UNIVERSIDADE ESTADUAL DO MARANHÃO

CENTRO DE CIÊNCIAS TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

MESTRADO EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

Luciano Marques Brito Reis

Detecção de Lesões de Câncer de Pele Utilizando Análise de Componentes Independentes e Análise Discriminante Linear

São Luís-MA

2017

LUCIANO MARQUES BRITO REIS

DETECÇÃO DE LESÕES DE CÂNCER DE PELE UTILIZANDO ANÁLISE DE COMPONENTES INDEPENDENTES E ANÁLISE DISCRIMINANTE LINEAR

Dissertação apresentada ao Curso de Mestrado em Engenharia da Computação e Sistemas da Universidade Estadual do Maranhão, como requisito parcial para obtenção do Grau de Mestre. Área de Concentração: Computação Aplicada.

Orientador: Prof. Dr. Lúcio Flávio de Albuquerque Campos

São Luís-MA

2017

Reis, Luciano Marques Brito

Detecção de lesões de câncer de pele utilizando análise de componentes independentes e análise discriminante linear. / Luciano Marques Brito Reis. – São Luís, 2017.

63 f.

Dissertação (Mestrado) – Curso de Pós-Graduação em Engenharia da Computação e Sistemas, Universidade Estadual do Maranhão, 2017.

Orientador: Prof. Dr. Lúcio Flávio de Albuquerque Campos.

1. Análise de componentes independentes. 2. Análise discriminante linear. 3. Câncer de pele.. I. Título.

CDU: 519.7:616.5-006

LUCIANO MARQUES BRITO REIS

DETECÇÃO DE LESÕES DE CÂNCER DE PELE UTILIZANDO ANÁLISE DE COMPONENTES INDEPENDENTES E ANÁLISE DISCRIMINANTE LINEAR

Dissertação apresentada ao Curso de Mestrado em Engenharia da Computação e Sistemas da Universidade Estadual do Maranhão, como requisito parcial para obtenção do Grau de Mestre. Área de Concentração: Computação Aplicada.

Aprovada em 13 de Fevereiro de 2017.

BANCA EXAMINADORA

Prof. Dr. Lúcio Flávio de Albuquerque Campos (Orientador)

Doutor em Biotecnologia

Universidade Estadual do Maranhão

Prof. Dra. Flávia Baluz Bezerra de Farias Nunes

Doutora em Enfermagem em Saúde Pública

Universidade Federal do Maranhão

Prof. Me. Pedro Brandão Neto

Mestre em Engenharia da Eletricidade

Universidade Estadual do Maranhão

Pai, José Bonifácio Lisboa Reis

Irmã, Bianka Marques Brito Reis

Irmão, Ludenberg Marquse Brito Reis

Namorada, Bruna Mendonça de Oliveira

Amiga, Solimar Carvalho Martins

AGRADECIMENTOS

A Deus, pois sem ele nada seria possível.

À minha família, pelo encorajamento e apoio.

Ao professor Lúcio Flávio de Albuquerque Campos pela amizade e principalmente, pela paciência, sem a qual este trabalho não se realizaria.

À minha namorada, Bruna Mendonça de Oliveira, pelo companheirismo e apoio incondicional.

À amiga, Solimar Carvalho Martins, pela força e incentivo.

Aos professores do Programa de Pós-Graduação em Engenharia da Computação e Sistemas - PECS - UEMA pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o nosso enriquecimento pessoal e profissional.

Aos amigos do PECS Alex, Charles, David, Dayane, Elzenir, Marcos.

"Nenhum homem realmente produtivo pensa
como se estivesse escrevendo uma
dissertação".
Albert Einstein

RESUMO

O câncer de pele vem se configurando como o mais frequente em toda população brasileira. A exposição excessiva à radiação solar é o fator principal de risco. As pessoas com pele branca tem esse risco aumentado. As taxas de incidência e mortalidade do câncer de pele vem aumentando cada vez mais, dessa forma o diagnóstico precoce dessa enfermidade é o método mais simples para a diminuição dessas taxas. Esse trabalho tem como objetivo apresentar um método CAD baseado em processamento de imagens para detecção de lesões de câncer de pele melanoma e não melanoma utilizando análise de componentes independentes para extração de características, o algoritmo de máxima relevância e mínima redundância para redução de dimensionalidade e análise discriminante linear para a classificação das imagens. A capacidade do método foi avaliada pela técnica de validação cruzada, e atingiu 100% de acurácia, 100% de sensibilidade e 100% de especifidade ao analisar todas as amostras.

Palavras-chave: Análise de Componentes Independentes, Análise Discriminante Linear, Câncer de Pele.

ABSTRACT

Skin cancer has become the most frequent in the Brazilian population. Excessive exposure to solar radiation is the major risk factor. White people have this risk increase. The rates of incidence and morbidity of skin cancer have been increasing, so the early diagnosis of this disease is the simplest method to reduce these rates. This work aims to present a CAD method based on image processing for detection of melanoma and non-melanoma skin cancer using independent component analysis for feature extraction, the algorithm of maximum relevance and minimum redundancy for dimensionality reduction and analysis Linear discriminant for the classification of images. The ability of the method was evaluated by the cross-validation technique, and reached 100% accuracy, 100% sensitivity and 100% specificity when analyzing all samples.

Keywords: Independent Components Analysis, Linear Discriminant Analysis, Skin Cancer.

LISTA DE FIGURAS

LISTA DE SIGLAS

ABCD Assimetria, Borda, Cor, Diâmetro

ABCDE Assimetria, Borda, Cor, Diâmetro, Evolução

BNN *Back-propagation Neural Network* (Rede Neural de Retro-propagação)

BSS *Blind Source Separation* (Separação Cega de Fontes)

CAD *Computer-Aided Diagnosis* (Diagnóstico Auxiliado por Computador)

CID Classificação de Imagens Digitais

DMC Distância Mínima ao Centróide

DNA *Deoxyribonucleic Acid* (Ácido Desoxirribonucleico)

ICA *Independent Component Analysis* (Análise de Componentes Independentes)

INCA Instituto Nacional José Alencar Gomes da Silva

KNN *k-Nearest Neighbor* (k-vizinhos mais próximos)

LDA *Linear Discriminant Analysis* (Análise Discriminante Linear)

mRMR *Max-Relevance and Min-Redundancy* (Máxima Relevância e Mínima Redundância)

NB *Naive Bayes*

PDI Processamento Digital de Imagens

RGB *Red, Green, Blue* (Vermelho, Verde, Azul)

ROI Região de Interesse

SBCD Sociedade Brasileira de Cirurgia Dermatológica

SBD Sociedade Brasileira de Dermatologia

SRM *Statistical Region Merging* (Fusão de Região Estatística)

SVM *Support Vector Machine* (Máquina de Vetor de Suporte)

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Câncer é o nome dado a um grupo de doenças malignas caracterizadas pelo crescimento anormal e descontrolado de células que sofreram alteração em seu material genético, em algum momento do seu ciclo celular. Essas células geneticamente modificadas podem invadir os tecidos e órgãos, espalhando-se para outras regiões do corpo (ROBBINS *et al.*, 2001), gerando diversos tipos de câncer. Os diferentes tipos de câncer são classificados em grandes categorias: os carcinomas, as leucemias, os linfomas e mielomas e os tumores do sistema nervoso central (PRADO, 2014).

Neste trabalho, será abordado apenas o câncer de pele, que é um câncer do tipo carcinoma, isto é, são tumores que se originam nas células epiteliais ou glandulares (adenocarcinoma) com forte tendência a invadir tecidos vizinhos.

De acordo com Stewart *et al.* (2016), em seu estudo realizado em 2012, houve mais de 232.000 novos casos de melanoma e cerca de 55.000 mortes. As taxas mais elevadas são em países com a população caucasiana, predominantemente, com mais de 80% desses novos casos, e cerca de 65% de mortes ocorreram na Oceania, Europa e América do Norte.

Segundo o INCA (2016), no Brasil, embora o câncer de pele seja o mais frequente e corresponda a 30% de todos os tumores malignos registrados no país, o melanoma representa apenas 3% das neoplasias malignas do órgão, apesar de ser o mais grave devido à sua alta possibilidade de metástase. Enquanto o não melanoma apresenta altos percentuais de cura, caso seja detectado precocemente, e é entre os tumores de pele, o de menor incidência e mais baixa mortalidade.

Ainda segundo o INCA (2016), nos estados da Região Nordeste, em 2016-2017, a estimativa é que surjam 25.410 novos casos de câncer de pele, sendo 940 melanomas

onde 550 são em homens e 390 em mulheres e 24.470 não melanomas, onde 11.720 são em homens e 12.750 em mulheres. Já no Estado do Maranhão estima-se que, no ano de 2016, haverá um surgimento de 1350 novos casos de câncer de pele, sendo 60 novos casos de melanomas, onde 30 serão em homens e 30 em mulheres. Como também 1290 casos de não melanomas, onde 750 serão em homens e 540 em mulheres.

Apesar da grande incidência do câncer de pele, se diagnosticado precocemente, verifica-se um alto índice de cura. Infelizmente a maioria dos cânceres de pele não causam sintomas incômodos, exceto, eventualmente, quando as lesões já se tornaram maiores (PEREIRA, 2012). Dessa forma, a melhor maneira para se fazer um diagnóstico precoce é através do autoexame da pele.

Mesmo o autoexame da pele sendo a maneira mais simples de detecção precoce do câncer de pele, vários pesquisadores de várias áreas vêm desenvolvendo técnicas auxiliares para o diagnóstico precoce do câncer de pele com o intuito de diminuir essa taxa de mortalidade.

As técnicas auxiliares são denominadas de Diagnóstico Auxiliado por Computador (CAD) com o objetivo de melhorar a acurácia do diagnóstico e agir como uma segunda opinião ao especialista fazendo com que seja aumentada a sensibilidade do diagnóstico.

O objetivo geral deste trabalho é a presentar uma técnica CAD, através da Análise de Componentes Independentes (ICA) somada com o algoritmo de máxima relevância e mínima redundância (mRMR) para a extração de características e encontrar o melhor conjunto das características obtidas, que por fim serão analisadas por um classificador, baseado em Análise Discriminante Linear (LDA), para decidir se imagens de lesões de câncer de pele são do tipo melanoma ou não melanoma. E como objetivos específicos, esta técnica visa auxiliar o especialista, agindo como uma segunda opinião, para que seja aumentada a sensibilidade do diagnóstico, assim como melhorar a acurácia do diagnóstico e diminuir a taxa de mortalidade relacionada a essa enfermidade.

Buscando melhor expor os tópicos propostos, o trabalho será dividido como segue:

No capítulo 2 tem a finalidade de mostrar alguns trabalhos relacionados ao tema em questão.

No capítulo 3 será feita uma fundamentação teórica sobre o assunto e as técnicas utilizadas neste trabalho, apresentando alguns conceitos e objetivos.

O capítulo 4 descreve a metodologia utilizada em todas as etapas.

O capítulo 5 apresenta as discussões obtidas baseadas nos resultados da metodologia proposta.

O capítulo 6 versa sobre a conclusão e trabalhos futuros.

Capítulo 2

TRABALHOS RELACIONADOS

A incidência do câncer de pele está ganhando, cada vez mais, proporções epidêmicas, inúmeros especialistas de diversas áreas vêm desenvolvendo técnicas auxiliares para o diagnóstico precoce do câncer de pele com o intuito de diminuir sua taxa de mortalidade. Neste capítulo, iremos discutir alguns trabalhos relacionados ao tema em questão.

2.1 Análise e classificação de imagens de lesões de pele por atributos de cor, forma e textura utilizando máquina de vetor de suporte

Soares (2008), em sua tese de doutorado, propôs a análise e classificação de imagens de câncer de pele por atributos de cor, forma e textura. Para a realização da técnica proposta, foi utilizado um base de dados composto de 122 imagens, que foram obtidas através de exames clínicos, dermatoscopia e videodermatoscopia. Estas imagens apresentavam lesões de peles classificadas, por médicos dermatologistas, como melanoma, não melanoma e benigno.

A extração das características de textura foi utilizada a transformada *Wavelet Packet*, ou seja, um conjunto de funções base representará a imagem em diferentes bandas de frequências, sendo cada uma com resoluções distintas correspondente a cada escala. Como descritores de textura das imagens, foram calculados para cada imagem 84 canais de energia, 84 variâncias dos canais e 8 ramos de frequência dominante. Este procedimento lhe proporcionou uma melhor escolha das janelas para a representação das imagens, assim

como a diminuição do tempo de treinamento e um melhor desempenho do sistema.

Para os descritores de cor foi utilizada a imagem gerada pela segmentação realizada pelo algoritmo *k-means*. Isto é, o algoritmo fornece uma classificação de informações de acordo com os próprios dados, baseada em análises e comparações entre seus valores numéricos. Os descritores foram calculados de acordo com a média dos *pixels* da cor da pele e a mádia dos *pixels* da cor da lesão. Extraindo assim, apenas as cores RGB da lesão sem a influência da pele, acrescido da variância em cada um dos canais, acrescido da variância em cada um dos canais para que se tenha a variância da cor de cada canal de cor na lesão.

Após a segmentação foram extraídas as características de forma, baseados nos descritores regionais e descritores de *Fourier* com o objetivo de representar a lesão por meio de seu contorno independentemente das invariâncias da escala, rotação, translação e dispersão do sinal. Os descritores regionais são medidas derivadas do perímetro e da área utilizadas para a separação das classes e os descritores de *Fourier* definem o contorno da lesão, criando assim, sua assinatura.

Na etapa de classificação, os dados extraídos foram utilizados como padrões de entrada para a máquina de vetor de suporte (SVM), que é uma técnica de aprendizagem de máquina baseada na teoria de aprendizagem estatística. O conjunto de características foi dividido em dois subconjuntos, um para treinamento e outro para teste. Dois testes foram realizados, o primeiro através da estratégia tradicional um contra todos e a tomada de decisão através de uma Máquina de Comitê na forma de mistura de especialistas. Já o segundo teste foi realizado através da estratégica denominada de Máquinas SVM Especialistas, de forma que novos atributos são atribuídos às máquinas, em caso de empate, para a determinação da categoria pertencente.

O sistema de classificação foi composto de 6 máquinas SVM especialistas, para que os testes fossem realizados de acordo com as estratégias definidas na classificação. Cada máquina representava uma classe: máquina 1 - melanoma, máquina 2 - benigna, máquina 3 - não melanoma. E para o caso de empate: máquina 4 - melanoma $\times$ benigno, máquina 5 - melanoma $\times$ não melanoma, máquina 6 - benigna $\times$ não melanoma.

Utilizando a transformada *Wavelet* para a extração dos descritores de textura, para as 3 classes de imagens de lesões de pele, foi obtida uma taxa de acerto global de $92,73\%$ para melanoma e 86% para lesões benignas e não melanoma. E a estratégia denominada

de máquinas SVM especialistas se mostrou mais eficiente que a estratégia tradicional, obtendo assim, uma taxa de acerto global de 100% para melanoma e 90% para lesões benignas e não melanoma.

2.2 Reconhecimento do câncer de pele do tipo melanoma

Frutuoso, Santo e Siqueira (2013) propuseram uma técnica de reconhecimento de câncer de pele do tipo melanoma. A base de dados utilizada foram obtidas através de (SILVA, 2012) e (ROSADO, 2009) e continha 41 imagens, sendo 12 classificadas como melanoma e 29 que não correspondem a melanoma.

Neste sistema, não se tinha como objetivo realizar o acompanhamento da dimensão do câncer, portanto a análise do diâmetro (regra D) foi eliminada do processo.

Na etapa de pré processamento foram extraídas as regiões de interesse das imagens com o objetivo de remover todos outros elementos que poderiam estar nas imagens. E a extração de características foi baseada na regra do ABCD (assimetria, borda cor e diâmetro).

A análise da variação de cor (regra C) foi realizada com o intuito de quantificar a variação de cor que a região de interesse possui. Para esta etapa, primeiro calculou-se o histograma de cada componente (RGB) da imagem colorida. Logo em seguida, foi calculado o desvio padrão para cada histograma. E por fim, a quantidade de pontos maiores de que um determinado limiar pré-estabelecido foi processada.

A análise da variação da borda (regra B), para quantificar as irregularidades da borda da imagem foi realizada das seguinte forma: após a detecção da região de interesse, a imagem foi divida em 4 partes e a primeira parte utilizada para fazer o processamento. Utilizando as operações de convolução com as máscaras visando destacar os *pixels* brilhantes circundados por *pixels* mais escuros, realizou-se o cálculo da borda. Depois, o cálculo do histograma de projeção horizontal e vertical. E para cada histograma de projeção, foi calculado a quantidade de máximos e mínimos locais contidos em seu vetor.

A análise da simetria (regra A), feita para realizar a comparação entre as duas metades da região de interesse da imagem, e assim verificar as diferenças entre elas, seguiu os seguinte passos: detecção da região de interesse; divisão da imagem ao meio

através do ponto médio da largura da imagem; e utilizando o momentos Hu (invariantes), comparou-se as duas metades.

Para a etapa de classificação das imagens foram utilizados os classificadores *k-Nearest Neighbor* (KNN), *Naive Bayes* (NB), Distância Mínima ao Centróide (DCM). O KNN é um classificador que faz previsões de acordo com os rótulos dos K vizinhos mais próximos da instância de teste, O Naive Bayes é um classificador que faz previsões de acordo com o teorema de Bayes e o DMC, Para cada classe é assumido um centróide, isto é, um objeto pertence a essa classe quando a distância entre ele e o centróide for menor que todas as distâncias entre os outros centróides restantes do espaço de características.

Os resultados obtidos para o reconhecimento de lesões de pele do tipo melanoma, utilizando os 3 algoritmos foram superior a 70%, o classificador KNN obteve uma taxa de acerto de $81,81\%$, o classificador NB $90,90\%$ e $71,03\%$ para DMC.

2.3 Wavelet and curvelet analysis for automatic identication of melanoma based on neural network classification

Mahmoud, Al-Jumaily, Takruri (2013), propuseram um sistema para análise automática de melanoma. O banco de dados de imagem utilizado continha 448 imagens de dermatoscopia classificadas como melanoma maligno e benigno. E foram obtidas através do *Sydney Melanoma Diagnostic Centre* no *Royal Prince Alfred Hospital* e através de sites da internet.

A etapa de pré-processamento foi realizada para remover objetos irrelevantes contidos em cada imagem. Os quadros pretos, que geralmente aparecem nas imagens durante o processo de digitalização, foram retirados. E as imagens foram suavizadas através dos filtros *Wiener* e mediana. Logo em seguida, no pós-processamento, com o objetivo de melhorar a forma da imagem, foi utilizado o algoritmo de equalização de histograma, técnica utilizada para realçar o contraste.

A segmentação foi implementada a partir da Região de Interesse Limiar (ROI), método que calcula o valor de intensidade a partir de imagens cinzas. E da Fusão de Região Estatística (SRM), baseada na teoria da fusão de região que começa em um ponto

de semente que é comparado com seus quatro pontos vizinhos ou *pixels*.

O procedimento de extração de características foi realizado através das transformadas *Wavelet* e *Curvelet*. A análise *Wavelet* utiliza os termos de aproximações e detalhes. As aproximações são os componentes de baixa frequência de alta escala, enquanto os detalhes são os componentes de alta frequência de baixa escala. A transformada de *Curvelet* é uma pirâmide não-padrão multi-escala com muitas direções e posições em cada escala de comprimento, e em escalas elevadas, a forma de onda *Curvelet* se torna tão fina que se parece com um elemento em forma de agulha. Dessa forma, com o aumento no nível de resolução a *Curvelet* se torna mais fina e menor no domínio espacial e mostra mais sensibilidade nas bordas das curvas, o que lhe permite captar as curvas numa imagem com uma maior eficácia.

A classificação foi realizada através da Rede Neural de Retro-Propagação (BNN), ou seja, uma generalização da regra de aprendizagem de *Widrow-Hoff*, classe de filtro adaptativo usada para imitar um filtro desejado, para redes de camadas múltiplas e funções de transferências diferenciáveis não lineares.

A precisão do reconhecimento obtida pelo classificador foi de $58,44\%$ para os coeficientes baseados em *Wavelet* e $86,57\%$ para os baseados em *Curvelet*.

Capítulo 3

FUNDAMENTAÇÃO TEÓRICA

3.1 Câncer de pele

3.1.1 Aspectos gerais

Assim como qualquer outro órgão, a pele também pode sofrer com os fenômenos patológicos básicos que são determinadas pelas variações morfológicas na pele. Cruz (2009), caracteriza o câncer de pele pelo crescimento anormal e descontrolado das células que compõem a epiderme da pele.

O câncer de pele atualmente é o tipo de câncer mais comum no mundo, chegando a atingir proporções epidêmicas (DIEPGEN *et al.*, 2012).

As neoplasias cutâneas podem ser classificadas como benignas e malignas. São denominadas benignas quando não se espalham para outra parte do corpo, nem invadem outros tecidos ou órgãos adjacentes. Já as denominadas malignas, tanto podem se alastrar para outra parte do corpo, como também invadir e danificar outros tecidos ou órgão adjacentes (COFEN, 2015).

3.1.2 Tipos de tumores cutâneos

Os tumores benignos raramente se transformam em câncer. Já as neoplasias malignas podem ser divididas em dois grupos: melanoma e não melanoma, onde, este último, pode ser do tipo carcinoma basocelular e carcinoma espinocelular.

3.1.2.1 Melanoma

O melanoma é um tipo de câncer de pele que, apesar de representar apenas 4% das neoplasias malignas de pele, possui um prognóstico limitado e alto poder metastático (AVILA; CRUZ; RIERA, 2016).

Mesmo com a baixa incidência, o câncer de pele melanoma é o mais agressivo tipo de câncer de pele, apesar de ser menos frequente do que os outros tumores. Onde, a idade, o sexo e a susceptibilidade individual também são importantes no desenvolvimento desse tipo de câncer. E cerca de 20% a 30% dos melanomas está associada à presença de nevo melanocítico prévio (INCA, 2016).

Existem quatro principais grupos de melanomas:

1. Extensivo Superficial

Localiza-se, geralmente, no tronco dos homens e nos membros inferiores das mulheres. Representa cerca de 70% dos casos de melanomas. E é frequentemente diagnosticado em pacientes entre 30 e 50 anos.

2. Nodular

É encontrada, frequentemente, na cabeça e no pescoço. É equivalente a cerca de 10% a 15% dos casos de melanoma. Costuma ser diagnosticado em pacientes entre 50 e 60 anos.

3. Lentigo Maligno

Afeta principalmente as áreas mais danificadas pelo sol como rostos, pescoço e braços. Ocorre em cerca de 5% dos casos de melanoma. A fase mais propensa a este tipo de melanoma é entre os 60 e 70 anos.

4. Lentiginoso Acral

Desenvolve-se nas plantas dos pés, palmas das mãos, sob as unhas e representa cerca de 5% dos casos de melanoma. E aparece, comumente, em pessoas de pele escura de qualquer idade.

3.1.2.2 Não melanoma

Dados do (INCA, 2016) demonstram que o câncer de pele não melanoma continua sendo o tumor mais incidente em ambos os sexos. Vale ressaltar que é provável que exista um sub-registro dessa neoplasia em função do subdiagnóstico, podendo assim subestimar as taxas de incidência e os números esperados de casos novos.

É o câncer mais frequente no Brasil e corresponde a 30% de todos os tumores malignos registrados no país. Apresenta altos percentuais de cura, se for detectado precocemente. Entre os tumores de pele, o tipo não melanoma é o de maior incidência e menor taxa de mortalidade (INCA, 2016).

Existem dois tipos de câncer de pele não melanoma:

1. Carcinoma Basocelular

É o tipo mais comum de câncer de pele, representando cerca de 70% dos casos diagnosticados. Desenvolve-se nas regiões mais expostas ao sol como rostos, braços, pescoços e etc. Como o surgimento está diretamente ligado a exposição solar, ocorrem geralmente a partir dos 40 anos de idade.

2. Carcinoma Espinocelular

É o segundo tipo de câncer de pele mais incidente, representado cerca de 20% dos tumores cutâneos. Assim como o basocelular, este tipo de câncer desenvolve-se nas regiões expostas mais expostas ao sol. Mas também pode surgir em feridas crônicas e cicatrizes antigas. Este tipo de câncer é diagnosticado com mais frequência em pessoas idosas.

3.1.3 Fatores de risco e prevenção

Segundo informações do INCA, o termo ”risco” é usado para definir a chance de uma pessoa sadia, exposta a determinados fatores, ambientais ou hereditários, desenvolver uma doença. Os fatores associados ao aumento do risco de se desenvolver uma doença são chamados fatores de risco. Alguns dos fatores de riscos associados ao câncer de pele são a exposição excessiva aos raios solares, exposição a fatores químicos, o histórico familiar de melanoma, xeroderma pigmentoso, nevo displásico entre outros.

De acordo com as normas e recomendações do INCA para a prevenção do câncer de pele (2013), as queixas mais comuns relacionadas ao câncer da pele são:

- mancha que coça, dói, sangra ou descama;
- ferida que não cicatriza em 4 semanas;
- sinal que muda de cor textura, tamanho, espessura ou contornos;
- elevação ou nódulo circunscrito e adquirido da pele que aumenta de tamanho e tem aparência perolada, translúcida, avermelhada ou escura.

A prevenção é classificada em dois níveis: A prevenção primária, refere-se a toda e qualquer ação voltada para redução da exposição da população a fatores de risco da doença, tendo como objetivo reduzir a sua ocorrência, por meio da promoção da saúde (hábitos saudáveis de vida) e proteção específica; E a prevenção secundária é o rastreamento (*screening*) do câncer, ou seja, uma avaliação de indivíduos assintomáticos, para classificá-los como candidatos a exames mais refinados de avaliação, com o objetivo de descobrir um câncer oculto ou uma afecção pré-maligna que pode ser curada com tratamento (CESTARI; ZAGO, 2005).

Diante disso, a mudança dos hábitos comuns de vida para hábitos saudáveis de vida, assim como após a detecção do surgimento da doença nos estágios iniciais é fundamental para que os índices de incidência e mortalidade por câncer no Brasil possam ser reduzidos. Nos últimos anos, principalmente devido à detecção precoce, houve uma grande melhora na sobrevida dos pacientes com melanoma. A sobrevida média mundial estimada em 5 anos é de 69%, sendo de 73% nos países desenvolvidos e de 56% nos países em desenvolvimento (BRASIL, 2013). Sendo assim, a prevenção do câncer de pele pode fazer uma grande diferença com relação a gravidade da lesão. Portanto, algumas ações podem ajudar a prevenir o câncer de pele, tais como, evitar exposição excessiva ao sol ou fontes artificiais de radiação ultravioleta, não se expor determinadas substâncias químicas, evitar o consumo de álcool, não fumar, praticar atividades físicas, ter uma alimentação saudável (PRADO, 2014).

De acordo com a Sociedade Brasileira de Dermatologia - SBD (2016), todos os casos de câncer de pele devem ser diagnosticados e tratados precocemente, inclusive os de baixa letalidade, que podem provocar lesões mutilantes ou desfigurantes em áreas expostas do corpo, causando sofrimento aos pacientes.

3.1.4 Métodos diagnóstico do câncer de pele

3.1.4.1 Diagnóstico clínico - Regra do ABCDE

Stolz *et al.* (1994) propuseram um método para a análise de lesões pigmentadas, denominada de regra do ABCD (Assimetria, Borda, Cor e Diâmetro). A regra do ABCD é utilizada para orientar médicos, profissionais da saúde e pacientes quanto ao reconhecimento das principais características de lesões cutâneas suspeitas (MÜLLER *et al.*, 2009).

Após o seu surgimento, a regra do ABCD vem sendo utilizada e comprovada ao longo dessas duas décadas. Mas, infelizmente, esta regra tem algumas exceções e limitações. Por exemplo, a regra do ABCD não é válida para os seguintes tipos de melanoma: o melanoma nodular, o melanoma amelanótico, o melanoma inicial e o melanoma nevoide em que as condições das lesões podem apresentar simetria, bordas regulares, cor homogênea e diâmetro inferior de 6mm (SILVA *et al.*, 2016). Diante disso, foi acrescido a letra E (evolução) ao ABCD. E assim, passou-se a avaliar outros tipos de modificações como espessura, coceira ou sangramento (BARCELOS; PIRES, 2009). Esse acréscimo corrobora com a acurácia diagnóstica do exame clínico do melanoma (WAINSTEIN ALBERTO; BELFORT, 2014).

O exame clínico da pele deve fazer parte do exame físico de rotina, mesmo que a queixa principal do paciente não esteja localizada na pele, especialmente as pessoas de pele branca, trabalhadores rurais, pescadores e outros profissionais com alta exposição à luz solar (INCA, 2003).

As ocorrências mais comuns relacionadas ao câncer de pele, são sinais que modificam a textura, espessura, tamanho, lesões que não cicatrizam, manchas que apresentem coceira, dor ou sangramento (INCA, 2003).Para realizar o diagnóstico clínico através da regra do ABCDE é necessário seguir os seguintes passos:

- A - Assimetria

  Ao dividir o sinal ao meio, verificar a simetria entre os lados. Lesões benignas são caracterizadas por lados semelhantes, enquanto que as malignas possuem lados diferentes.

- B - Borda

  As bordas das lesões benignas apresentam-se uniformes, bem definidas. Já as lesões malignas apresentam bordas mal definidas, irregulares.

- C - Cor

  As benignas possuem uma cor uniforme. Por outro lado, as malignas apresentam cores variadas.

- D - Diâmetro

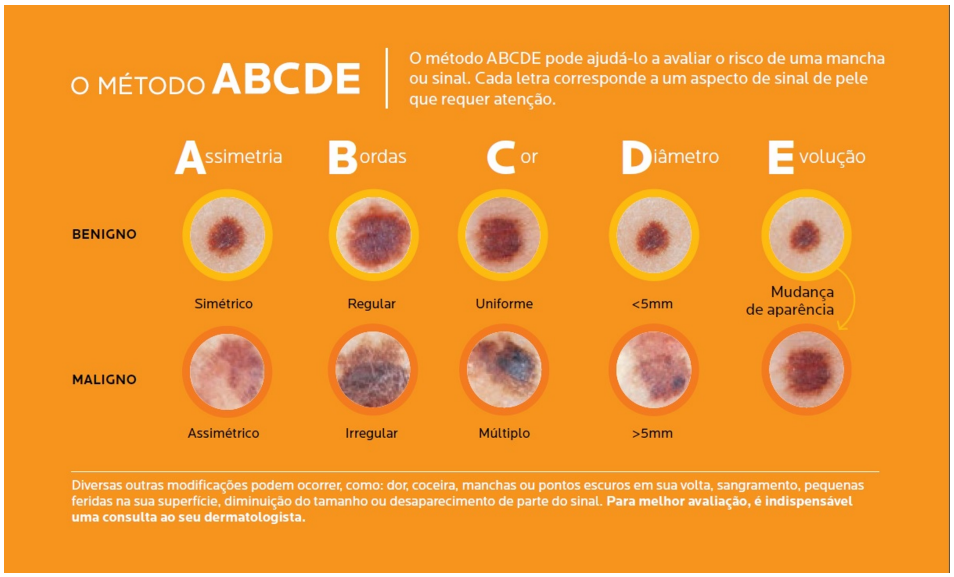
  Lesões benignas são menores que 6 mm de diâmetro. Enquanto as malignas maiores que 6 mm.

- E - Evolução

  As lesões malignas tendem a modificar suas características, coçar ou sangrar.

A figura (3.1) exemplifica as verificações feitas através da regra do ABCDE.

Figura 3.1: Regra do ABCDE



Fonte: (NEWSWIRE, 2015)

3.1.4.2 Diagnóstico dermatoscópico

A dermatoscopia ou microscopia de epiluminescência surgiu em 1933, quando Hinselman propôs a utilização de um colposcópio para a realização de exames de alta potência em mucosas e lesões cutâneas. Em 1951, Goldman utilizou a microscopia de superfície para o diagnóstico de lesões de pele pigmentadas. 20 anos depois, em 1971, a microscopia resurgiu com Mackie, para o diagnóstico de lesões de pele. E, em 1981, Fritsch e Pechlaner, aperfeiçoaram a técnica para diferenciar lesões de pele benignas e malignas. Em vista

disso, a dermatoscopia ou microscopia de epiluminescência surgiu como exame auxiliar *in vivo*, que tem papel fundamental na realização do diagnóstico precoce e amplifica a acurácia diagnóstica do melanoma (REZZE; SA; NEVES, 2006).

Frange, Arruda, Daldon (2009) define a dermatoscopia como um método diagnóstico não invasivo auxiliar na avaliação das lesões pigmentadas da pele. É realizada mediante o emprego do dermatoscópio (microscópio de epiluminescência), instrumento que permite o aumento de 10 vezes das lesões cutâneas. Portanto, a dermatoscópia é de suma importância para o diagnóstico clínico, pelo fato de ser utilizada para visualizar estruturas morfológicas de lesões pigmentadas, não visíveis a olho nu e correlacionadas com características histológicas específicas (MARIANTE, 2003).

Segundo a Sociedade Brasileira de Cirurgia Dermatológica (SBCD) exite dois tipos de exames dermatoscópico:

- A dermatoscopia das lesões pigmentadas que é feita no consultório, normalmente durante a consulta ou como procedimento individualizado.

- O mapeamento corporal, realizado por especialista nesse exame, na qual são utilizados aparelhos mais sofisticados e serve para fotografar todo o indivíduo e detectar precocemente modificações ou aparecimento de pintas malignas.

3.1.4.3 Diagnóstico histopatológico

Em 1828, a histopatologia começou a ser utilizada para a elaboração de bases da patologia celular pelo médico Rudolph Virchow. E está baseada no critério morfológico arquitetural e celular, sendo considerada a melhor técnica para o diagnóstico morfológico (STIVAL *et al.*, 2005).

A histopatologia ou anatomopatologia é a análise microscópica de tecidos que são removidos de pacientes quando é realizada uma biópsia (coleta de amostras de tecidos, para estudos laboratoriais) (EDUCAÇÃO, 2013).

O exame histopatológico pode afirmar com precisão a natureza da lesão, já que algumas lesões de pele podem ter características semelhantes. Assim, a descrição histopatológico deve reportar informações úteis para estabelecer o estádio, tratamento e prognóstico de um melanoma (ACOSTA *et al.*, 2008).

De acordo com a American Cancer Society (2016) existem diversos tipos de biópsia cutâneas para diagnosticar o câncer de pele histopatologicamente, e essa escolha depende

do tamanho da área afetada e da localização no corpo. Alguns tipos de biópsias são discutidos a seguir:

- Biópsia por raspagem

  O médico raspa as camadas superiores da pele com uma pequena lâmina cirúrgica. Este tipo de biópsia é utilizada no diagnóstico de muitos tipos de doenças de pele e em sinais de amostragem quando o risco de melanoma é muito baixo;

- Biópsia por punção

  O médico usa uma ferramenta que se parece com um cortador de biscoito redondo minúsculo para remover uma amostra mais profunda da pele.

- Biópsias incisional e excisional

  A biópsia incisional remove apenas uma porção do tumor. A biópsia excisional remove todo o tumor. Este tipo de biópsia é utilizado para examinar um tumor localizado nas camadas mais profundas da pele.

- Biópsias ópticas

  Tipo de biópsia que examina o tumor sem a remoção de amostras.

3.2 Processamento digital de imagens - PDI e diagnóstico auxiliado por computador - CAD

Por Processamento Digital de Imagens (PDI) entende-se a manipulação de uma imagem por computador de modo que a entrada e a saída do processo sejam imagens. O objetivo de se usar processamento digital de imagens é melhorar o aspecto visual de certas feições estruturais para o analista humano e fornecer outros subsídios para a sua interpretação, inclusive gerando produtos que possam ser posteriormente submetidos a outros processamentos (CÂMARA *et al.*, 1996).

Conforme Silva (2001), a principal função do processamento digital de imagens de sensoriamento remoto é a de dar instrumentos para melhorar a identificação e a retirada de informações contidas nas imagens, para interpretação posterior. Nesse contexto, sistemas de computação são utilizados para atividades interativas de análise e manipulação das imagens brutas. O que resulta desse processo é a produção de outras imagens, estas já dotadas de informações já conhecidas, retiradas e realçadas a partir das imagens brutas.

A informação de interesse é caracterizada em função das características dos objetos ou padrões que compõem a imagem. Portanto, retirar informações de imagens envolve a identificação de objetos ou padrões. A maior parte dessa atividade requer grande capacidade cognitiva por parte do intérprete, devido à complexidade dos processos envolvidos e à falta de algoritmos computacionais precisos o bastante para realizá-lo de forma automática.

O ser humano é dono de um sistema visual com capacidade incrível de reconhecer padrões. No entanto, ele dificilmente é capaz de processar o enorme volume de informações presentes numa imagem. Vários tipos de degradações e distorções, inerentes aos processos de aquisição, transmissão e visualização de imagens, contribuem, ainda mais, para a limitação dessa capacidade do olho humano.

O principal objetivo do processamento de imagens é o de encerrar essas limitações humanas, ajudando na retirada de informações a partir de imagens. Nesse contexto, o processamento digital deve ser visto como um estágio preparatório, embora quase sempre obrigatório, da atividade de interpretação das imagens (BRYS, 2008).

O processo de tomada de decisão de especialistas pode ser uma tarefa muito complicada, principalmente quando os fatores conhecidos e recursos disponíveis não são tão evidentes (MARTINEZ, 2007). De acordo com Poel *et al.* (2007) a detecção de anormalidades em imagens médicas é, geralmente, um processo complexo suscetível a vários erros.

Diagnóstico auxiliado por computador (CAD) é definido como um diagnóstico realizado pelo especialista, utilizando o resultado de análises quantitativas automatizadas de imagens como auxílio para tomada de decisões diagnósticas, e tem como objetivo melhorar a acuidade do diagnóstico, assim como a consistência da interpretação da imagem, mediante o uso da resposta do computador como referência (SEIXAS; SAADER, 2005). Tal tecnologia, vem ganhando cada vez mais espaço nas mais diferentes modalidades diagnósticas, com a finalidade de reduzir o tempo de leitura dos exames, com aumento da especificidade e, principalmente, da sensibilidade (CAPOBIANCO; JASINOWODOLINSKI; SZARF, 2008). É sempre importante deixar claro que o CAD deve ser utilizado apenas como uma ferramenta para se obter informações adicionais acerca da imagem, e que o diagnóstico final será dado pelo especialista (AZEVEDO-MARQUES, 2001).

A análise de imagens requer tempo, além do que o custo do treinamento de es-

pecialistas na análise de um número grande de imagens é muito alto. Por conta disso, pesquisadores têm desenvolvido sistemas CAD. De um modo geral, os sistemas CAD utilizam-se de técnicas provenientes de duas áreas do conhecimento, visando a retirada e quantificação de características de uma imagem em formato digital (RODRIGUES, 2008):

- Visão computacional: utiliza o processamento de imagem para realce, segmentação e retirada de características;
- Inteligência artificial: envolve métodos para seleção de características, estatísticas e reconhecimento de padrões.

Os sistemas CAD podem ser usados com o propósito de triagem. Como na condição de triagem a probabilidade de positivo verdadeiro é, relativamente, baixa, e usam leitura manual que além de tediosa é demorada, o sistema de análise automática de imagens pode indicar imagens anormais ou duvidosas em posterior análise pelo interprete. O problema da interpretação de imagens é baseado em técnicas de Inteligência Artificial. Apesar de ser improvável que o desenvolvimento de interpretação completamente automatizada ocorra brevemente, sistemas que oferecem interpretação parcialmente automatizada são viáveis. Esses sistemas resolvem subtarefas de uma tarefa de interpretação global.

O processamento global envolve a computação levando em consideração a imagem em sua totalidade, sem levar em consideração o conteúdo local especifico. O objetivo é realçar a imagem para a visualização humana ou para posterior análise pelo computador. Deste modo o intérprete consegue perceber melhor pequenas mudanças na resolução de contraste dentro da sub-região de interesse, porém ao mesmo tempo eles sacrificam a resolução em outras áreas da imagem.

Com base em diversas pesquisas relacionadas, é notória a importância do uso de recursos tecnológicos no auxílio de diagnósticos de diversas doenças. Qualquer que seja o procedimento realizado no paciente, o uso de equipamentos tanto de hardware como de software melhora muito o processo e a tomada de decisão. A utilização da tecnologia na investigação de problemas contribui cada vez mais para auxiliar os especialistas no aumento da acurácia dos diagnósticos (FURGERI; RODRIGUES; SILVA, 2013).

O uso de *softwares* CAD melhora a eficácia na determinação de anormalidades, o que implicaria na redução de biópsias e outros procedimentos invasivos. Como cada procedimento tem sua funcionalidade própria, podendo assim, haver a necessidade de mais

de um procedimento para se obter o diagnóstico final, é necessário agregar o maior número de funcionalidades possíveis ao *software* CAD. Desse modo, aumentando as chances de detecção de anormalidades. Lembrando que, o devido treinamento aos especialistas é necessário para que o diagnóstico seja o mais preciso possível.

3.3 Extração de características

O objetivo da extração de características é definir um objeto por meio de suas medidas para que objetos semelhantes sejam reconhecidos na mesma categoria e objetos diferentes em categorias diferentes (DUDA; HART; STORK, 2012). Com isso, os métodos utilizados para a extração de características visam contornar este problema e representar os dados com certa precisão.

Se o espaço de características contiver apenas as mais notáveis, o classificador poderá agir de forma mais rápida e necessitará de menos memória (BIANCHI, 2006).

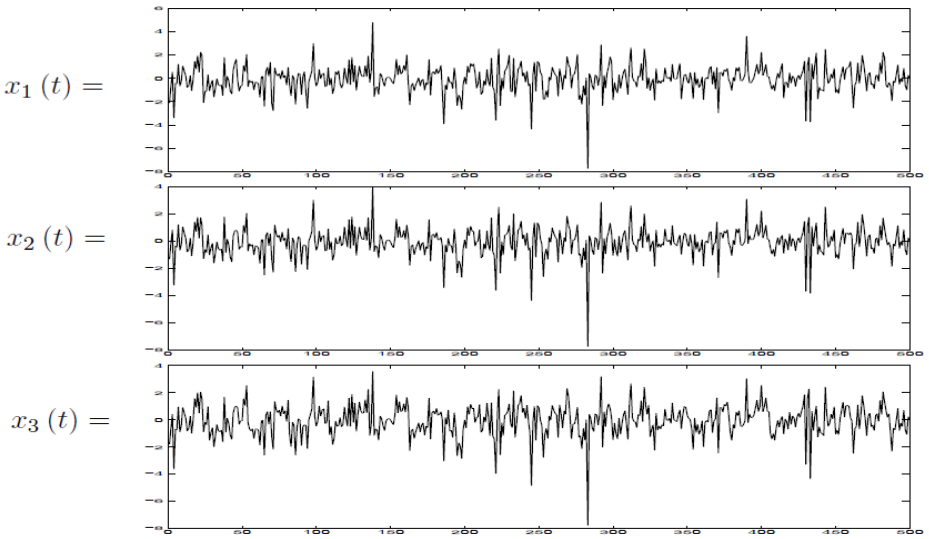
3.3.1 Análise de componentes independentes - ICA

A análise de componentes independentes (ICA) é uma técnica estatística e computacional que revela fatores ou componentes que fundamentam um conjunto de variáveis aleatórias, medições ou sinais multivariados observados. A ICA define um modelo para gerar dados multivariados observados, que são reconhecidos como misturas de variáveis lineares ou não lineares de algumas variáveis latentes desconhecidas. Essas variáveis latentes são assumidas como mutuamente independentes e não gaussianas, assim como são denominadas de componentes independentes de dados multivariados observados (HYVÄRINEN; KARHUNEN; OJA, 2004).

A origem da ICA se dá por volta das décadas 50 e 70, através dos trabalhos *Anályse Générale des Liaisons Stochastiques* (DARMOIS, 1953) e *Characterization Problems in Mathematical Statist* (KAGAN; RAO; LINNIK, 1973), respectivamente, quando as variáveis aleatórias começaram a serem caracterizadas em estruturas lineares. Mas somente nas décadas de 80, com *Space or Time Adaptive Signal Processing by Neural Networks Models* (HERAULT; JUTTEN, 1986), e 90, com *Independent Component Analysis - a new concept?* (COMON, 1994) a análise de componentes independentes passa a ser utilizada para modelagem de redes neurais e para a separação de sinais de áudios na telecomunicação.

A ICA é uma técnica largamente utilizada, principalmente, para o tratamento de problemas de separação cegas de sinais de fontes independentes, denominado também de separação cega de fontes ou *Blind Source Separation* (BSS), que está associada ao modelo *Cocktail Party*. Por exemplo, suponha que estejam três pessoas em uma sala conversando, e em pontos diferentes desta mesma sala, instala-se três microfones para que gravem a conversa durante um intervalo de tempo t, figura (3.2).

Figura 3.2: Sinais capturados pelos microfones.



Fonte: (HYVÄRINEN; OJA, 2000)

Denotando cada conversa gravada por $x_1(t), x_2(t)$ e $x_3(t)$ e cada amostra capturada por cada microfone como $s_1(t), s_2(t)$ e $s_3(t)$. Dessa forma, tem-se

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t) \\ x_2(t) &= a_{12}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t) \\ x_3(t) &= a_{13}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t) \end{aligned} \tag{3.1}$$

onde, a_{ij} são coeficientes reais. E assim, o problema consiste em encontrar $s_1(t), s_2(t)$ e $s_3(t)$.

A equação (3.1) também pode ser representada na forma matricial:

$$X = AS \tag{3.2}$$

Sendo,

$$X = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} \qquad A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{32} & a_{33} \end{pmatrix} \qquad S = \begin{pmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{pmatrix}$$

e A é a matriz de coeficientes de mistura.

Dessa forma, o modelo apresentado pela equação (3.2) refere-se a geração de dados observados baseado no processo de mistura das componentes independentes.

A matriz de dados X é considerada uma combinação linear das componentes não-gaussianas (independentes), tais que, $X = A * S$, sendo que as colunas de S contêm as componentes independentes e A é a matriz de mistura. Sendo assim, a técnica ICA tenta separar os dados, estimando uma matriz de separação $W = A^{-1}$, tal que a matriz de componentes possa ser encontrada a partir dos dados observados, isto é,

$$X * W = S \tag{3.3}$$

3.3.1.1 Restrições

Para que a técnica ICA possa estimar as componentes independentes é necessário que sejam observadas as seguintes restrições:

1. As componentes independentes devem ser estatisticamente independentes.

Para que duas ou mais variáveis aleatórias sejam consideradas estatisticamente independentes, elas devem satisfazer o critério de mínima informação mútua, ou seja, o valor de uma variável não fornece nenhuma informação sobre o valor de outra variável.

A independência estatística também pode ser definida através das funções de probabilidades das variáveis aleatórias. Sejam x_1 e x_2 variáveis aleatórias, $p(x_1, x_2)$ a função densidade de probabilidade conjunta (fdp) e $p_1(x_1)$ e $p_2(x_2)$ as funções densidades de probabilidade marginal de x_1 e x_2. Diz-se que duas variáveis aleatórias são estatisticamente independentes se a função densidade de probabilidade conjunta for igual ao produto das funções densidades de probabilidade marginal, isto é,

$$p(x_1, x_2) = p_1(x_1) * p_2(x_2) \tag{3.4}$$

Uma característica importante para a independência estatística entre duas variáveis

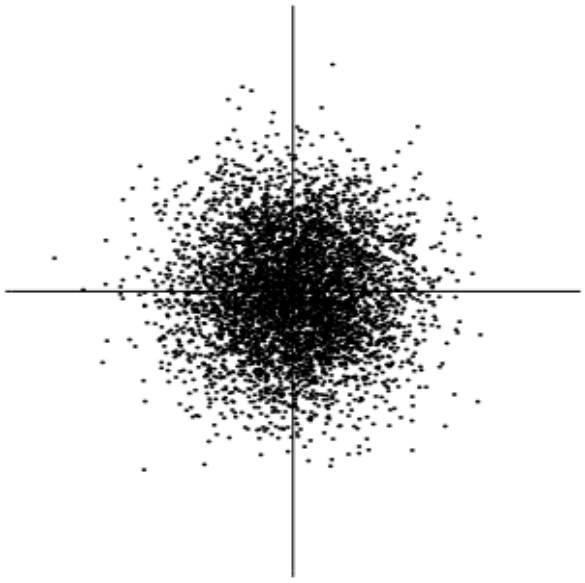
aleatórias é a não correlação. Duas variáveis aleatórias x_1 e x_2 são ditas não correlacionadas se sua covariância for igual a zero,

$$E\{x_1, x_2\} - E\{x_1\} E\{x_2\} = 0 \tag{3.5}$$

2. As componentes independentes devem possuir distribuições de probabilidade não-gaussianas.

Conforme ilustra a figura (3.3), as distribuições gaussianas são simétricas, e dessa forma as suas propriedades serão idênticas em qualquer direção quando rotacionada. E dessa forma, a estimação da matriz de mistura das duas componentes originais será impossível (FRANCO, 2008).

Figura 3.3: Distribuição gaussiana de variáveis independentes



Fonte: (HYVÄRINEN; OJA, 2000)

3.3.1.2 Ambiguidades

Em relação às componentes independentes, a técnica ICA, apresenta algumas ambiguidades.

1. Não se pode determinar as variâncias das componentes independentes.

De acordo com Hyvärinen, Karhunen, Oja (2004), sendo S e A desconhecidos, qualquer multiplicador escalar em uma das fontes s_i poderia ser cancelado dividindo a

coluna correspondente a_i de A pelo mesmo escalar.

2. Não se pode determinar a ordem das componentes independentes.

Segundo Silva (2010), como A e S são desconhecidos, pode-se alterar a ordem da soma dada pela equação (3.1), obtendo assim o mesmo resultado na combinação linear.

3.3.1.3 Estimação das componentes Independentes

A ICA consegue estimar a matriz de mistura e recuperar as componentes independentes através das informações obtidas pelas misturas observadas. A estimação das componentes independentes S pode ser obtida através da matriz de mistura A, conforme a equação (3.3).

Assim, pode-se expressar uma combinação linear x_i por

$$y = b^T X \tag{3.6}$$

e pela equação (3.2), reescreve-se a equação (3.6) da seguinte forma

$$y = b^T AS \tag{3.7}$$

Veja que y é uma combinação linear das componentes independentes com vetor $q = b^T A$ e,

$$y = q^T S \tag{3.8}$$

Sendo b uma linha de A^{-1}, então y será uma das componentes independentes, e neste caso, um dos elementos de q será igual a um e todos os outros serão iguais a zero. No entanto, sendo X conhecido, b não pode ser determinado exatamente, porém pode-se estimar seu valor (CAMPOS, 2013).

O Teorema Central do Limite afirma que a soma de variáveis aleatórias independentes, $y_i = x_1 + \ldots + x_n$, é uma variável com distribuição que se aproxima da distribuição gaussiana quando $n \to \infty$. Dessa forma, se considerando duas variáveis aleatórias estatisticamente independentes e distribuídas igualmente, a soma dessas variáveis aleatórias terá uma distribuição mais próxima da distribuição gaussiana do que qualquer distribuição das variáveis originais (SILVA, 2010). Sendo assim, tem-se que y será mais gaussiana que qualquer componente s_i e menos gaussiana quando se igualar a uma das s_i.

Então, deve-se variar b e observar a distribuição de y. Ao se encontrar um vetor b que maximize a não-gaussianidade de y, ter-se-á encontrado uma componente independente (LEITE, 2005).

3.3.1.4 Medidas para a não gaussianidade

Para se estimar as componentes independentes de uma mistura, as variáveis aleatórias devem possuir distribuições não-gaussianas. Dessa forma, para utilizar a não-gaussianidade na estimativa da técnica ICA, serão discutidas apenas duas importantes medidas de não-gaussianidade, a curtose, a entropia e a negentropia.

1. Curtose

A curtose é um parâmetro que descreve a forma de uma função densidade de probabilidade. Visto que uma distribuição gaussiana possui curtose normalizada igual a zero, ela também pode ser utilizada como medida de não gaussianidade de uma variável aleatória (SILVA, 2009).

A Curtose é uma medida clássica de não gaussianidade para a estimativa do ICA e é definida como o cumulante de quarta ordem de uma variável aleatória

$$kurt\left(y\right) = E\left\{y^4\right\} - 3\left(E\left\{y^2\right\}\right)^2 \tag{3.9}$$

onde $E\left\{y^4\right\}$ é o cumulante de quarta ordem e é a variância $\left(E\left\{y^2\right\}\right)^2$

Supondo que y tenha a variância unitária, ou seja, $\left(E\left\{y^2\right\}\right)^2 = 1$, tem-se que

$$kurt\left(y\right) = E\left\{y^4\right\} - 3 \tag{3.10}$$

e, dessa forma, a função $kurt\left(y\right)$ é uma versão normalizada do cumulante de quarta ordem.

Um fato de extrema importância para calcular a gaussianidade de uma variável com distribuição gaussiana, é a medida da curtose ser igual a zero, isto é, o cumulante de quarta ordem é $3\left(E\left\{y^2\right\}\right)^2$. Já que, para variáveis não gaussianas, a medida da curtose é diferente de zero (ABIB *et al.*, 2013).

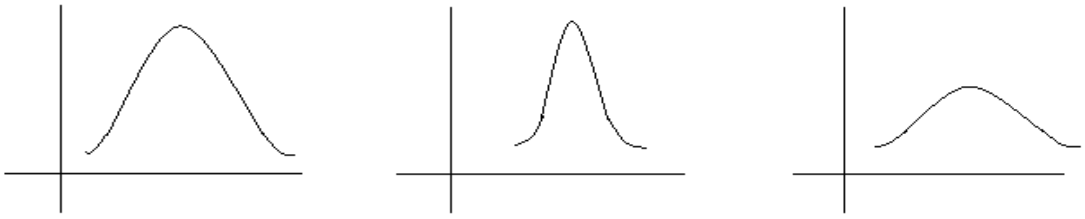
- $kurt\left(y\right) = 0$ - variável gaussiana ou mesocúrtica, figura (3.4a). A função de densidade de probabilidade tem distribuição normal.

- $kurt\,(y) > 0$ - variável super-gaussiana ou leptocúrtica, figura (3.4b). A função de densidade de probabilidade tem distribuição mais elevada e mais concentrada que a normal.

- $kurt\,(y) < 0$ - variável sub-gaussiana ou platicúrtica, figura (3.4c). A função de densidade de probabilidade tem distribuição é menos elevada e menos concentrada que a normal.

Figura 3.4: Formas da curtose

(a) Mesocúrtica (b) Leptocúrtica (c) Platicúrtica



Fonte: (ESTATSITE.COM, 2016)

2. Entropia

A entropia de uma variável aleatória é uma medida da quantidade de informação requerida sobre a média para descrever a variável aleatória, ou seja, está relacionada com o grau de informação que esta variável fornece. O conceito de entropia está associado a medida de incerteza que desta variável, assim sendo, quanto maior for sua entropia, maior será a desorganização e aleatoriedade desta variável (COVER; THOMAS, 2012).

A entropia $H\,(y)$ de uma variável aleatória y com função densidade de probabilidade $f\,(y)$ é definida da seguinte forma

$$H\,(y) = -\int f\,(y) \log\,(y)\ \ dy \tag{3.11}$$

Como uma variável aleatória com distribuição gaussiana tem maior entropia que qualquer outra variável com qualquer distribuição, pode-se utilizar a entropia como medida de não gaussianidade (DIAS, 2014).

3. Negentropia

A negentropia está relacionada com a quantidade de informação teórica de uma variável dada pela entropia (SILVA, 2010). Fundamenta-se na não gaussianidade da distribuição utilizada pela técnica ICA para estimar as componentes independentes (MORIMITSU; TUESTA, 2015).

Seja y_{gauss} uma variável aleatória com distribuição gaussiana e mesma matriz de covariância que y, define negentropia da seguinte forma

$$H(y) = H(y_{gauss}) - H(y) \tag{3.12}$$

A medida da negentropia é sempre não negativa. Assumindo o valor zero apenas se a variável aleatória y, possuir distribuição gaussiana e é constante para transformações lineares inversíveis.

Apesar da negentropia ser uma ótima medida de não gaussianidade, ela apresenta exige uma robustez computacional. Já que para calcular o seu valor em uma variável extremamente difícil, havendo a necessidade de técnicas auxiliares para sua estimação aproximada. Uma dessas técnicas de aproximação é feita utilizando o momento de quarta ordem ou curtose

$$J(y) \approx \frac{1}{12} E\left\{y^3\right\}^2 + \frac{1}{48} kurt(y)^2 \tag{3.13}$$

Assim, percebe-se que a curtose é mais robusta que a negentropia. Logo, segundo Neves (2012) é mais conveniente utilizar o método baseado em expectâncias para a aproximação da negentropia

$$J(y) \approx k_1 \left(E\left\{G_1(y)\right\}\right)^2 + k_2 \left(E\left\{G_2(y)\right\} - E\left\{G_2(v)\right\}\right)^2 \tag{3.14}$$

onde k_1 e k_2 são constantes positivas, v é uma variável gaussiana de média zero e variância unitária, e $G_1(y) = y^3$ e $G_2(y) = y^4$.

3.3.1.5 FastICA

O FastICA é um algoritmo proposto por Hyvärinen com o objetivo de estimar o vetor aleatório S de componentes independentes, através de uma matriz de separação W, inversa da matriz de mistura A (SANTOS; MONTESCO; JUNIOR, 2014).

O algoritmo FastICA foi desenvolvido como intuito de fornecer um método computacional mais rápido para a estimação das componentes independentes (MARCHINI; HEATON; RIPLEY, 2013). Assim, de acordo com o modelo generativo dado pela equação (3.2), as medidas em X tendem a ser mais gaussianas do que os componentes em S e por consequência, encontrar a matriz de separação W que maximiza a não-gaussianidade das fontes (ARAUJO; CAMPOS; FURTADO, 2014). Dessa forma, Esta aproximação pode ser dada da seguinte forma:

$$J_{G(y)} = |E_y \{G(y)\} - E_v \{G(v)\}|^2 \tag{3.15}$$

Onde G é uma função não linear, v é uma variável aleatória gaussiana normalizada, assume-se que y assumido como normalizado e com variância unitária. Algumas características do algoritmo FastICA são citadas abaixo:

- A convergência é cúbica (ou pelo menos quadrática), o que faz com que o algoritmo seja mais rápido em relação aos métodos baseados em gradiente descendente, onde sua convergência é linear.
- A não utilização de parâmetros faz com que a utilização do FastICA seja mais simples.
- A possibilidade de encontrar as componentes independentes de quase toda distribuição não gaussiana através da função não linear g.
- O desempenho pode ser otimizado de acordo com a escolha da função não linear g.
- Pode ser utilizado quando se quer estimar apenas algumas componentes independentes. Pois elas podem ser estimadas uma a uma.

3.4 Seleção de características mais significativas

A seleção de características é indispensável em situações onde se precisa distinguir um subgrupo de dados adequados dentre um grupos grande de características (CAMPOS, 2001).

Teoricamente, o maior número de características em uma base, serviria para uma discriminação melhor, algo que ná prática isso nem sempre acontece (PAPPA, 2002). A

definição da quantidade de características a serem utilizadas em um sistema de reconhecimento de padrões é extremamente importante, uma vez que esse procedimento favorece o desempenho do classificador, além de reduzir o custo computacional e o tempo na etapa da classificação (LINHARES *et al.*, 2016).

Sendo a dimensionalidade do espaço de características grande, isso pode gerar o fenômeno conhecido como a maldição da dimensionalidade (SANTOS; OLIVEIRA; DUTRA, 2005). A maldição da dimensionalidade está ligada ao aumento exponencial do esforço computacional devido a quantidade de características a serem estimadas.

A maioria dos algoritmos fazem a seleção de características apenas pelo critério de relevância, no entanto tem se observado que apenas esse critério de seleção não é suficiente (LEE; MONARD; WU, 2005). Pois além de selecionar as características relevantes se faz também necessário selecionar as características mais redundantes que precisam ser eliminadas do espaço de características.

Conforme Campos (2013) para se encontrar a caracterização ótima, o algoritmo deve buscar a melhor forma de encontrar este subespaço, sendo ela pela classificação por algum critério e selecionando as n melhores ou escolher um subconjunto das características pequeno que não afetem a precisão do classificador.

3.4.1 Máxima relevância e mínima redundância - mRMR

O algoritmo de Máxima Relevância e Mínima Redundância (mRMR) é utilizado para reduzir a quantidade de características de um dado conjunto, isto é, este algoritmo seleciona desse dado conjunto as características que são mais importantes e que menos se repetem. Essas características que não são selecionadas, são consideradas como irrelevantes e podem ser retiradas desse conjunto sem que haja uma alteração nos resultados finais da classificação, já que existem outras características que tem a mesma finalidade.

Para a seleção das características utiliza-se a medida de máxima relevância, e é feita através da informação mútua (I) entre as variáveis de cada característica v_i e de classe c

$$\max D\left(v, c\right), \quad D = \frac{1}{|v|} \sum_{v_i \in d} I\left(v_i; c\right) \tag{3.16}$$

onde v é o vetor de características.

E como é provável que as características selecionadas apresentem uma grande de-

pendência entre suas características (DING; PENG, 2005), utiliza-se a medida de mínima redundância para selecionar as características mutuamente exclusivas, ou seja,

$$\min R(v), \quad R = \frac{1}{|v|^2} \sum_{v_i, v_j \in v} I(v_i, v_j) \tag{3.17}$$

Portanto, a união dos métodos representados pelas equações (3.16) e (3.17) descritos acima, é denominada de Máxima Relevância e Mínima Redundância (PENG; LONG; DING, 2005). E dessa forma, define-se o operador $\Phi(D, R)$ para fazer a combinação entre D e R. E logo em seguida, para otimizá-los simultaneamente

$$\max \Phi(D, R), \quad \Phi = D - R \tag{3.18}$$

3.5 Classificação de imagens digitais - CID

A classificação de imagens é baseada nas características retidas dos objetos de estudo. As técnicas de aprendizagem de máquina para a classificação de informações obtidas através das características são denominadas de classificação supervisionada. As técnicas de classificação necessitam de dados rotulados para a realização do aprendizado a partir das características extraídas. Uma vez que as caraterísticas das imagens analisadas já foram extraídas e estão disponíveis, o classificador tem o objetivo de distingui-lo e agrupá-las em determinados grupos de acordo com suas semelhanças.

3.5.1 Análise discriminante linear - LDA

A proposta inicial de Fisher (1936), para a discriminação, e depois, a classificação entre dois ou mais grupos, era transformar as observações multivariadas, através de combinações lineares das variáveis originais, em observações univariadas de tal forma que, minimizasse a classificação equivocada de um indivíduo em uma população, sendo que este pertence à outra (REGAZZI, 2000). Sendo assim, de acordo com Cortivo (2015), a análise discriminante linear é uma técnica para a classificação supervisionada com o objetivo de encontrar um orientação para que as amostras projetadas sejam bem separadas.

Com base nas características dos grupos conhecidos as chamadas funções discriminantes são construídas com o objetivo de separar os grupos tanto quanto possível (FILZ-

MOSER; JOOSSENS; CROUX, 2006). A Análise Discriminante Linear, considera que a distribuição de probabilidade das amostras é conhecida e pode ser representada pela média e dispersão das amostras (XAVIER *et al.*, 2011). A função discriminante linear é uma combinação linear de características originais a qual se caracteriza por produzir separação máxima entre duas populações (VARELLA, 2008). Assim, sejam μ_i vetores de médias e $\sum$ matriz de covariâncias comuns das populações π_i, a função discriminante linear de Fisher de um vetor aleatório X que realiza a separação máxima entre as duas populações será dada por:

$$D(X) = L' * X = [\mu_1 - \mu_2]' * \sum^{-1} * X \tag{3.19}$$

sendo que,

$$X = [X_1 \; X_2 \; \ldots \; X_p] \quad e \quad \mu = [\mu_1, \mu_2] \tag{3.20}$$

onde

L é o vetor discriminante;

X é o vetor aleatório de características das populações;

μ é o vetor de médias p-variado;

$\sum$ é a matriz comum de covariâncias das populações π_1 e π_2.

O valor da função discriminante linear de Fisher para uma observação x_0 é dada por:

$$D(x_0) = [\mu_1 - \mu_2]' * \sum^{-1} * x_0 \tag{3.21}$$

O ponto médio entre as duas médias populacionais univariadas μ_1 e μ_2,

$$m = \frac{1}{2}[D(\mu_1) + D(\mu_2)] \tag{3.22}$$

Logo, o princípio de classificação das características baseada na função discriminante linear de Fisher é:

$$Alocar\ x_0\ em \begin{cases} \mu_1,\ se\ D(x_0) = [\mu_1 - \mu_2]' * \sum^{-1} * x_0 \geq m \\ \\ \mu_2,\ se\ D(x_0) = [\mu_1 - \mu_2]' * \sum_{-1} * x_0 < m \end{cases} \tag{3.23}$$

Supondo-se que as populações π_1 e π_2 tem a mesma matriz de covariâncias $\sum$, estima-se a matriz comum de covariâncias S_c:

$$S_c = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right] * S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right] * S_2 \tag{3.24}$$

onde,

S_c é a estimativa da matriz comum de covariâncias $\sum$;

n_1 é o número de observações da população π_1;

n_2 é o número de observações da população π_2;

S_1 é a estimativa da matriz de covariância da população π_1;

S_2 é a estimativa da matriz de covariância da população π_2;

Portanto, a função discriminante linear amostral de Fisher é obtida pela substituição dos parâmetros μ_1, μ_2 e $\sum$ e pelas respectivas quantidades amostrais $\overline{x_1}$ e $\overline{x_2}$ e S_c, ou seja,

$$D(x) = \hat{L}' * x = [\overline{x_1} - \overline{x_2}]' * S_c^{-1} * x \tag{3.25}$$

onde,

$D(x)$ e a função discriminante linear amostral de Fisher;

$\hat{L}'$ é a estimativa do vetor discriminante;

$\overline{x_1}$ é a média amostral da população π_1;

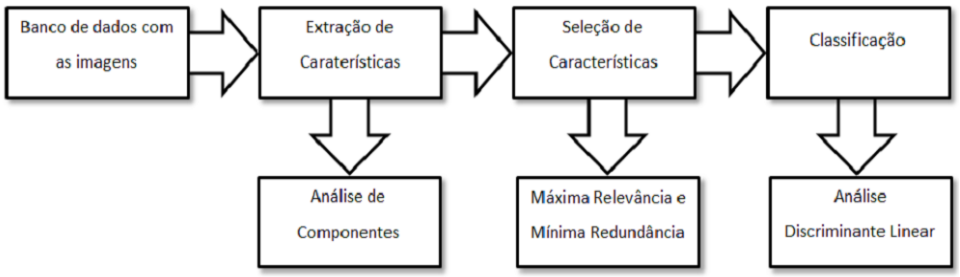
$\overline{x_2}$ é a média amostral da população π_2.

Capítulo 4

METODOLOGIA PROPOSTA E RESULTADOS

O método proposto foi desenvolvido a partir das técnicas demonstradas no capítulo anterior. O processo se inicia com a extração das características das imagens adquiridas utilizando a análise de ICA, que somada ao algoritmo mRMR para fazer a redução da dimensionalidade selecionando as características mais significantes e por fim, será utilizado a técnica LDA para fazer a classificação das imagens como melanoma ou não melanoma. Como mostra a figura (4.1).

Figura 4.1: Diagrama de blocos do método proposto



Fonte: (REIS, 2017)

4.1 Aquisição de dados

As bases de dados utilizadas neste método, são bases de domínio público e podem ser adquiridas em DermIS (2012) e DermQuest (2012). E são usadas para fornecer

informações sobre diagnósticos, relatos de casos e informações adicionais sobre inúmeras doenças de pele.

Foram selecionadas 206 imagens de câncer de pele, sendo que através do DermIs (2012), adquiriu-se 43 imagens de câncer de pele diagnosticas como Melanoma e 26 como não-Melanoma. Já em DermQuest (2012), obteve-se 76 imagens de câncer de pele diagnosticadas como Melanoma e 61 como não-Melanoma. No entanto, só foram utilizadas 204 imagens (duplicidade).

A figura (4.2) apresenta quatro amostras utilizadas na aplicação do método proposto:

Figura 4.2: Imagens adquiridas

(a) DermIs - melanoma

(b) DermIs - não-melanoma

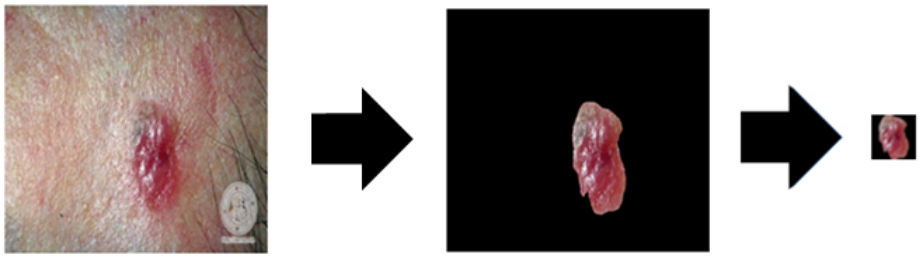
(c) DermQuest - melanoma

(d) DermQuest - não-melanoma

Fonte: (DERMIS, 2012), (DERMQUEST.COM, 2012)

Como as imagens originais possuíam muitas informações desnecessárias e dimensões diferentes, elas foram contornadas de acordo com a região de interesse e redimensionadas para que ficassem com 25 pixels de largura e 25 pixels de altura, ou simplesmente, redimensionadas para 25×25. Conforme mostra a figura (4.3).

Figura 4.3: Contorno da região de interesse e redimensionamento da imagem



Fonte: (REIS, 2017)

Após fazer o redimensionamento das imagens, foram eliminadas duas imagens, uma diagnosticada melanoma e a outra como não-melanoma, pois verificou-se que estavam duplicadas. Dessa forma, obteve-se uma matriz X_m, dos casos de melanoma, com 118 linhas e 525 colunas, ou apenas, com dimensão 118×525. E uma matriz X_{nm}, dos casos de não-melanoma, com dimensão 86×525.

4.2 Extração de características

Nesta etapa, o objetivo é extrair as características das Regiões de Interesse (ROI) que representem de maneira mais significante as diferenças entre os dados dos grupos analisados. Assim, essas características devem garantir que as Regiões de Interesse sejam classificadas corretamente como Melanoma ou não-Melanoma.

Para a extração desta características foi utilizada a Análise de Componentes Independentes (ICA), a imagem analisada é decomposta em uma combinação linear de imagens-base de modo que esta passa a ser representada por um vetor cujos elementos são os coeficientes de cada componente independente na mistura. Tais vetores são as características extraídas das imagens a partir da ICA (LEITE, 2013). Dessa forma de acordo com o modelo generativo dado pela equação (3.2), tem-se

$$x_i = a_{i1}s_1 + \ldots + a_{in}s_n \quad para\ todo \quad i = 1, \ldots, n \tag{4.1}$$

onde

- x_i - Dado aleatório da imagem analisada;

- a_{ij} - Coeficiente de mistura (reais);
- s_i - Imagens-base.

Seja

$$X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T \tag{4.2}$$

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \tag{4.3}$$

$$S = \begin{bmatrix} s_1 & \cdots & s_n \end{bmatrix}^T \tag{4.4}$$

Portanto, a Análise de Componentes Independentes permite derivar um algoritmo de seleção de variáveis independentes do modelo baseado em um teste de dependência estatística. A estratégia é aplicar a ICA para estimar a independência das entradas e então proceder um teste estatístico para determinar o subgrupo desejado de variáveis de entrada.

Dessa forma, a matriz X, de dimensão 204×525, utilizada no modelo ICA foi obtida pela união da matriz X_m, dos casos de melanoma, de dimensão 118×525, com a matriz X_{nm}, dos casos de não-melanoma, de dimensão 86×525. Assim, através da matriz X, e utilizando o algoritmo FastICA, obteve-se a matriz A, de dimensão 204×204, 204 linhas e 204 colunas, com as características de cada amostra. Sendo que, cada coluna da matriz A corresponde a uma característica, e cada linha a uma amostra.

4.3 Seleção das características mais significantes

Para fazer a seleção das características da matriz A foi aplicado, primeiramente, o algoritmo de máxima relevância, conforme com a equação (3.16), que organizou cada característica, em ordem decrescente de significância, ou seja, do mais significante para o menos significante. Logos após, como a dependência entre essas características ainda podem ser grandes, aplica-se o algoritmo de mínima redundância, pela equação (3.17), para que a redundância, isto é, a similaridade entre essas características sejam a menor

possível. E por fim, através da equação (3.18), fez-se a otimização para que a seleção das características obtivesse um resultado satisfatório.

Sendo assim, a redução do vetor de características de cada amostra foi realizada utilizando o algoritmo de Máxima Relevância e Mínima Redundância, ou seja, as características foram organizadas da mais significantes para as que menos se repetem.

Os dados da matriz foram distribuídos em ordem decrescente de representatividade, o que possibilita definir o número de características a serem utilizadas no classificador com a finalidade de encontrar o vetor de melhor desempenho (LINHARES *et al.*, 2016).

As características irrelevantes podem ser removidas sem comprometer o resultado da classificação, pois neste contexto, são consideradas redundantes, ou seja, implicam na presença de outra característica com a mesma funcionalidade, e não trazem nenhuma informação nova ao vetor de características (ARAUJO; CAMPOS; FURTADO, 2014).

Os testes foram realizados incrementando 1 característica para cada iteração.

4.4 Classificação

A etapa de classificação das amostras foi feita utilizando a Análise Discriminante Linear, isto é, a partir de um conjunto de imagens de treinamento, gera um conjunto de vetores característicos que representa a imagem no espaço de características (SILVA, 2016).

A Análise Discriminante Linear, considera que a distribuição de probabilidade das amostras é conhecida e pode ser representada pela média e dispersão das amostras (XAVIER *et al.*, 2011). Com isso, cria-se uma função discriminante linear para produzir, de acordo com suas características, uma máxima separação entre os subgrupos de imagens com melanoma e não melanoma.

Como a técnica LDA busca preservar o máximo de informação discriminatória possível dos dados originais, obtêm-se a direção na qual a discriminação entre as classes se mantém a maior possível (NETO, 2015). Sendo assim, visando encontrar um sistema de coordenadas ótimo, a LDA aplica uma transformação linear que garanta a máxima separabilidade entre as classes do conjunto de dados (SANTOS, 2005).

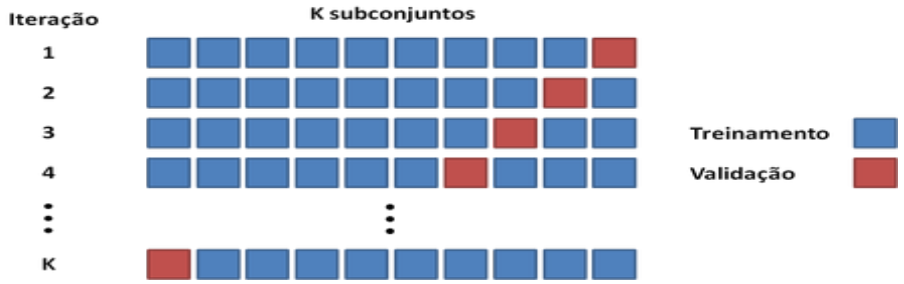
Por esse motivo, a técnica de análise discriminante linear foi utilizada na classificação das imagens. As características extraídas das imagens através da ICA e devidamente

selecionado pelo algoritmo mRMR serviram como dados de entrada para a análise de discriminante linear.

Com base nesse conjunto de características, que equivalem às 204 imagens previamente classificadas por meio de análise visual, criou-se a função discriminante de Fisher, de acordo com a equação (3.19).

Após a classificação da Análise Discriminante Linear foi utilizada a técnica de validação cruzada k-fold. A validação cruzada k-fold baseia-se na divisão em k subconjuntos mutuamente exclusivos de um conjunto total de dados, onde são utilizados $(k-1)$ subconjuntos para a estimação dos parâmetros, ajuste do classificador, e 1 para avaliação (JUNIOR, 2016).

Figura 4.4: Técnica de validação cruzada



Fonte: (COUTO, 2013)

Veja, pela figura (4.4), que o processo se repete k vezes, alternando os subconjuntos de tal forma que em cada iteração, um subconjunto seja utilizado como avaliação.

A técnica tem como característica fazer uma melhor varredura da base de dados, atenuando dessa maneira problemas causados por diferenças na base (JR, 2011).

Portanto, as amostras foram divididas em 10 subconjuntos, onde 6 continham 20 amostras e 4 continham 21 amostras, com o objetivo de realizar o teste de validação cruzada 10-fold cross validation, ou seja, 9 subconjuntos foram utilizados para o ajuste do classificador e 1 para a avaliação.

A tabela (4.1) mostra os resultados obtidos no processo de classificação, para cada algumas características analisadas.

Tabela 4.1: Desempenho do classificador

Características	VP	VN	FP	FN	Acc (%)	Sens (%)	Spec (%)
20	101	37	17	49	67,65	67,33	68,52
40	101	48	17	38	73,04	72,66	73,85
60	100	55	18	31	75,98	76,34	75,34
80	100	60	18	26	78,42	79,36	76,92
100	104	70	14	16	85,29	86,67	83,33
150	114	83	4	3	96,57	97,44	95,40
185	**118**	**86**	**0**	**0**	**100**	**100**	**100**
204	118	86	0	0	100	100	100

Fonte: (REIS, 2017)

Dessa forma, analisando a tabela (4.1), pode-se verificar que de acordo com o aumento quantidade de características, aumenta-se também as porcentagens da acurácia, especificidade e sensibilidade. E para os vetores de 185 características, a acurácia é de 100, a especificidade de 100% e a sensibilidade de 100%. Sendo que a partir do vetor deste vetor de 185 características, foi possível concluir que todas as 204 amostras, foram classificadas corretamente, sendo 118 melanomas (Verdadeiro-Positivo) e 86 não melanomas (Verdadeiro-Negativo), obtendo assim, 100% de acurácia, especificidade e sensibilidade.

4.5 Métricas e desempenho

A matriz de confusão foi utilizada para exibir uma medida eficiente de classificação. Uma vez que mostra o número de classificação correta para cada grupo determinado (MATOS *et al.*, 2009).

A matriz de confusão é uma matriz quadrada de números definidos em linhas e colunas que expressam o número de unidades da amostra atribuído a uma categoria particular relativa à categoria atual (SUAREZ; CANDEIAS, 2012). A diagonal principal da matriz de confusão exibe os dados classificados corretamente.

A qualidade dos testes apresentados, foi feita através dos cálculos das medida de acurácia, especificidade e sensibilidade. Onde a acurácia (Acc) é a capacidade do método acertar o diagnóstico, a sensibilidade (Sens) é a capacidade do método em reconhecer os doentes, e a especifidade (Spec) é a capacidade do método reconhecer os saudáveis.

Para tais medições, são utilizadas as seguintes fórmulas:

$$Acc = \frac{V_P + V_N}{V_P + V_N + F_P + F_N} \tag{4.5}$$

$$Sens = \frac{V_P}{V_P + V_N} \tag{4.6}$$

$$Spec = \frac{V_N}{V_N + F_P} \tag{4.7}$$

sendo,

V_P é o número de verdadeiros positivos.

V_N é o número de verdadeiros negativos.

F_P é o número de falsos positivos.

F_N é o número de falsos negativos.

Capítulo 5

DISCUSSÕES

Com relação ao diagnóstico das imagens utilizadas, previamente classificadas por especialistas, o método proposto nesse trabalho se mostrou satisfatório, visto que, com o aumento das características analisadas a acurácia, a sensibilidade e a especificidade alcançaram uma taxa de acerto de 100%.

Para efeito de comparação com os trabalhos relacionados, pode-se dizer que o método proposto obteve resultados melhores. Mas não é possível afirmar que esta técnica é mais eficiente que as outras, tendo em vista que a utilização das bases de dados de imagens foram diferentes.

Sobre a relação com o diagnóstico clínico, esta técnica, é uma tentativa de encerrar as dificuldades encontradas pelos especialistas no diagnóstico do tumor. Uma vez que este diagnóstico é feito em várias etapas, onde o especialista deve fazer uma análise cuidadosa baseando-se apenas nas informações dos pacientes e no seu conhecimento sobre o caso.

Em relação ao diagnóstico dermatoscópico, o método proposto busca auxiliar o especialista na detecção precoce do câncer de pele, tendo em vista, em algumas cidades brasileiras não disponibilizam dos aparelhos necessários para a realização deste exame.

Como a realização de uma biópsia que é um procedimento invasivo necessário para o estudo histopatológico e causa inúmeros transtornos aos paciente. A aplicabilidade desta técnica pode ser de extrema importância para que o diagnóstico do câncer de pele seja realizado sem danos extras.

Considerando que a detecção precoce é fundamental para se obter a cura do câncer de pele. Diante de todas as situações citadas acima, o método para o diagnóstico do câncer de pele apresentado neste trabalho, visa facilitar a detecção deste tipo de câncer

no seu estágio inicial e assim, proporcionar o tratamento adequado aos pacientes.

Capítulo 6

CONCLUSÃO

Nessa dissertação foi utilizada uma técnica para auxiliar o diagnóstico precoce de lesões de câncer de pele denominada de Diagnóstico Auxiliado por Computador, onde foi utilizada um total de 204 imagens de lesões de câncer de pele previamente diagnosticadas como melanoma e não melanoma. Para tanto, empregou-se a análise de componentes independentes para a extração de características das imagens, para reduzir a dimensionalidade e selecionar as características mais significantes e menos redundantes foi utilizado o algoritmo de máxima relevância e mínima redundância. A classificação das imagens previamente diagnosticadas foi realizada pela Análise Discriminante Linear e a validação do método utilizado se deu a partir da validação cruzada k-fold.

Os resultados apresentados na seção (4.4), demonstram que a técnica utilizada alcançou um desempenho satisfatório. Uma vez que a técnica ICA se mostrou eficiente para a extração das características das imagens, o algoritmo de mRMR selecionou as características de forma que a classificação pela técnica LDA fosse de maneira mais precisa. Com um vetor de 185 características, a técnica obteve uma acurácia de 100%, a especificidade de 100% e a sensibilidade de 100%. E a partir do aumento das características, a técnica continuou se mostrando bastante eficaz.

Apesar dos bons resultados previamente apresentados, outros testes deverão ser realizados em outras bases de dados, para que dessa forma, aumente a confiabilidade da técnica proposta. Obtendo assim, uma melhor compreensão no diagnóstico e visando o desenvolvimento de *softwares* para que sejam testados em clínicas e hospitais auxiliando na diminuição da taxa de mortalidade para este e outros tipos de câncer.

Diante do resultado satisfatório obtido este trabalho foi submetido para a Revista

de Ciências da Computação e para a Revista de Engenharia de Computação e Sistemas Digitais.

REFERÊNCIAS BIBLIOGRÁFICAS

ABIB, G. d. C. A. *et al.* **Separação cega de fontes acústicas em ambientes com reverberação: testes e análises.** 2013.

ACOSTA, Á. E. *et al.* **Melanoma: patogénesis, clínica e histopatología**. *Asociación Colombiana de Dermatología y Cirugía Dermatológica*, p. 85, 2008.

ALMEIDA, A. B. d. **Usando o computador para processamento de imagens médicas**. *Revista Informática Médica*, v. 1, n. 6, 1998.

ARAUJO, W. B. D.; CAMPOS, L. F. A.; FURTADO, A. S. **Método de detecção de câncer de ovário utilizando padrões proteômicos, análise de componentes independentes e máquina de vetores de suporte**. In: *XIV Workshop de Informática Médica*. [S.l.: s.n.], 2014. v. 14.

AVILA, M.; CRUZ, C. d. O.; RIERA, R. **Evidências de revisões sistemáticas cochrane sobre prevenção e tratamento de melanoma**. *Revista Diagnóstico e Tratamento*, v. 21, n. 2, p. 84, 2016.

AZEVEDO-MARQUES, P. M. de. **Diagnóstico auxiliado por computador na radiologia**. *Radiologia Brasileira*, SciELO Brasil, v. 34, n. 5, p. 285–293, 2001.

BARCELOS, C. A. Z.; PIRES, V. **An automatic based nonlinear diffusion equations scheme for skin lesion segmentation**. *Applied Mathematics and Computation*, Elsevier, v. 215, n. 1, p. 251–261, 2009.

BIANCHI, M. F. de. ***Extração de características de imagens de faces humanas através de wavelets, PCA e IMPCA***. Tese (Doutorado) — Universidade de São Paulo, 2006.

BRASIL. ***Ministério da Saúde - Secretaria de Atenção à Saúde. Portaria n. 357, de 8 de abril de 2013***. [S.l.], 2013. Disponível em: <http://bvsms.saude.gov.br/bvs/saudelegis/sas/2013/prt0357_08_04_2013.html>. Acesso em: 14 fev. 2017.

BRYS, L. M. ***Página dinâmica para aprendizado do sensoriamento remoto***. Tese (Doutorado) — Universidade Federal do Rio Grande do Sul, 2008.

CÂMARA, G. *et al.* **SPRING: Integrating remote sensing and GIS by object-oriented data modelling**. *Computers & graphics*, Elsevier, v. 20, n. 3, p. 395–403, 1996.

CAMPOS, L. F. d. A. ***Método de deteção de câncer em mamas densas Utilizando diagnóstico auxiliado por Computador***. Tese (Doutorado) — Programa de Pós-Graduação em Biotecnologia - RENORBIO, 2013.

CAMPOS, T. E. de. ***Técnicas de seleção de características com aplicações em reconhecimento de faces***. Tese (Doutorado) — Universidade de São Paulo, 2001.

CAPOBIANCO, J.; JASINOWODOLINSKI, D.; SZARF, G. **Diagnóstico auxiliado por computador na detecção de nódulos pulmonares pela tomografia computadorizada com múltiplos detectores: estudo preliminar de 24 casos**. *J Bras Pneumol*, SciELO Brasil, v. 34, n. 1, p. 27–33, 2008.

CESTARI, M. E. W.; ZAGO, M. M. F. **A prevenção do câncer e a promoção da saúde: um desafio para o Século XXI**. *Rev Bras Enferm*, v. 58, n. 2, p. 218–21, 2005.

CO-LOCALIZADA, M. ***Novas possibilidades na caracterização de minérios***. 41–45 p. Tese (Doutorado) — PUC-Rio, 2007.

COFEN, C. F. d. E. ***Fisiopatologia do câncer***. 2015. Disponível em: <http://biblioteca.cofen.gov.br/wp-content/uploads/2015/03/cap2-fisiopatologia-do-cancer.pdf>. Acesso em: 14 fev. 2017.

COMON, P. **Independent component analysis, a new concept?** *Signal processing*, Elsevier, v. 36, n. 3, p. 287–314, 1994.

COUTO, E. ***Bias vs. variância (Parte 2)***. 2013. Disponível em: <https://ericcouto.wordpress.com/2013/07/18/bias-vs-variancia-parte-2/>. Acesso em: 10 set. 2016.

COVER, T. M.; THOMAS, J. A. ***Elements of information theory***. [S.l.]: John Wiley & Sons, 2012.

DARMOIS, G. **Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire**. *Revue de l'Institut international de statistique*, JSTOR, p. 2–8, 1953.

DERMATOLÓGICA, S. B. de C. ***Dermatoscopia***. Disponível em: <https://www.sbcd.org.br/procedimentos/390>. Acesso em: 14 fev. 2017.

DERMIS, S. D. I. ***Online atlases***. 2012. Disponível em: <http://dermis.net/dermisroot/en/home/index.htm>. Acesso em: 10 set. 2016.

DERMQUEST.COM. ***Online database***. 2012. Disponível em: <https://www.dermquest.com/>. Acesso em: 10 set. 2016.

DIAS, D. M. B. **Separação de sinais gaussianos e não gaussianos usando ica e beamforming**. 2014.

DIEPGEN, T. *et al.* **Occupational skin cancer induced by ultraviolet radiation and its prevention**. *British Journal of Dermatology*, Wiley Online Library, v. 167, n. s2, p. 76–84, 2012.

DING, C.; PENG, H. **Minimum redundancy feature selection from microarray gene expression data**. *Journal of bioinformatics and computational biology*, World Scientific, v. 3, n. 02, p. 185–205, 2005.

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

EDUCAÇÃO, P. D. ***Histopatologia: o que é?*** 2013. Disponível em: <https://www.portaleducacao.com.br/medicina/artigos/53117/histopatologia-o-que-e>. Acesso em: 14 fev. 2017.

ESTATSITE.COM. ***Curtose***. 2016. Disponível em: <https://estatsite.com/tag/estatistica-descritiva/>. Acesso em: 10 set. 2016.

FILZMOSER, P.; JOOSSENS, K.; CROUX, C. **Multiple group linear discriminant analysis: robustness and error rate**. In: *Compstat 2006-Proceedings in Computational Statistics*. [S.l.]: Springer, 2006. p. 521–532.

FRANCO, A. L. **Aplicação da análise de componentes independentes em estudo de eventos em finanças**. 2008.

FURGERI, S.; RODRIGUES, S. C.; SILVA, S. M. da. **Tecnologias associadas ao diagnóstico do Câncer de Mama**. *Reverte-Revista de Estudos e Reflexões Tecnológicas da Faculdade de Indaiatuba*, n. 11, 2013.

HERAULT, J.; JUTTEN, C. **Space or time adaptive signal processing by neural network models**. In: AIP PUBLISHING. *Neural networks for computing*. [S.l.], 1986. v. 151, n. 1, p. 206–211.

HYVÄRINEN, A.; KARHUNEN, J.; OJA, E. ***Independent component analysis***. [S.l.]: John Wiley & Sons, 2004. v. 46.

HYVÄRINEN, A.; OJA, E. **Independent component analysis: algorithms and applications**. *Neural networks*, Elsevier, v. 13, n. 4, p. 411–430, 2000.

INCA. **Estimativa 2016: incidência de câncer no Brasil**. *Instituto Nacional de Câncer José Alencar Gomes da Siva. Coordenação de Prevenção e Vigilância*, Ministério da Saúde, v. 11, 2016.

INCA, I. N. d. C. **Prevenção do câncer da pele**. *Revista Brasileira de Cancerologia*, v. 49, n. 4, p. 203, 2003.

JR, C. S. **Ajuste de análise discriminante linear semi-supervisionada através de salidação cruzada ponderada**. *Proceedings Seminário Interno da disciplina de Reconhecimento de Padrões*, p. 103, 2011.

JUNIOR, E. E. R. **Estratégias para Classificação Binária Um estudo de caso com classificação de e-mails**. 2016.

KAGAN, A. M.; RAO, C. R.; LINNIK, Y. V. **Characterization problems in mathematical statistics**. Wiley, 1973.

LEE, H. D.; MONARD, M. C.; WU, F. C. **Seleção de atributos relevantes e não redundantes usando a dimensão fractal do conjunto de dados**. In: *Anais do V Encontro Nacional de Inteligência, XXV Congresso da Sociedade Brasileira de Computação, Porto Alegre, RS*. [S.l.: s.n.], 2005. p. 444–453.

LEITE, I. C. C. **Análise de componentes independentes aplicada a avaliação de imagem radiográfica de sementes**. Universidade Federal de Lavras, 2013.

LEITE, L. ***Análise de componentes independentes aplicada à identificação de regiões lesionadas em mamogramas***. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2005.

LINHARES, J. d. N. *et al.* **Método computacional para o diagnóstico precoce da granulomatose de wegener**. *Revista de Informática Teórica e Aplicada*, v. 23, n. 1, p. 277–292, 2016.

MARCHINI, J.; HEATON, C.; RIPLEY, B. **Fastica: Fastica algorithms to perform ica and projection pursuit**. *R package version*, p. 1–2, 2013.

MARIANTE, J. C. S. ***Alterações Clínicas, Dermatoscópicas, Histopatológicas e Imuno-histoquímicas de Nevos Melanocíticos Irradiados com Raios Ultravioleta B***. Tese (Doutorado) — CLÍNICA MÉDICA, 2003.

MARTINEZ, A. C. ***Desenvolvimento de novas técnicas para redução de falso-positivo e definição automática de parâmetros em esquemas de diagnóstico auxiliado por computador em mamografia***. Tese (Doutorado) — Universidade de São Paulo, 2007.

MATOS, P. F. *et al.* ***Relatório técnico "Métricas de avaliação"***. São Carlos, Brasil, 2009. Disponível em: <http://conteudo.icmc.usp.br/pessoas/taspardo/TechReportUFSCar2009a-MatosEtAl.pdf>. Acesso em: 10 set. 2016.

MORIMITSU, H.; TUESTA, E. F. **Análise comparativa das abordagens de estimativa do modelo FastICA por maximização da negentropia e da verossimilhança**. 2015.

MÜLLER, K. R. *et al.* **Avaliação do aprendizado dos pacientes sobre a regra do ABCD: um estudo randomizado no sul do Brasil**. *An. bras. dermatol*, v. 84, n. 6, p. 593–598, 2009.

NETO, D. A. S. ***Análise da assimetria e irregularidade de borda entre lesões melanocíticas***. Tese (Doutorado) — Universidade de São Paulo, 2015.

NEWSWIRE, P. ***Campanha skinchecker***. 2015. Disponível em: <http://www2.prnewswire.com.br/imgs/pub/2015-08-14/original/2569.jpg>. Acesso em: 14 fev. 2017.

NIE, D. **Classification of melanoma and clark nevus skin lesions based on medical image processing techniques**. In: IEEE. *Computer Research and Development (ICCRD), 2011 3rd International Conference on*. [S.l.], 2011. v. 3, p. 31–34.

PAPPA, G. ***Seleção de atributos usando algoritmos genéticos multiobjetivos***. Tese (Doutorado) — Dissertação (Mestrado em Informática Aplicada)-Pontifícia Universidade Católica do Paraná. Curitiba, 2002.

PATWARDHAN, S. V.; DHAWAN, A. P.; RELUE, P. A. **Classification of melanoma using tree structured wavelet transforms**. *Computer Methods and Programs in Biomedicine*, Elsevier, v. 72, n. 3, p. 223–239, 2003.

PENG, H.; LONG, F.; DING, C. **Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy**. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 27, n. 8, p. 1226–1238, 2005.

PEREIRA, G. A. ***Câncer de pele***. 2012. Disponível em: <http://www.cancerdepele.net.br/cancer-de-pele>. Acesso em: 10 set. 2016.

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.

QUEIROZ, J. E. R. de; GOMES, H. M. **Introdução ao processamento digital de imagens**. *RITA*, v. 13, n. 2, p. 11–42, 2006.

REGAZZI, A. J. ***Análise multivaria da dados***. 2000. INF-766 - notas de aula.

REZZE, G. G.; SA, B. C. S. D.; NEVES, R. I. **Dermatoscopia: o método de análise de padrões**. *Anais brasileiros de dermatologia*, Sociedade Brasileira de Dermatologia, v. 81, n. 3, p. 261–268, 2006.

ROBBINS, S. L. *et al.* **Fundamentos de robbins: patologia estrutural e funcional**. In: *Fundamentos de robbins: patologia estrutural e funcional*. [S.l.]: Guanabara Koogan, 2001.

RODRIGUES, C. I. **Sistemas CAD em Patologia Mamária**. *Faculdade de Engenharia da Universidade do Porto, Setembro*, 2008.

ROSADO, L. **Sistema automático para diagnóstico de lesões cutâneas baseado em imagens dermoscópicas**. *Mestrado, Instituto Superior Técnico-Engenharia Biomédica, Universidade Técnica de Lisboa, Lisboa*, 2009.

SANTOS, A. R. d. ***Identificação de faces humanas através de PCA-LDA e redes neurais SOM***. Tese (Doutorado) — Universidade de São Paulo, 2005.

SANTOS, H. C. d.; MONTESCO, C. A. E.; JUNIOR, M. C. R. **Classificação de sinais EGG combinando redes neurais e análise de componentes independentes**. 2014.

SANTOS, J. C.; OLIVEIRA, J. R. d. F.; DUTRA, L. V. **Uso de Algoritmos Genéticos na Seleção de Atributos para Classificação de Regiões**. In: *GeoInfo*. [S.l.: s.n.], 2005. p. 253–261.

SEIXAS, F. L.; SAADER, D. C. M. ***Diagnóstico Auxiliado por Computador***. Rio de Janeiro, Brasil, 2005. Disponível em: <http://www.midiacom.uff.br/~debora/fsmm/trab-2005-2/CAD.pdf>. Acesso em: 10 set. 2016.

SILVA, A. L. **Redução de características para clssificação de imagens de faces**. 2016.

SILVA, A. P. da. ***Separação Cega de misturas convolutivas no domínio do tempo utilizando clusterização***. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2009.

SILVA, A. P. O. d. **Uma implementação da análise de componentes independentes em plataforma de hardware reconfigurável**. Universidade Federal do Rio Grande do Norte, 2010.

SILVA, C. T. X. *et al.* **Análise dos fatores epidemiológicos, clínico-patológicos e expressão das proteínas OCT4 e NANOG em amostras de melanoma cutâneo**. Universidade Federal de Goiás, 2016.

SILVA, S. F. d. ***Dermatology atlas***. 2012. Disponível em: <http://www.atlasdermatologico.com.br>. Acesso em: 14 fev. 2017.

STIVAL, C. O. *et al.* **Avaliação comparativa da citopatologia positiva, colposcopia e histopatologia: destacando a citopatologia como método de rastreamento do câncer do colo do útero**. *Rev Bras Anal Clin*, v. 37, p. 215–8, 2005.

SUAREZ, A. F.; CANDEIAS, A. L. B. **Avaliação de acurácia da classificação de dados de sensoriamento remoto para o município de maragogipe**. 2012.

VARELLA, C. A. A. **Análise Discriminante**. *Análise Multivariada Aplicada às Ciências Agrárias*, 2008.

VENNILA, G. S.; SURESH, L. P.; SHUNMUGANATHAN, K. **Dermoscopic image segmentation and classification using machine learning algorithms**. In: IEEE. *Computing, Electronics and Electrical Technologies (ICCEET), 2012 International Conference on*. [S.l.], 2012. p. 1122–1127.

WAINSTEIN ALBERTO; BELFORT, F. ***Melanoma: Prevenção, diagnóstico, tratamento de acompanhamento***. 2. ed. [S.l.]: Atheneu, 2014.

XAVIER, A. C. *et al.* **Análise discriminante e classificação de imagens 2d de ultrassonografia mamária**. In: CITESEER. *VII Workshop de Visão Computacional*. [S.l.], 2011. p. 67–72.

YUAN, X. *et al.* **SVM-based texture classification and application to early melanoma detection**. In: IEEE. *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. [S.l.], 2006. p. 4775–4778.