–28, 2015.

MANDELBROT, B. B.; MANDELBROT, B. B. *The fractal geometry of nature*. [S.l.]: WH freeman New York, 1982. v. 1.

MARINS, J. C. B.; FERNÁNDEZ-CUEVAS, I.; ARNAIZ-LASTRAS, J.; FERNANDES, A.; SILLERO-QUINTANA, M. Aplicaciones de la termografía infrarroja en el deporte:Una revisión. *Revista internacional de Medicina y Ciencias de la Actividad Física del Deporte*, Universidad Autónoma de Madrid. Comunidad Virtual Ciencias del Deporte, 2015.

MARQUES, R.; RESMINI, R.; CONCI, A.; LIMA, R.; FONTES, C. A. P. Método para segmentação manual de imagens térmicas para geração de ground truth. In: *XXXII Congresso da Sociedade Brasileira de Computacao, Curitiba-PR*. [S.l.: s.n.], 2012. p. 9.

MEDIUM. *Como lidar com dados desbalanceados em problemas de classificação*. 2019. Disponível em: <https://medium.com/data-hackers/ como-lidar-com-dados desbalanceados-em-problemas-de-classificaçao-17c4d4357ef9>. Acesso em: 19 Out. 2020.

MEDIUM. *Implementing the Particle Swarm Optimization (PSO) Algo- rithm in Python*. 2020. Disponível em: <https://medium.com/analytics-vidhya/ implementing-particle-swarm-optimization-pso-algorithm-in-python-9efc2eb179a6>. Acesso em: 01 Nov. 2020.

MILARÉ, C. R.; BATISTA, G. E.; CARVALHO, A. C. d. L. de. Avaliaç ao de uma abordagem hıbrida para aprender com classes desbalanceadas: Resultados experimentaiscom o indutor cn2. In: *IV Congresso da Academia Trinacional de Ciências*. [S.l.: s.n.],2009. v. 2, p. 15.

MILOSEVIC, M.; JANKOVIC, D.; PEULIC, A. Thermography based breast cancer detection using texture features and minimum variance quantization. *EXCLI journal*, LeibnizResearch Centre for Working Environment and Human Factors, v. 13, p. 1204, 2014.

MORE, A. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*, 2016.

MUŠTRA, M.; GRGIĆ, M.; DELAČ, K. Breast density classification using multiple feature selection. *automatika*, Taylor & Francis, v. 53, n. 4, p. 362–372, 2012.

NETO, H. d. M. D.; LIMA, R. d. C. F. de. Automated breast cancer detection via thermography. ENEBI, 2015.

NETO, O. P. S.; SILVA, A. C.; PAIVA, A. C.; GATTASS, M. Automatic mass detectionin mammography images using particle swarm optimization and functional diversityindexes. *Multimedia Tools and Applications*, Springer, v. 76, n. 18, p. 19263–19289, 2017.

NG, E.; KEE, E.; ACHARYA, U. R. Advanced technique in breast thermography analysis.In: IEEE. *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. [S.l.], 2006. p. 710–713.

NG, E.-K. A review of thermography as promising non-invasive detection modality for breast tumor. *International Journal of Thermal Sciences*, Elsevier, v. 48, n. 5, p. 849–859,2009.

NG, E. Y.; SUDHARSAN, N. Computer simulation in conjunction with medical thermography as an adjunct tool for early detection of breast cancer. *BMC cancer*, Springer,v. 4, n. 1, p. 17, 2004.

NOVA, R. d. L. V. *Uso de imagens termográficas de mama para análise de patologias através da comparação entre diversos classificadores estatísticos*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2017.

NUNES, A. P.; SILVA, A. C.; PAIVA, A. C. de. Detection of masses in mammographic images using simpson's diversity index in circular regions and svm. In: SPRINGER. *International Workshop on Machine Learning and Data Mining in Pattern Recognition*.[S.l.], 2009. p. 540–553.

OLIVERA, A. R. Comparação de algoritmos de aprendizagem de máquina para cons-trução de modelos preditivos de diabetes não diagnosticado. 2016.

OSHIRO, T. M. *Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica*. Tese (Doutorado) — Universidade de São Paulo, 2013.

PEREIRA, R. B.; CARVALHO, A. P. d.; ZADROZNY, B.; MERSCHMANN, L. H. d. C. Information gain feature selection for multi-label classification. 2015.

PORKODI, R. Comparison of filter based feature selection algorithms: An overview. *International journal of Innovative Research in Technology & Science*, v. 2, n. 2, p. 108–113, 2014.

POZZOLO, A. D.; CAELEN, O.; WATERSCHOOT, S.; BONTEMPI, G. Racing for unbalanced methods selection. In: SPRINGER. *International conference on intelligent data engineering and automated learning*. [S.l.], 2013. p. 24–31.

PRADO, B. B. F. d. Influência dos hábitos de vida no desenvolvimento do câncer. *Ciência e Cultura*, Sociedade Brasileira para o Progresso da Ciência, v. 66, n. 1, p.21–24, 2014.

PRATI, R. C. *Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos*. Tese (Doutorado) — Universidade deSão Paulo, 2006.

QUEIROZ, K.; ARAUJO, M.; LIMA, R. de. Análise da influência da segmentação sobreo problema da classificação de anomalias em imagens termográficas de mama. 2014.

QUEIROZ, K. F. F. d. C.; ARAÚJO, M. C. de; ESPÍNDOLA, N. A.; SANTOS, L. C.;SANTOS, F. G.; LIMA, R. d. C. F. de. Developing and using computational frameworksto conduct numerical analysis and calculate temperature profiles and to classify breast abnormalities. In: *Biomedical Computing for Breast Cancer Detection and Diagnosis*. [S.l.]:

IGI Global, 2021. p. 230–249.

QUEIROZ, K. F. F. d. C. *et al. Desenvolvimento e implementação de uma ferramenta computacional de uso médico para análise de imagens termográficas*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2016.

RASTGHALAM, R.; POURGHASSEM, H. Breast cancer detection using mrf-based probable texture feature and decision-level fusion-based classification using hmm on thermography images. *Pattern Recognition*, Elsevier, v. 51, p. 176–186, 2016.

RESMINI, R. *Classificação de Doenças da Mama Usando Imagens por Infravermelho*.Tese (Doutorado) — PhD Thesis, Federal Fluminense University. Niteroi, Rio de Janeiro,Brazil, 2016.

RING, E.; AMMER, K. The technique of infrared imaging in medicine. *Thermology international*, v. 10, n. 1, p. 7–14, 2000.

RING, F. *Thermal imaging today and its relevance to diabetes*. [S.l.]: SAGE Publications, 2010.

RINGNÉR, M. What is principal component analysis? *Nature biotechnology*, Nature Publishing Group, v. 26, n. 3, p. 303–304, 2008.

RODRIGUES, A. L.; SANTANA, M. d.; AZEVEDO, W.; BEZERRA, R.; SANTOS, W.d.;LIMA, R. d. Seleção de atributos para apoio ao diagnóstico do câncer de mama usando imagens termográficas, algoritmos genéticos e otimização por enxame de partículas. *IISimpósio de Inovação em Engenharia Biomédica (SABIO 2018), Recife, Brazil*, 2018.

SAMPAIO, W. B.; DINIZ, E. M.; SILVA, A. C.; PAIVA, A. C. D.; GATTASS, M. Detection of masses in mammogram images using cnn, geostatistic functions and svm. *Computersin Biology and Medicine*, Elsevier, v. 41, n. 8, p. 653–664, 2011.

SANCHES, I. J. Sobreposição de imagens de termografia e ressonância magnética: uma nova modalidade de imagem médica tridimensional. Universidade Tecnológica Federal do Paraná, 2009.

SCHOLZ, M. Approaches to analyse and interpret biological profile data. 2006.

SERRANO, R. C.; ULYSSES, J.; RIBEIRO, S.; LIMA, R.; CONCI, A. Using hurst coefficient and lacunarity to diagnosis early breast diseases. In: *17 th International Conferenceon Systems, Signals and Image Processing*. [S.l.: s.n.], 2010. p. 550–3.

SHANG, W.; HUANG, H.; ZHU, H.; LIN, Y.; QU, Y.; WANG, Z. A novel feature selection algorithm for text categorization. *Expert systems with applications*, Elsevier, v. 33, n. 1, p. 1–5, 2007.

SILVA, H.; FARIA, P.; CARRÃO, L.; AMADO, S.; ALMEIDA, H.; ESPANHA, M. Desenvolvimento de um sistema inteligente para padronização e caracterização de imagens térmicas de diferentes regiões anatómicas. 2015.

SIMÕES, A. da S.; COSTA, A. H. R. Classificação de laranjas baseada em padrõesvisuais. *Anais do Simpósio Brasileiro de Automaçao Inteligente*, 2003.

SMITH, L. I. *A tutorial on principal components analysis*. [S.l.], 2002.

SONG, F.; GUO, Z.; MEI, D. Feature selection using principal component analysis.In: IEEE. *2010 international conference on system science, engineering design and manufacturing informatization*. [S.l.], 2010. v. 1, p. 27–30.

SOUZA, J. C. *et al.* Diagnóstico de câncer de mama a partir de imagens de mamografia2d utilizando descritores de forma 3d. Universidade Federal do Maranhão, 2018.

STOLL, B. B.; DARÓS, L. V.; CURY, D.; MENEZES, C. S. de. Análise preditiva em bases desbalanceadas e comparação de técnicas de pré-processamento estudo de caso mooc. *Revista de Sistemas e Computação-RSC*, v. 10, n. 1, 2020.

STROBL, C.; BOULESTEIX, A.-L.; KNEIB, T.; AUGUSTIN, T.; ZEILEIS, A. Conditional variable importance for random forests. *BMC bioinformatics*, Springer, v. 9, n. 1, p. 307, 2008.

SUN, S.; HUANG, R. An adaptive k-nearest neighbor algorithm. In: IEEE. *2010 seventh international conference on fuzzy systems and knowledge discovery*. [S.l.], 2010. v. 1, p. 91–94.

TSAI, C.-F.; HSU, Y.-F.; YEN, D. C. A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, Elsevier, v. 24, p. 977–984, 2014.

UNLER, A.; MURAT, A. A discrete particle swarm optimization method for feature selection in binary classification problems. *European Journal of Operational Research*,Elsevier, v. 206, n. 3, p. 528–539, 2010.

VAPNIK, V. N. T he nature of statistical l earning t heory. *New York: Springer2V erlag*,1995.

VASCONCELOS, J. de; SANTOS, W. D.; LIMA, R. d. C. F. de. Analysis of methods of classification of breast thermographic images to determine their viability in the early breast cancer detection. *IEEE Latin America Transactions*, IEEE, v. 16, n. 6, p. 1631–1637, 2018.

VASCONCELOS, J. H. d. *et al. Investigações sobre métodos de classificação parauso em termografia de mama*. Dissertação (Mestrado) — Universidade Federal dePernambuco, 2017.

WANG, D.; TAN, D.; LIU, L. Particle swarm optimization algorithm: an overview. *Soft Computing*, Springer, v. 22, n. 2, p. 387–408, 2018.

WANG, X.; PALIWAL, K. K. Feature extraction and dimensionality reduction algorithmsand their applications in vowel recognition. *Pattern recognition*, Elsevier, v. 36, n. 10, p. 2429–2439, 2003.

YUAN, R.; LI, Z.; GUAN, X.; XU, L. An svm-based machine learning method for accurate internet traffic classification. *Information Systems Frontiers*, Springer, v. 12, n. 2, p. 149–156, 2010.

ZHU, W.; ZENG, N.; WANG, N. *et al.* Sensitivity, specificity, accuracy, associated confi-dence interval and roc analysis with practical sas implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, v. 19, p. 67, 2010.

ZOU, J.; HAN, Y.; SO, S.-S. Overview of artificial neural networks. In: *Artificial Neural Networks*. [S.l.]: Springer, 2008. p. 14–22.

APÊNDICE A – PARÂMETROS UTILIZADOS NOS CLASSIFICADORES

Logistic Regression:

- Regularization Type: Lasso
- Strength: 400

Random Forest:

- Number of trees: 6
- Number of attributes considered at each split: 3
- Limit depth of individual trees: 3
- Do not split subsets smaller than: 4

KNN:

- Number of neighbors: 4
- Metric: Euclidiana
- Weight: Distance

SVM:

- Regressio Cost: 1,00
- Complexy bound: 0,45
- Polynomial: RBF com g=0,02 c=0 d=3
- Numerical Tolerance: 0,0010
- Iteration Limit: 100

NN:

- Neuron in hidden layers: 100
- Activation: tanh
- Solver: ADAM

- Regularization: 0,0001
- Maximal number of iterations: 200