ARTHUR DIEGO DIAS ROCHA

Detecção e classificação de lesões em imagens de mamografia usando classificadores SVM, wavelets morfológicas e seleção de atributos

Recife

2016

ARTHUR DIEGO DIAS ROCHA

Detecção e classificação de lesões em imagens de mamografia usando classificadores SVM, wavelets morfológicas e seleção de atributos

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Engenharia Biomédica da Universidade Federal de Pernambuco como parte dos requisitos para a obtenção do título de Mestre em Engenharia Biomédica.

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 22 de fevereiro de 2016. A versão original encontra-se em acervo reservado na Biblioteca do CTG-UFPE e na Biblioteca Central da UFPE.

Orientador: Prof. Dr. Wellington Pinheiro dos Santos

Co-orientador: Prof. Dr. Ricardo Emmanuel de Souza

Recife

2016

Catalogação na fonte
Bibliotecária Maria Luiza de Moura Ferreira, CRB-4 / 1469

R672d Rocha, Arthur Diego Dias.
Detecção e classificação de lesões em imagens de mamografia usando classificadores SVM, wavelets morfológicas e seleção de atributos / Arthur Diego Dias Rocha. 2016.
106 folhas, il.

Orientador: Prof. Dr. Wellington Pinheiro dos Santos.
Coorientador: Prof. Dr. Ricardo Emmanuel de Souza.

Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-graduação em Engenharia Biomédica, 2016.
Inclui Referências.

1. Engenharia Biomédica. 2. Mamografia. 3. Seleção de atributos. 4. Câncer de mama. 5. Processamento de imagem. I. Santos, Wellington Pinheiro dos (Orientador). II. Souza, Ricardo Emmanuel de (Coorientador). III. Título.

610.28 CDD (22. ed.) UFPE/BCTG/2016-107

*ATA DA **DÉCIMA NONA** DEFESA DE **DISSERTAÇÃO DE MESTRADO**, DO PROGRAMA DE PÓS-GRADUAÇÃO EM **ENGENHARIA BIOMÉDICA** DO CENTRO DE **TECNOLOGIA E GEOCIÊNCIAS** DA UNIVERSIDADE FEDERAL DE PERNAMBUCO, NO DIA 22 DE FEVEREIRO DE 2016.*

Aos 22 (vinte e dois) dias do mês de fevereiro de dois mil e dezesseis (2016), às 14 horas e 30 minutos, na sala 2, do Programa de Engenharia Biomédica, área 2, da Universidade Federal de Pernambuco, em sessão pública, teve início a defesa da Dissertação de Mestrado intitulada "AVALIAÇÃO DE TÉCNICAS DE REDUÇÃO DE ATRIBUTOS PARA CLASSIFICAÇÃO DE LESÕES EM IMAGENS DE MAMOGRAFIAS", do(a) aluno(A) **ARTHUR DIEGO DIAS ROCHA**, na área de concentração de Computação Biomédica, sob a orientação do Prof. Dr. Wellington Pinheiro dos Santos. O(A) mestrando(a) Arthur Diego Dias Rocha cumpriu todos os demais requisitos regimentais para a obtenção do grau de Mestre em Engenharia Biomédica. A Banca Examinadora foi aprovada pelo colegiado do programa de pós-graduação Em Engenharia Biomédica, na sua 01º Reunião ordinária, ocorrida em 19 de janeiro de 2016 e homologada pela Diretoria de Pós-Graduação, através do Processo Nº. 23076.003338/2016-19 em 25/01/2016, composta pelos Professores: Prof. Dr. Wellington Pinheiro dos Santos, do Departamento de Engenharia Biomédica da Universidade Federal de Pernambuco – UFPE, na qualidade de presidente, Prof. Dr. Ricardo Emmanuel de Souza, do Departamento de Engenharia Biomédica da Universidade Federal de Pernambuco – UFPE, a Profª. Drª. Alana Elza Fontes da Gama, do Departamento de Engenharia Biomédica da Universidade Federal de Pernambuco – UFPE. Após cumpridas as formalidades, o (a) candidato (a) foi convidado (a) a discorrer sobre o conteúdo da Dissertação. Concluída a explanação, o/a candidato/a foi argüido pela Banca Examinadora que, em seguida, reuniu-se para deliberar e conceder ao mesmo a menção (Aprovado/Reprovado/em exigência) APROVADO da referida Dissertação. A Banca Examinadora sugeriu que o título da dissertação fosse alterado para "DETECÇÃO E CLASSIFICAÇÃO DE LESÕES EM IMAGENS DE MAMOGRAFIA USANDO CLASSIFICADORES SVM, WAVELETS MORFOLÓGICAS E SELEÇÃO DE ATRIBUTO". E, para constar, lavrei a presente Ata que vai por mim assinada, Secretário (a) de Pós-Graduação, e pelos membros da Banca Examinadora.

Recife, 22 de fevereiro de 2016.

Juliana Alves de Farias – Secretária PPGEB

BANCA EXAMINADORA

Prof. Wellington Pinheiro dos Santos

Prof. Ricardo Emmanuel de Souza

Profª. Alana Elza Fontes da Gama

Dedico esta dissertação a uma pessoa que sempre demostrou ter todo o orgulho de mim, pelas minhas conquistas na vida. Sempre mencionando, onde quer que fosse, toda a satisfação que tinha por mim e por toda a família. Sempre me apoiou na minha vida acadêmica, minha graduação, durante o meu mestrado, orgulho dobrado por eu fazer parte de um grupo de robótica e só tenho a agradecer por ter me ajudado a entrar na minha segunda graduação, que tanto gosto. Foi quem me deu os meus primeiros intrumentos musicais e que acabou iniciando a paixão enorme que eu tenho pela música. Foi guia em muitas viagens pelo país inteiro com seu velho hábito de fazer muita graça com lugares, pessoas e costumes diferentes, sempre com a maior alegria. Uma pessoa que mesmo quando já não era mais tanto ela mesma, foi um dos maiores apoios em mais uma conquista na minha vida, quando fui morar com minha noiva, futura esposa. Esta é uma dedicatória em memória de uma pessoa que se foi e faz muita falta. Exatamente no meio do ano de 2015, o meu pai deixou esta Terra e foi em busca do seu tão procurado paraíso. O que me entristece é o fato de que ele nunca me verá concluir o mestrado, nunca me verá trabalhando em um emprego que me faça feliz, nunca verá eu me casar e pior ainda nunca poderá algum dia ver o neto que ele tanto queria ter. Este trabalho é dedicado em memória do meu pai, Alexandre Rocha, que se foi cedo, mas que deixou muitas lembranças boas e realizou muitas obras positivas na minha vida e na vida da nossa família.

Agradecimentos

São várias as pessoas que passam pela vida, e que contribuem de alguma forma com alguma coisa, por menor que seja, no nosso ser. E a todas essas pessoas só tenho a agradecer. Agradeço a minha esposa, Bianca, por toda a paciência em todos os sentidos possíveis, por ter sempre ficado ao meu lado nos momentos bons, nos ruins e nos vários momentos de dificuldade, sempre demonstrando todo o amor que possui por mim e que eu espero ser capaz de demonstrar redobrado e por toda a vida. Agradeço à minha mãe Maria e meu irmão Lucas por sempre estarem presentes na minha vida, mesmo depois de ter me mudado, continuamos sempre próximos e em sintonia para que coisas boas apareçam na vida. Agradeço à minha avó Letícia e ao meu avô Rubem por sempre me apoiarem em tudo e por compreenderem a minha falta de tempo por causa dos estudos. Agradeço a toda a minha família, tios, primos, e todos os outros. Agradeço a minha sogra Hosani, por ser uma ótima sogra e bastante preocupada com o meu bem-estar e da minha família também. Agradeço ao meu orientador Wellington por toda a paciência e seu jeito tranquilo de ser, que me transmitiu todos os ensinamentos, sempre me orientando com toda a sabedoria, contribuindo massivamente para a minha formação acadêmica e pessoal. Agradeço à FACEPE pelo financiamento desta pesquisa. Agradeço aos meus colegas de mestrado e graduação e, também, a todos aqueles que de alguma forma, por menor que seja, trouxeram alguma contribuição na minha vida. Obrigado.

"Eu preferiria ser um fracasso em algo que amo do que um sucesso em algo que odeio."

(George Burns)

Resumo

ROCHA, Arthur Diego Dias. **Detecção e classificação de lesões em imagens de mamografia usando classificadores SVM, wavelets morfológicas e seleção de atributos**. 2016. 108 f.

O câncer de mama é o mais comum entre as mulheres no mundo e no Brasil, depois do de pele não melanoma. De acordo com o Instituto Nacional de Câncer, em 2013 foram registradas 14.388 mortes devido a esta moléstia. O câncer de mama é uma preocupação não somente nacional, mas mundial. O método utilizado para a sua detecção é a mamografia, que é uma técnica de imagem que utiliza a emissão Raios-X incidentes na mama e capta a parte da radiação não absorvida pelos tecidos mamários. A mamografia é um exame de difícil análise pelo motivo de, em muitos casos, a densidade tecidual do tumor ser bastante parecida com a densidade de alguns tecidos saudáveis da mama. Uma abordagem interessante é a utilização de técnicas computadorizadas de auxílio ao diagnóstico, ou seja, ferramentas baseadas em processamento de imagens e inteligência computacional projetadas para o apoio ao profissional radiologista. Estudos prévios demonstram que considerar a dominância tecidual mamária nas ferramentas computacionais de apoio ao diagnóstico melhora consideravelmente as taxas de acerto. Para este trabalho, é proposta a construção de um sistema de classificação de tumores de mama baseado descritores de *Zernike* como um descritor de forma das lesões de mama, associado às máquinas de vetor de suporte como classificador. São comparadas diferentes técnicas de seleção de atributos com o objetivo de reduzir o custo computacional do sistema, mas sempre levando em conta a necessidade de se manter altas taxas de acerto, já que isto pode refletir em erros de diagnóstico de câncer de mama. Através dos dados analisados, é notado que a técnica linear de análise de componentes principais (aliada à transformada de *wavelets* morfológica como etapa de pré-processamento) se mostrou uma ótima técnica para realização de redução de atributos com um menor impacto nas taxas de acerto do sistema de apoio ao diagnóstico do câncer de mama, onde são obtidas taxas de médias de redução de acerto em torno de 2% (uma queda média de aproximadamente 95% para 93%), onde a redução do tamanho do vetor de atributos é de cerca de 64% (dentre os diferentes tipos de tecido, são selecionados de 70 a 89 atributos do total de 224).

Palavras-chaves: Mamografia. Seleção de Atributos. Câncer de Mama. Processamento de Imagem

Abstract

ROCHA, Arthur Diego Dias. **Detection and classification of breast lesions in mammography images using SVM classifiers, morphological wavelets and feature selection**. 2016. 108 p.

Breast cancer is one of the most common type of cancer among women. According to Brazil's national institute of cancer, in 2013 it was registered 14,388 deaths due to this disease. Breast cancer is not only a national but worldwide concern. The most used method to its detection is mammography which is an image technique that uses X ray emission and measures the non-absorbed radiation by the breast internal tissues. Mammography is a hard to analyze image exam, mainly because in many cases tumor's density is much alike some of the healthy tissues' density. An interesting approach is the use of computer-aided techniques for diagnosis, meaning the use of image processing and computational intelligence tools designed to support and aid radiologists in their tasks. Previous studies show that considering the different types of breast tissue dominance improves considerably the rate of correct classification by these computational tools. It is proposed for this work the development of a breast tumor classification system based on *Zernike* descriptors as shape descriptors of these breast lesions along with support vector machines as machine learning algorithms for classification. Some feature selection techniques are compared for reducing the whole system computational cost but always taking into consideration that the classification rates must be kept as high as possible. Of the techniques studied in this work, principal components analysis along with morphological wavelet transform for image preprocessing has shown itself as a great technique for feature reduction with lesser impact on classification rates. It was achieved a mean 2% loss in those rates (from about 95% to 93% as mean values) with a mean feature reduction of about 64% (in the range of 70 to 89 features from 224).

Keywords: Mammography. Feature Selection. Breast Cancer. Image Processing

Lista de figuras

Lista de algoritmos

Lista de tabelas

Lista de abreviaturas e siglas

AG	Algoritmo Genético
BE	Busca Evolucionária
BF	*Best First*
CAD	*Computer-Aided Diagnosis* - Diagnóstico Auxiliado por Computador
CFS	*Correlation-based Feature Selection* - Seleção de Atributos Baseada em Correlação
DNA	*Desoxiribonucleic Acid* - Ácido Desoxirribonucleico
GI	Ganho de Informação
MLP	*MultiLayer Perceptron* - Perceptron Multicamadas
PCA	*Principal Component Analisys* - Análise de Componentes Principais
PSO	*Particle Swarm Optimization* - Otimização por Enxame de Partículas
SVM	*Support Vector Machine* - Máquina de Vetor de Suporte
WEKA	*Waikato Environment for Knowledge Analysis*

Lista de símbolos

μ	Média
σ	Desvio Padrão
$\mathcal{H}_0$	Hipótese Nula
ρ	Raio da circunferência
ϕ	Ângulo azimutal

Sumário

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

O uso de imagens médicas para construção e solidificação do diagnóstico de doenças tem crescido consideravelmente na prática clínica da atualidade. É importante notar que mesmo com todo o avanço tecnológico e científico nos últimos tempos, ainda existem áreas de atenção básica à saúde, diagnóstico e tratamento com grandes deficiências, no que diz respeito à eficácia e velocidade.

Mais especificamente na área de diagnóstico por imagens mamográficas, esses procedimentos se dão de maneira, por muitas vezes, bastante dependentes de expertise individual de profissionais, carecendo de um padrão de qualidade menos subjetivo e menos sujeito a falhas devido aos fatores humanos quando da inspeção visual dos mamogramas, principalmente quando submetido à análise de grandes conjuntos de mamogramas, o que pode levar à fadiga do profissional. (GANESAN et al., 2013)

Um exemplo de exame para diagnóstico por imagem é a mamografia. Este exame, particularmente, utiliza ondas de Raios X para a formação de sua imagem; e desde sua descoberta em 1895 pelo alemão Roëntgen, este tipo de radiação tem sido utilizado na formação de diversos tipos de imagem, principalmente na radiografia convencional, que vem sendo utilizado até a atualidade. É amplamente conhecido o risco de exposição de tecidos vivos a este tipo de radiação, que pertence ao grupo das radiações ionizantes, que são aquelas que por sua capacidade de ionizar a matéria (principalmente as moléculas de água) podem provocar, por efeitos secundários, a desestabilização da cadeia de DNA e com isso, desencadear mutações e provocar até doenças como o próprio câncer.

Existem outros tipos de exames por imagem para detecção de tumores de mama, nomeadamente a ressonância magnética, o ultrassom, a tomografia por impedância elétrica e a termografia de mama são alguns exemplos. Cada uma dessas técnicas se baseia em princípios físicos diferentes e sem a utilização de radiações ionizantes, porém muitos deles estão se desenvolvendo em ambientes de pesquisa acadêmica e não estão solidificados na prática clínica, outros são utilizados na prática, mas não possuem o poder e principalmente não são difundidos como a própria mamografia que é a técnica mais recorrida na atualidade com uma maior "massa crítica" e *know-how* por parte dos profissionais radiologistas.

Em comparação com a radiografia tradicional, a mamografia utiliza os Raios X não para diferenciar a estrutura óssea em contraste aos chamados tecidos moles, mas sim para detectar lesões de mama, como por exemplo as microcalcificações, nódulos mamários e distorções arquiteturais (SAMPAT; MARKEY; BOVIK, 2003), onde câncer de mama é qualquer forma de tumor maligno que se desenvolve em algum tecido da mama (AKAY, 2009).

De acordo com o Instituto Nacional do Câncer (INCA), no Brasil, a mamografia é o método preconizado para o rastreamento na rotina da atenção integral à saúde da mulher, onde a mamografia de rotina é recomendada para mulheres de 50 a 69 anos de idade a cada 2 anos (INCA, 2015a), mesmo embora a biópsia seja a única forma de diagnosticar com total certeza esta moléstia, a mamografia permite a identificação de tumores de mama antes do paciente ou médico serem capazes de senti-los (auto-exame e palpação clínica da mama). (MAGGIO, 2004; HELA et al., 2013; GROMET, 2008; GILBERT et al., 2008)

1.1 Motivação e Justificativa

Segundo o Instituto Nacional do Câncer (INCA), o câncer de mama é o mais comum entre as mulheres no mundo e no Brasil, depois do de pele não melanoma, respondendo por cerca de 25% dos casos novos a cada ano. O mesmo também acomete homens, porém é raro, representando apenas 1% do total de casos da doença. Relativamente raro antes dos 35 anos, acima desta idade sua incidência cresce progressivamente, especialmente após os 50 anos. Estatísticas indicam aumento da sua incidência tanto nos países desenvolvidos quanto nos em desenvolvimento. Em 2015 são estimados 57.120 novos casos no Brasil e em 2013 foram registradas 14.388 mortes, sendo 14.207 mulheres. (INCA, 2015b) Inserido neste cenário, ainda agrega-se a condição de que quando descoberto em estágios iniciais, o prognóstico torna-se relativamente favorável e promissor.

Talvez, o maior problema associado à mamografia seja a dificuldade na sua interpretação por parte do médico radiologista e isto se dá majoritariamente pelo diminuto contraste na imagem entre as massas tumorais e os tecidos saudáveis que efetivamente constituem a mama, principalmente os de composição predominantemente glandular, pelo fato de essas estruturas possuírem uma grande proximidade em suas densidades radiológicas, e semelhantemente, coeficientes de atenuação da radiação muito próximas

(GANESAN et al., 2013; HELA et al., 2013). Condicionado a essa subjetividade na análise dessas imagens por parte dos radiologistas, tem-se um crescente índice de falsos positivos e falsos negativos. Uma das formas de se tentar minimizar esses falsos positivos ou negativos é utilizar um protocolo de dupla checagem das imagens por diferentes profissionais, onde a vantagem desta abordagem seria uma diminuição do fator de subjetividade introduzido pela análise por parte de um único profissional, porém o viés trazido por este método é a dificuldade em alocar mais do que um profissional para esta análise e com isso, todas as imposição de custo e tempo para a realização do diagnóstico. (BLANKS; WALLIS; MOSS, 1998)

Nesse contexto, uma alternativa bastante interessante é o uso de sistemas inteligentes como uma segunda avaliação no apoio ao diagnóstico. Esses sistemas são ferramentas computacionais capazes de "inferir", através de exemplos anteriores, informações sobre casos de teste, isto é, através da construção de "conhecimento" a partir de uma etapa de treinamento com exemplos já classificados de um dado conjunto, essas máquinas possuem a capacidade de extrapolação, ou seja, conseguem classificar um novo exemplo nunca analisado antes, tendo como restrição que essa amostra pertença ao grupo populacional do qual também pertence o conjunto de exemplos utilizados no treinamento (o referido conjunto, por sua vez, deve ser estatisticamente representativo da população como um todo).

Em (FERNANDES, 2015) foi realizada uma análise comparativa entre alguns sistemas CAD (*Computer-Aided Diagnosis*) para o apoio ao diagnóstico médico do câncer de mama levando em consideração questões relacionadas às diferenças de densidade tecidual dos diversos tipos de composições das mamas. A partir das abordagens e contribuições realizadas no trabalho mencionado, são propostos para este, a continuação e refinamento dos resultados obtidos através da comparação de duas abordagens de pré-processamento dessas mamografias, bem como do uso de apenas do descritor de forma, um único tipo de classificador e principalmente o uso de técnicas para redução de atributos, tendo em vista que conforme a quantidade de entradas (o vetor de atributos) do sistema de classificação (neste caso as máquinas de vetores de suporte) cresce, também o faz de forma significativa o tempo de treinamento, tendo em vista que a arquitetura da própria rede é dependente da quantidade de entradas.

1.2 Objetivos

Um sistema inteligente pode ser uma ferramenta determinante para melhora da capacidade de realização de diagnóstico por um profissional médico. Esses tipos de sistema podem funcionar como um apoio ao especialista, de forma a destacar detalhes que, caso contrário, poderiam estar passando despercebidos. No entanto, o desenvolvimento de um sistema inteligente de apoio ao diagnóstico clínico, como o de câncer de mama, não é uma tarefa simples e, portanto, deve ser meticulosamente validado. Um dos problemas desses sistemas é o *trade-off* entre custo computacional (que poderia inviabilizar a utilidade prática de todo o sistema) e o desempenho de classificação (que, quando baixo, inviabilizaria todo o sistema, tendo em vista que poderia levar a perda de vidas humanas).

1.2.1 Objetivo Geral

Para este trabalho é proposta a realização de uma análise comparativa entre técnicas de redução e seleção de atributos para a otimização do custo computacional de treinamento de máquinas de aprendizado utilizadas para a classificação de lesões de mama a partir de imagens de mamografia reais da base IRMA, partindo do pressuposto que o impacto de redução do desempenho de classificação dessas lesões seja o mínimo possível para não inviabilizar a utilidade prática do sistema proposto.

Também é parte do escopo deste trabalho a comparação de duas diferentes abordagens para a realização do pré-processamento das imagens mamográficas com o objetivo de realçar características da mesma para que fique evidente na geração do vetor de atributos e posterior classificação pela máquina de aprendizado.

1.2.2 Objetivos Específicos

Pretende-se com este trabalho comparar especificamente duas abordagens de pré-processamento da imagem mamográfica, através do uso da transformada de wavelets utilizando na primeira abordagem bancos de filtros lineares passa-altas e passa-baixas para os níveis de detalhes e aproximações respectivamente e na segunda abordagem, o uso de filtros baseados em morfologia matemática. Também é do escopo deste trabalho, a

comparação de algumas técnicas de redução de atributos, com o objetivo de diminuição da dimensionalidade dos dados do vetor de entrada para o treinamento do classificador. Outro objetivo deste trabalho é a realização de testes estatísticos para se averiguar a eficácia dos métodos minimizando os efeitos negativos na capacidade de extrapolação da rede.

Para o adequado cumprimento das metas propostas, é necessária a realização de alguns objetivos específicos, como a revisão de técnicas de redução de atributos, bem como a revisão da fundamentação teórica básica para emprego das técnicas propostas neste trabalho e dentro desta fundamentação teórica, observa-se a necessidade de estudos sobre toda a parte computacional e técnicas empregadas na construção do sistema inteligente de apoio ao diagnóstico, bem como a revisão de conceitos biológicos atrelados ao problema estudado. Demais objetivos específicos para realização deste trabalho concernem a familiarização com a biblioteca de métodos do software WEKA, bem como a aplicação do software convIRMA para utilização no sistema de classificação utilizado neste estudo. Posteriormente, o sistema deve ser construído e testado, para posteriormente os dados de classificação com e sem redução de atributos sejam coletados para a realização de análise estatística e posterior discussão acerca dos resultados obtidos com a aplicação das técnicas propostas.

1.3 Metodologia

Este trabalho visa o desenvolvimento de software e algoritmos aplicados a problemas da área de saúde, para isto utiliza software e bibliotecas já existentes e valida suas proposições através da simulação (ou experimentos computacionais). A parte de classificação e redução de atributos foi realizada utilizando a ferramenta WEKA e *add-ons* do mesmo software importados para aplicações escritas em java e em haskell. As etapas de pré-processamento e extração de atributos da imagem foram realizadas utilizando o software ConvIRMA. Os testes estatísticos foram realizados no software R. As etapas de seleção de atributos e classificação (por sua natureza estocástica) foram realizadas com repetições de trinta vezes para cada caso e a partir disto, poder se obter respaldo estatístico nos testes a serem realizados (teste de Wilcoxon).

1.4 Contribuições Esperadas

Ao término deste projeto, são esperadas contribuições acerca do levantamento da técnica (ou técnicas) mais relevante para ser aplicada ao problema de seleção e redução de atributos para o sistema inteligente de apoio ao diagnóstico de lesões de mama aqui proposto.

Este trabalho tem a pretensão de também selecionar a melhor técnica (dentre duas opções) de pré-processamento das imagens de mamografia.

1.5 Organização do Documento

Esta dissertação está organizada da seguinte maneira: Este primeiro capítulo reserva-se a realizar uma introdução geral ao trabalho, expondo a motivação e justificativa de realização do mesmo, dentre outros. O segundo capítulo trata de trabalhos relacionados, onde são relatadas técnicas de redução de atributos e aplicações em classificação de câncer de mama.

Para o terceiro capítulo são abordados os temas referentes à revisão dos fundamentos teóricos pertinentes relacionados ao contexto biológico da anatomia e fisiologia da mama, bem como sobre o exame de mamografia, a base de imagens e outros tópicos relacionados diretamente com a mamografia e o câncer de mama.

Para o quarto capítulo são revisados os conceitos relacionados às técnicas computacionais utilizadas neste trabalho, dentre elas: Redes neurais artificiais e máquinas de aprendizado supervisionado, bem como as técnicas utilizadas para redução de atributos, inclusive os algoritmos de busca e otimização utilizados para esta tarefa e outros.

O quinto capítulo é reservado para o aprofundamento da descrição da metodologia utilizada na realização dos experimentos computacionais (simulações), sempre visando a repetibilidade do experimento científico (sendo este um dos principais pilares da pesquisa científica, a repetibilidade do experimento).

No sexto capítulo são demonstrados e analisados os resultados obtidos nos experimentos realizados. Por fim, no sétimo capítulo, são expostas as considerações finais, bem como conclusões, dificuldades encontradas e sugestões de melhorias e trabalhos futuros.

2 Trabalhos Relacionados

Alguns esforços têm sido realizados por parte da comunidade acadêmica em prol da exploração de diferentes técnicas para redução de atributos em sistemas de classificação automática de lesões de mama.

No trabalho realizado por (THANGAVEL; VELAYUTHAM, 2012) foi proposta uma técnica de seleção de atributos não supervisionada baseada na medida de entropia de *Rough Sets*. A técnica foi comparada com algumas técnicas supervisionadas e avaliadas com uma métrica baseada em *Fuzzy C-Means Clustering*. Foi mostrado que a técnica é capaz de remover efetivamente atributos redundantes e que os subconjuntos de atributos retornados por este método não supervisionado eram similares aos das técnicas supervisionadas. Neste trabalho foram utilizados descritores de *Haralick* que são descritores de textura baseados na geração de momentos a partir da matriz de co-ocorrência de níveis de cinza. A base de imagens utilizada foi a base MIAS. As taxas de acerto encontradas foram em na faixa de 69% a 74,5%.

(CHAKRAVARTY et al., 2013) prupuseram a criação de uma função objetivo baseada na técnica de Evolução Diferencial para seleção de atributos que combina a variação intraclasse e a distância interclasse utilizando multiplicadores de Lagrange e medidas estatísticas como métricas. O classificador utilizado foram as máquinas de vetor de suporte e o sistema foi aplicado a bases de dados de diversas origens e, dentre elas, a base *Wisconsin Breast Cancer*. Para esta base, foram alcançadas taxas de acerto no valor de 97,8%.

A base WDBC, diferentemente das bases MIAS, DDSM ou IRMA, não é uma base de imagens mamográficas. A WDBC contem 699 instâncias de lesões de mama descritas por 10 atributos referentes à análise citológica de células da região suspeita obtidas por biópsia de aspiração por agulha fina.

Em (OSAREH; SHADGAR, 2009) também foi utilizada a base WDBC e foram combinadas diversas técnicas de aprendizado de máquina (como as máquinas de vetor de suporte, *K-nearest neighbours* e redes neurais probabilísticas) com técnicas como avaliação de razão sinal-ruído, seleção de atributos baseada em seleção sequencial direta e a análise de componentes principais (PCA). As melhores taxas de acerto alcançadas foram

98,8% e 96,33% utilizando máquinas de vetor de suporte com kernel RBF (*Radial-Basis Function*).

Outro trabalho a utilizar a base de dados WDBC foi o realizado por (HASAN; TAHIR, 2010), onde foram utilizadas técnicas baseadas em análise de componentes principais (PCA) combinadas com redes neurais artificiais como classificadores e para a referida base de dados, foram obtidas taxas de acerto de até 97,98%.

Em (KEYVANFARD; SHOOREHDELI; TESHNEHLAB, 2011) foi realizado um trabalho de classificação automática de lesões de mama utilizando imagens de ressonância magnética. Para isto, a técnica de *Fuzzy C-Means* é utilizada para determinar e realçar as bordas do tumor, de onde são extraídos atributos de forma e de textura da dada região de interesse. Então, os algoritmos genéticos são utilizados como técnica de busca para selecionar o melhor subconjunto de atributos para os diferentes classificadores utilizados (algumas redes neurais artificiais e as máquinas de vetor de suporte). As imagens utilizadas forma obtidas no Hospital Milad no Tehran, Iran. Taxas de acerto alcançadas oscilaram entre 87% e 97%.

O trabalho realizado por (PéREZ et al., 2014) utiliza as bases de imagens de mamografias *Breast Cancer Digital Repository* (BCDR) e *Digital Database for Screening Mammography* (DDSM) que foi obtida a partir da base *Digital Database for Screening Mammography* (IRMA). As técnicas de seleção de atributos utilizadas foram baseadas em discretização χ^2, ganho de informação, *one-rule* e *relief*. Também foi utilizada uma outra técnica proposta pelos autores em trabalhos anteriores chamada *RMean* que é baseada em uma função de ponderação para indexação de atributos relevantes. Os classificadores utilizados foram as redes *Multi-Layer Perceptron* (MLP), máquinas de vetor de suporte, classificador *Naive-Bayes* e classificador baseado em análise discriminante linear (LDA). A combinação com obtenção de melhores resultados é a baseada em MLP e RMean, para o qual foram obtidas áreas abaixo da curva ROC (AUC) na faixa de 0,78 a 0,8562.

O próximo trabalho analisado se utilizou de atributos curvilineares, de textura, de *Gabor*, bem como atributos multi-resolução. Sendo combinados com as técnicas de busca por algoritmos genéticos e *adaptive floating search* e classificação com LDA. A base mamográfica utilizada foi a DDSM. Foi chegada à conclusão de que a técnica de

adaptive floating search (que naturalmente é mais desenvolta que os algoritmos genéticos para problemas de menor escala como o proposto no trabalho em questão) obteve melhor performance com o valor de área abaixo da curva ROC de 0,93 (SUN; BABBS; DELP, 2005).

Um trabalho de seleção de atributos aplicado a detecção de câncer de mama através de imagens de ultrassonografia é o proposto por (NAYEEM et al., 2014). Neste trabalho foram utilizados mais de 50 atributos e para seleção destes foi utilizada a técnica de *Multi-Cluster Feature Selection* (MCFS). Foi utilizado um classificador de representação esparsa (SRC). Sendo, as imagens, obtidas das universidades de Thomas Jeferson, Cincinnati e Yale, sendo os diagnósticos confirmados por biópsia. Como resultado, foram obtidas taxas de acerto de classificação de 87,52% a 93,31%.

O trabalho realizado por (MUñOZ-MEZA; GóMEZ, 2013) é também de redução em atributos gerados em imagens de ultrassonografia de mama. Para este trabalho, foram utilizadas imagens oriundas da base *Breast Ultrassound* (BUS). Foram utilizadas técnicas baseadas em análise de componentes principais (PCA) e informação mútua (MI) para seleção de atributos. Como classificador, foi utilizada a análise discriminante de Fisher (FLDA). A melhor performance de classificação obtida por este trabalho foi no emprego da técnica de informação mútua (área abaixo da curva ROC de 0.951 e 13 atributos), enquanto que para o conjunto total de atributos, a área abaixo da curva ROC foi de 0,657 com uma quantidade de 524 atributos.

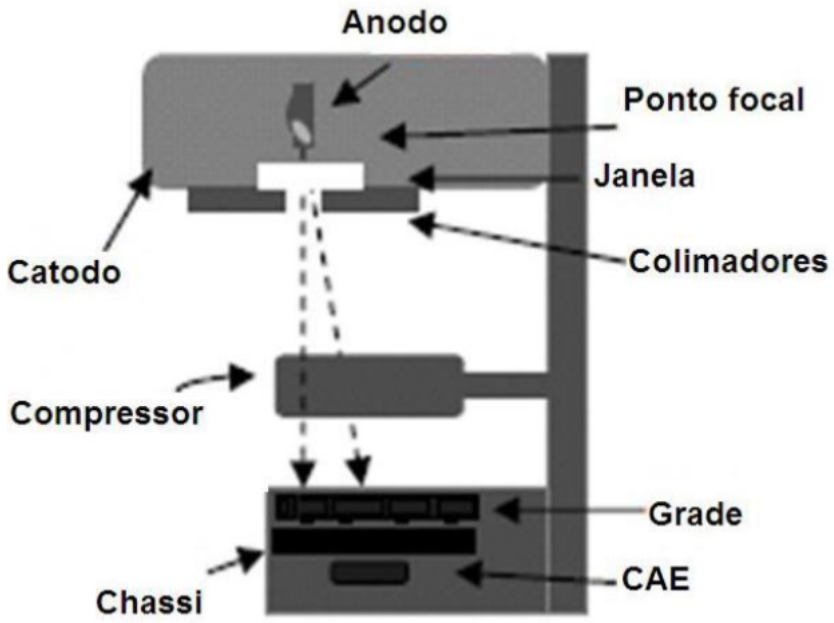
3 Imagens Mamográficas

Presente predominantemente, mas não exclusivamente, em mulheres, o câncer de mama se apresenta como um vilão para o bem-estar e qualidade de vida, quando não fatal. Normalmente indolores em estágios não avançados, os tumores de mama acabam passando despercebidos ou ignorados, trazendo complicações futuras que poderiam ser evitadas. O autoexame da mama é o método mais simples e de grande importância no processo de identificação do tumor. Porém, cistos ou tumores de dimensões inferiores (normalmente em estágios iniciais) e principalmente microcalcificações são de difícil detecção por este meio. Existem diversas técnicas para diagnóstico por imagem de câncer de mama, entre eles temos ultrassom e termografia. Contudo, a mamografia permanece sendo a técnica mais utilizada para a detecção não invasiva do câncer de mama. Quando há a suspeita não confirmada de câncer, técnicas invasivas, como a biópsia, são utilizadas como meio de confirmar, ou não, a patologia.

A mamografia, ou exame mamográfico, é uma técnica de produção de imagem da estrutura interna da mama. Nela, radiação é emitida pelo mamógrafo a partir do anodo, guiado pela direção do catodo e sendo focalizada pelos colimadores, vide figura 1. Ao incidir sobre o seio, que deve estar depositado entre o chassi e o compressor que uniformiza a distribuição da estrutura mamária sobre o aparelho, uma parcela dessa energia é transmitida e é captada em um receptor que é responsável por converter a informação de intensidade de radiação transmitida em imagem. A imagem pode ser analógica (filmes radiológicos) ou digital, onde, nesse processo de aquisição de imagem, existe um sistema de captação que converte o sinal em uma grandeza elétrica e consequentemente converte-o para o domínio digital, podendo ser processado por softwares de computador.

A importância da mamografia se dá pelo fato de ser uma técnica não invasiva com a capacidade de detectar estruturas de pequenas dimensões (cistos, massas, microcalcificações) que não poderiam ser identificadas por outra técnica não invasiva. Todavia, a experiência do profissional que analisa a imagem com o objetivo de diagnosticar a patologia (radiologista) influencia bastante no resultado e é um atributo de quantificação quase impossível. Por esse motivo, sistemas de apoio ao diagnóstico vêm sendo propostos com o intuito de auxiliar o

Figura 1 – Esquema do Mamógrafo



profissional e aprimorar cada vez mais o processo de diagnóstico da patologia, visando sempre o bem-estar do paciente.

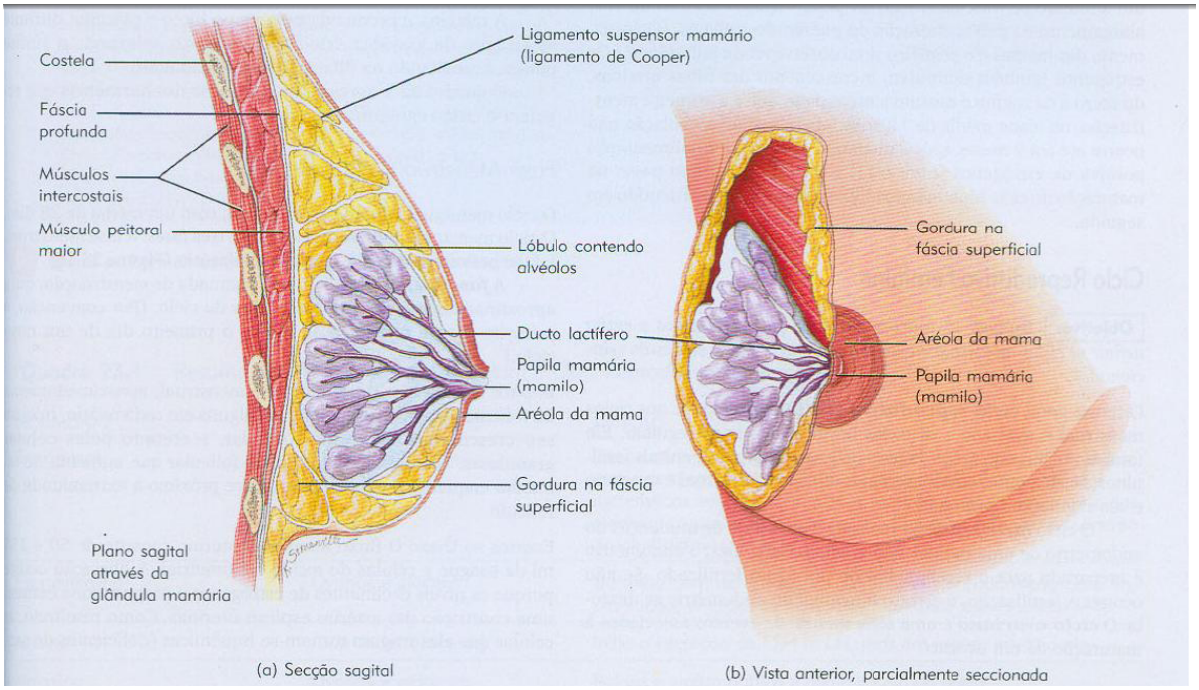
Computadores digitais representam uma ferramenta fundamental nesse processo de auxílio ao diagnóstico. Utilizando-se imagens mamográficas digitalizadas, podem-se construir softwares que combinam técnicas de processamento digital de imagens e inteligência computacional como ferramentas para a promoção de um diagnóstico mais preciso, eficiente e rápido por parte do profissional médico.

3.1 Anatomia e Fisiologia da Mama

A mama, ou seio, é uma estrutura presente nos mamíferos, sua principal função é a lactação, ou amamentação dos recém-nascidos. É uma estrutura também presente nos indivíduos masculinos, porém apresentando hipotrofia, tamanho reduzido e sem função aparente. A estrutura da mama é constituída por tecido adiposo, tecido conjuntivo e tecido glandular mamário.

As mamas começam a desenvolver-se na puberdade. Esse desenvolvimento é estimulado pelos estrogênios do ciclo sexual feminino mensal; os estrogênios estimulam o crescimento da parte glandular das mamas, além do depósito de gordura que concederá

Figura 2 – Estrutura Mamária



Fonte: (TORTORA, 2007)

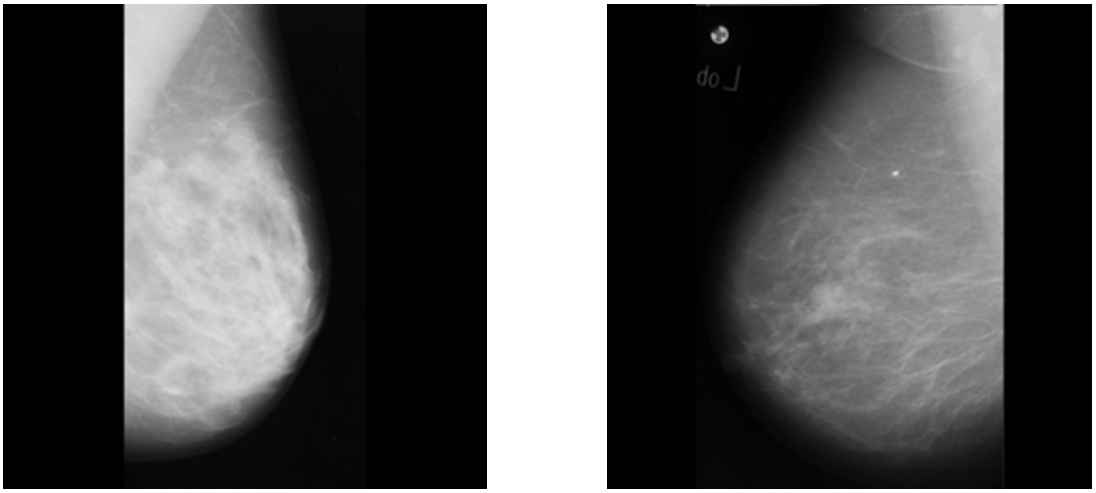
massa às mamas. Além disso, ocorre um crescimento bem mais intenso durante o estado de altos níveis de estrogênio da gravidez e só então o tecido glandular torna-se inteiramente desenvolvido para a produção de leite (GUYTON; HALL, 2006).

As glândulas mamárias são glândulas sudoríparas modificadas que se situam sobre os músculos peitoral maior e serrátil anterior e são ligadas a eles por tecido conjuntivo (TORTORA, 2007). Conforme figura 2 pode-se observar as estruturas de composição da mama, com destaque para o tecido adiposo presente em predominância em localização mais periférica (e que varia conforme a idade e tipo físico da mulher), mais centralizadas estão as glândulas mamárias e na porção anterior, estão os tecidos músculo-esqueléticos.

Internamente, cada glândula mamária consiste de 15 a 20 lobos organizados radialmente e separados por tecido adiposo e faixas de tecido conjuntivo (ligamentos de Cooper) que suportam as mamas (TORTORA, 2007).

Lactação é o nome dado à principal função das glândulas da mama que compreende os atos de secreção e ejeção do leite materno. No período anterior à puberdade, ambos os homens e mulheres possuem glândulas mamárias subdesenvolvidas aparentando leves

Figura 3 – Imagem comparativa entre mamogramas de acordo com idade



(a) Mama jovem (b) Mama idosa

Fonte: Base de mamografias MiniMias (SUCKLING et al., 1994)

elevações na região torácica. Com o surgimento da puberdade, as mamas femininas começam a desenvolver-se, sob a influência dos estrogênios.

As mamas são estruturas de composição dinâmica ao longo da vida, isto é, para mulheres mais jovens, as mamas são formadas pelos tecidos glandulares, adiposos e conjuntivos, porém, com o passar da idade, o tecido adiposo passa a ser predominante na estrutura. O tecido adiposo é menos denso que os demais, o que acaba tornando-o mais escuro na imagem. Em contrapartida, a imagem de um seio mais jovem tende a ser mais claro, vide figura 3, o que acaba dificultando fortemente a distinção desse tipo de estrutura saudável do seio com possíveis tumores e massas malignas (que também tendem a ser claros na imagem), esse é um dos motivos pelo qual não se aconselha a mamografia para mulheres jovens (salvo casos onde já se há a suspeita de presença do tumor e o exame é utilizado como confirmação).

3.2 Formação da Imagem Mamográfica

Para a formação da imagem das estruturas internas aos seios, primeiramente, é necessário que haja algum tipo de sinal (no sentido mais amplo possível) que seja capaz de atravessar e interagir com os tecidos internos dos seios, dessa forma, pode-se extrair informação a respeito das condições internas da estrutura.

Em 1895, Wilhelm Conrad Roentgen descobriu um tipo de radiação pertencente ao espectro eletromagnético altamente energética, essa radiação (que ficou conhecida como Raio-X) possui energia na ordem de keV (kilo elétron-volt) até poucas centenas de keV. Considerando-se esses atributos, os raios-x são chamadas também de radiações ionizantes pelo fato de terem o poder de ionização da matéria (principalmente a molécula de água que é predominante no corpo humano). Os raios-x possuem um enorme poder de penetração nos chamados tecidos moles (matéria orgânica de menor densidade), por esse motivo, os raios-x passaram a ser, e ainda são, bastante utilizados na produção de imagens de estruturas ósseas do corpo humano. Posteriormente, essa radiação passou a ser utilizada na produção de imagens de estruturas formadas predominantemente por tecidos moles (como as mamas, por exemplo), para isso, a dose radiológica e a energia do fóton emitida devem ser ajustados para obter penetração na matéria (tecido) ao mesmo tempo em que mantém certo nível de contraste entre as estruturas de tecido mole.

Devido ao fato dessa radiação ser ionizante, ela pode ser prejudicial ao corpo e provocar doenças como o próprio câncer, por exemplo. Por isso, exige-se uma maior eficiência no diagnóstico da doença, evitando a necessidade de maior exposição à radiação na condição de repetição do exame.

Para a formação da imagem, é necessário que sejam gerados raios-x. Essa radiação é direcionada à mama (que deve estar comprimida em um suporte), após interagir com a mama (parte da radiação é absorvida, parte é espalhada e outra parte atravessa), a radiação é captada em receptores que após um processo de aquisição do sinal (para mamógrafos digitais) se transforma em imagem.

É importante destacar que a quantidade de energia que atravessa o seio depende dos tecidos que o compõe (os coeficientes de transmitância radiológica são diferentes para os diferentes tecidos), por exemplo: o tecido adiposo atenua menos a radiação em comparação com microcalcificações ou a própria massa tumoral mais densa. Deve-se destacar que imagens são sinais bidimensionais e que trazem informação planificada das estruturas tridimensionais dos seios, o que é responsável pela sobreposição da “penumbra” gerada pela transmissão de radiação através das camadas de tecidos internos ao seio. Aliado a isso, imagens mamográficas são bastante ruidosas e diversos artefatos acabam prejudicando um

pouco o processo de diagnóstico, deste modo, é de fundamental importância que sistemas de apoio ao diagnóstico sejam utilizados.

3.3 Base de Imagens de Mamografias

Para a validação do sistema CAD, é de interesse que sejam utilizadas imagens reais de mamografias e que as mesmas já estejam previamente classificadas de acordo com a expertise de profissionais radiologistas.

Todas as imagens mamográficas utilizadas neste trabalho foram obtidas da base de imagens IRMA[1] (Image Retrieval in Medical Applications) (OLIVEIRA et al., 2008; DESERNO et al., 2012).

A base IRMA, criada pela Universidade de Tecnologia de Aachen (RWTH Aachen), é a unificação de outras bases de imagens mamográficas disponíveis publicamente para pesquisa, como DDSM (Digital Database for Screening Mammography), MIAS (Mammographic Image Analysis Society), LLNL (Lawrence Livermore National Laboratory) e RWTH (Rheinisch-Westf alische Technische Hochschule, Aachen University, Aachen, Germany Department of Radiology).

As imagens que compõem esta base são *patches* redimensionados para terem a resolução de 128x128 *pixels* que se referem às regiões de interesse, isto é, neles estão contidas as lesões (ou ausência das mesmas, no caso de regiões de tecido normal), onde estas estão classificadas em benignas ou malignas. Além do diagnóstico, também estão presentes informações acerca da densidade do tecido que compõe a mama e o tipo de lesão (massa circunscrita, espiculada, microcalcificações, dentre outros).

[1] Cortesia de TM Deserno, Dept. de Informática Médica, RWTH Aachen, Alemanha

4 Técnicas Computacionais

Este capítulo destina-se a revisar parte do arcabouço teórico necessário para a compreensão das técnicas computacionais empregadas neste trabalho, desde as etapas iniciais de como representar imagens digitais, passando por técnicas de processamento das mesmas (filtragens e transformada de Wavelets), extração de atributos utilizando momentos de Zernike, Máquinas de Vetor de Suporte como classificadores e as técnicas aqui empregadas para a seleção dos atributos mais relevantes.

4.1 Imagens Digitais

Uma imagem pode ser definida como uma função bidimensional, $f(x, y)$, onde x e y são coordenadas espaciais do plano e a amplitude de f para qualquer par de coordenadas (x, y) é chamado de intensidade ou nível de cinza da imagem naquele ponto. Quando x, y e f são todas quantidades discretas e finitas, chama-se a imagem de digital (GONZALEZ; WOODS, 2002). Uma enorme vantagem do uso de imagens digitais sobre as analógicas (como por exemplo as mamografias analógicas em filme) é a possibilidade do uso de computadores digitais para o processamento das mesmas. A computação digital tem evoluído de forma considerável tanto em poder de processamento, quanto no desenvolvimento de técnicas e algoritmos para manipulação dessas imagens.

Para um dada imagem, *pixel* (*picture element*) é a unidade espacial básica que é representada pelo par ordenado (x, y) e que possui nível de intensidade $f(x, y)$. Resolução é o nome dado às dimensões máximas de uma dada imagem na vertical e na horizontal. Para imagens digitais de oito bits tem-se 256 (podendo ser escalado no intervalo de [0,255] ou [0,1], ambos os casos *unsigned*) valores possíveis de intensidade que representam o nível de cinza do dado pixel.

Filtragem é um dos exemplos dos tipos de processamento de imagens. Supondo uma imagem $f(x, y)$ e uma máscara $h(x, y)$, o resultado $g(x, y)$ derivado da convolução de f por h, isto é, $g(x, y) = (f * h)(x, y)$, onde o operador $*$ denota a operação de convolução, é a imagem filtrada. Exemplos comuns de máscaras são a da média, $h_m = \frac{1}{9} \times$

$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$, muito utilizada para gerar imagens de aproximações, ou atenuar ruídos e possui a característica de "borrar" uma imagem, funcionando como uma filtro passa-baixas. Máscaras de detalhes também são bastante utilizadas, como por exemplo a de detalhes verticais (h_v) ou horizontais (h_h), que são exemplos de filtros passa altas, $h_v = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ $h_h = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$

Outra técnica de processamento de interesse para este trabalho é o tratamento por morfologia matemática. Tem-se como operações básicas primitivas a erosão e a dilatação.

A área de morfologia matemática é fundamentada na teoria dos conjuntos (para representação de objetos na imagem) e trata de algoritmos de processamento de imagens ligados à forma dos objetos nelas presentes (GONZALEZ; WOODS, 2002).

Para compreender as operações primitivas da morfologia matemática, primeiramente é necessário entender o que são os elementos estruturantes. Elementos estruturantes são pequenos conjuntos ou sub-imagens utilizados para avaliar uma imagem com relação a alguma propriedade de interesse, de maneira mais prática, um elemento estruturante funciona como uma máscara que percorre a imagem pixel a pixel computando a propriedade procurada para cada um desses pixels em toda a imagem. Normalmente esses elementos são simétricos e com origem em seu ponto central (ou centro de massa) e mesmo em tratamentos de imagens em níveis de cinza, esses elementos costumam ser homogêneos (assumindo um único valor em todos os pontos de interesse). Exemplos comuns de elementos estruturantes são vizinhança 4-conectada, vizinhança 8-conectada ou quadrados e círculos de variadas dimensões (GONZALEZ; WOODS, 2002).

A operação de erosão em imagens binárias corresponde (de uma maneira simplificada) em "varrer" toda a imagem original (para facilitar, chamada imagem A) e computando se para cada um desses pixels, todos os seus vizinhos possuem nível lógico alto quando o mesmo vizinho na posição relativa do elemento estruturante também possui nível alto (isto deve ser verdadeiro para todos os vizinhos), em outras palavras checar para toda a imagem A os pontos onde o elemento estruturante está contido nos objetos desta imagem em questão. No caso de imagens em níveis de cinza o processo é semelhante, com a

diferença de que o pixel em questão assumirá o valor do menor vizinho dentro da máscara do objeto estruturante observado na região da imagem A. A erosão em imagens possui o efeito de "consumir", deixar menores os objetos ou até mesmo remover completamente aqueles que não são grandes o suficiente para conter o elemento estruturante e a partir deste ponto de vista, esta operação pode ser enxergada como um filtro que remove elementos ou detalhes pequenos o suficiente (GONZALEZ; WOODS, 2002).

A operação de dilatação é o oposto da erosão. Dilatar uma imagem tem o efeito de alargar objetos, agregando mais pixels de nível alto para as vizinhanças onde antes era apenas *background*. Em imagens binárias, esta operação é realizada percorrendo a imagem e para cada pixel observa-se a vizinhança e quando pelo menos um pixel dentro da região do elemento estruturante for de nível alto na imagem original, então a imagem de saída sinalizará nesta posição um nível alto, caso contrário sinalizará nível baixo. Para imagens em níveis de cinza, a operação é semelhante mas com a diferença de retornar valor do pixel maior dentro da janela formada pelo elemento estruturante.

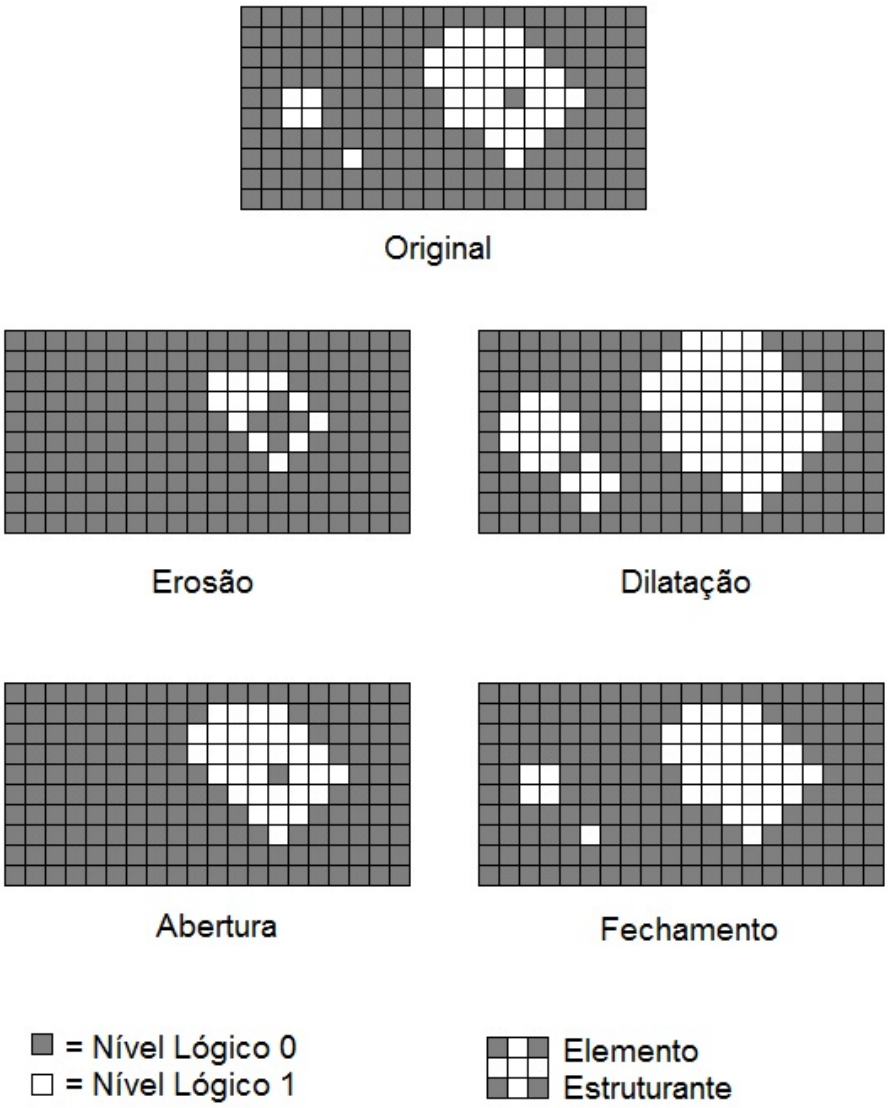
A partir das operações primitivas, são formuladas operações mais complexas, como por exemplo abertura, fechamento e gradiente morfológico, dentre muitas outras operações.

De forma bastante sucinta, abertura de ordem n é a operação correspondente a realizar n erosões seguidas de n dilatações, tendo o efeito de suavizar contornos, remove ligações estreitas e elimina pequenos objetos. Analogamente, fechamento de ordem n corresponde a n operações de dilatação seguidas de n erosões, tendo o efeito de também suavizar contornos, mas normalmente conecta extremidades desconexas suficientemente próximas e preenche espaços ou buracos. Gradiente morfológico, por sua vez, resulta da subtração da imagem original dilatada pela mesma original erodida, possui a característica de enfatizar bordas das regiões, áreas mais homogêneas tendem a não aparecerem no resultado (subtraem-se), essa capacidade de destacar contornos e bordas em detrimento de regiões mais homogêneas resulta em um efeito de gradiente.

Na figura 4 estão apresentadas as operações básicas de morfologia matemática para o exemplo de imagem original presente no topo da figura. O elemento estruturante utilizado foi a vizinhança 4-conectada. Percebe-se que a operação de erosão tem a propriedade de remover *pixels* localizados mais à borda dos objetos de maneira a eliminar totalmente

aqueles de diminutas dimensões. A operação de dilatação, opostamente, tem a propriedade de agregar mais *pixels* às bordas dos objetos existentes. As operações de abertura e fechamento têm a capacidade de (dependendo da ordem, do elemento estruturante e dimensões dos objetos, da mesma maneira que as outras operações) respectivamente eliminar pequenos objetos e pequenos buracos nas imagens originais.

Figura 4 – Operações Básicas de Morfologia Matemática: Erosão, Dilatação, Abertura e Fechamento



4.2 Transformada de Wavelets

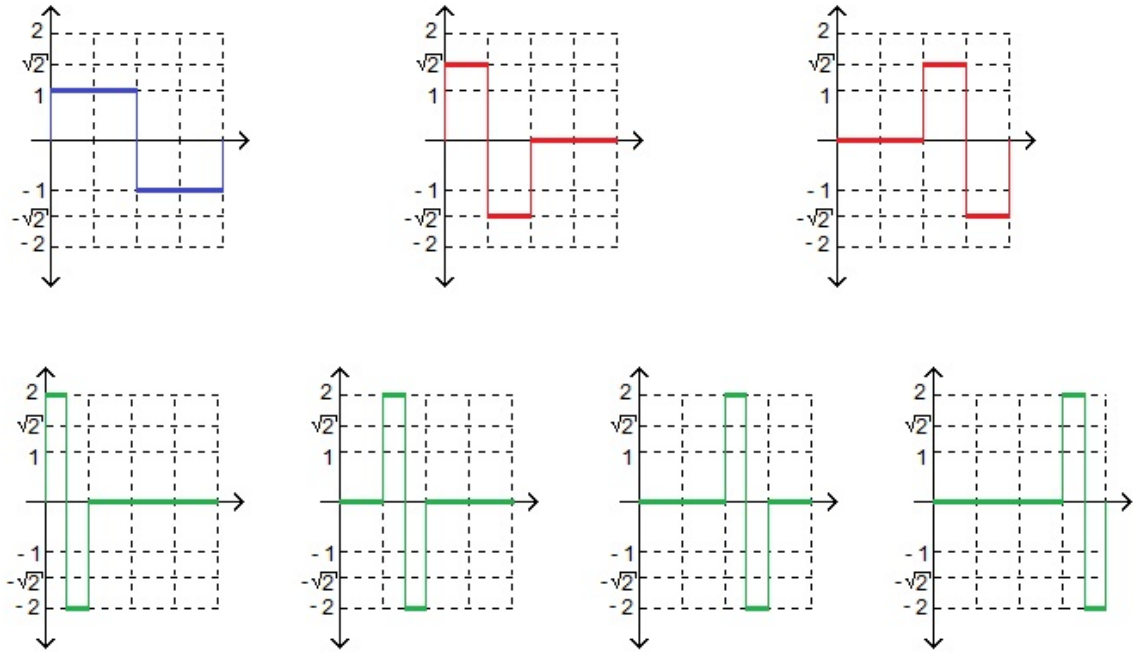
Em processamento de sinais, transformada de wavelets é um dos vários tipos de transformadas matemáticas que mapeiam um sinal de um domínio para outro. Um outro exemplo de transformada bastante utilizada na prática de processamento de sinais ou imagens é a transformada de Fourier (OPPENHEIM; SCHAFER, 1999). A grande diferença entre os dois tipos de transformada é que a de Fourier tem a capacidade de trazer um sinal da representação no domínio do tempo e passar a representá-lo no domínio da frequência, já a transformada de wavelets tem a enorme vantagem de levar o sinal para o domínio do tempo/frequência, ou seja, além de ser uma ferramenta capaz de analisar o conteúdo harmônico de um sinal, ela é poderosa o suficiente para apresentar essas informações ao longo do tempo.

Uma wavelet (ondaleta) é uma oscilação em formato de onda limitada no tempo, ou seja, sua amplitude inicia em zero e vai crescendo e decrescendo ao longo do tempo para no final decrescer para zero novamente. A partir de uma wavelet mãe podem ser geradas wavelets filhas através de deslocamento e compressão/dilatação no tempo. Essas ondas formam uma base ortogonal de forma semelhante às senoides que formam a base da transformada de Fourier, com a diferença de que essas senoides possuem apenas informação de frequência. O fato da onda ser limitada no tempo permite que de acordo com um fator de escala (comprimir ou dilatar no tempo) a resolução do componente de informação extraído por esta ondaleta possa se mover ao longo do tempo. Da mesma maneira, o conteúdo harmônico da ondaleta permite que a mesma seja capaz de extrair informação específica do conteúdo de frequência do sinal no ponto específico do tempo. Posto de outra maneira, cada wavelet filha é deslocada ao longo do sinal (através do processo de convolução) e extrai informações especificas de frequência em cada ponto no tempo do sinal, sendo a resolução mais grosseira para a banda do sinal de baixa frequência e mais refinada para a banda de alta frequência.

O tipo de wavelet mais simples é a wavelet de Haar (HAAR, 1910; CHUI, 1992) que representa uma família de wavelets de formato quadrado diferenciando em escala e posição ao longo do tempo, conforme figura 5, pode-se notar que essas funções são ortogonais entre sí (multiplicar e integrar no tempo par a par resulta em zero). A desvantagem desse

tipo de wavelet é o fato da mesma não ser contínua (logo não é diferenciável), mas essa propriedade pode ser vantajosa no sentido de destacar transições repentinas.

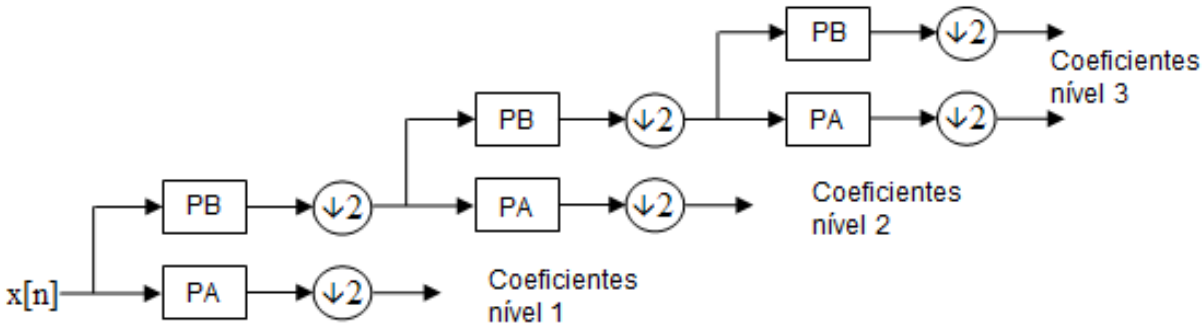
Figura 5 – Exemplo de algumas wavelets de Haar, a wavelet mãe no canto superior esquerdo e as filhas resultam de mudanças de escala e deslocamento no tempo



A maneira mais comum de se calcular a transformada discreta de wavelets é através da aplicação de bancos de filtros passas-altas (PA) e passas-baixas (PB), o diagrama de blocos está exposto na figura 6. O operador ($\downarrow 2$) representa o que se chama de *downsampling* e reduz a quantidade de elementos do sinal pela metade, embora essa operação à primeira vista pareça representar uma perda de informação, na verdade, com o uso de filtros ortogonais é possível ser realizada a reconstrução perfeita do sinal original junto com o operador inverso de *upsampling* ($\uparrow 2$). Quanto mais níveis de coeficientes forem gerados, significa que o sinal gerado será subdividido em mais bandas de frequências (cada saída do diagrama da figura 6 representa uma banda).

Este mesmo procedimento pode ser utilizado em imagens, tendo em vista que imagens são sinais bidimensionais cujo o domínio é o espaço e não o tempo. A diferença

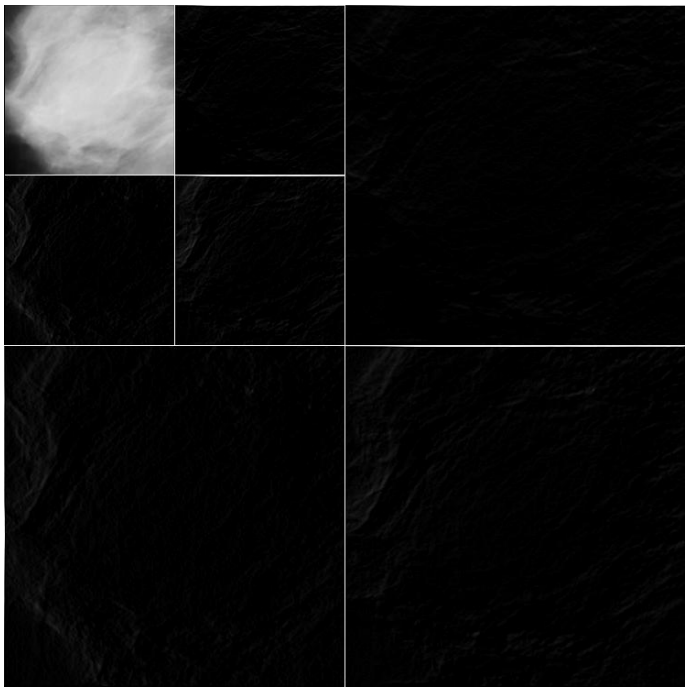
Figura 6 – Diagrama de blocos do banco de filtros para realizar a transformada de wavelets em três níveis



neste caso seriam os filtros passa-altas, que pelo fato das imagens serem bidimensionais então existem três filtros do tipo passa-altas (ou filtros de detalhes), referentes aos detalhes verticais, horizontais e diagonais. Um exemplo de transformada de wavelets de dois níveis em uma imagem pode ser observada na figura 7. Percebe-se que a imagem divide-se primeiramente em quatro quadrantes e o quadrante superior esquerdo, por sua vez, divide-se em mais quatro quadrantes. Cada subdivisão dessa representa um nível da transformada, onde o quadrante inferior esquerdo representa os coeficientes de detalhes verticais, o superior direito os detalhes horizontais e o inferior direito os diagonais. No quadrante superior esquerdo do último nível de decomposição encontram-se os coeficientes de aproximações de todo o processo. É importante perceber que a imagem resultante da transformada possui dimensões idênticas à imagem original, isso por causa do processo de *downsampling* que é reduz pela metade a quantidade de linhas e colunas. Caso não ocorresse esse processo de *downsampling*, cada imagem de coeficiente gerada em cada um dos níveis teria as mesmas dimensões da imagem original, isso significaria que para o exemplo da figura 7 seriam sete imagens diferentes com a mesma dimensão da original.

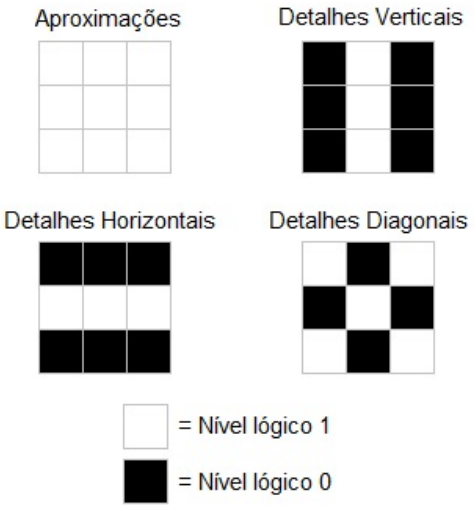
A transformada apresentada na figura 7 utiliza um banco de filtro digitais lineares para implementar a transformada mais simples (Haar), porém esses filtros lineares podem ser substituídos por operadores morfológicos e, nesse caso, a transformada de wavelets passaria a ser uma transformada chamada morfológica. Um conjunto de filtros morfológicos a ser utilizado para cumprir tal tarefa seria, por exemplo, uma sequência de fechamentos de ordem n seguidos de aberturas e fechamentos ambos dessa mesma ordem (para filtros passa-baixas também chamados de filtros de aproximações). No caso de filtros de detalhes

Figura 7 – Transformada de wavelets de dois niveis uma imagem de tumor de mama



(verticais, horizontais e diagonais), poderia ser utilizada a operação de gradiente morfológico de ordem n, em todos os casos diferenciando o elemento estruturante, conforme figura 8.

Figura 8 – Exemplos de elementos estruturantes utilizados em transformada de wavelets morfológica (em imagens de 8-bits, o nível lógico alto é representado pelo valor 255)



4.3 Momentos de Zernike

Os momentos de Zernike podem ser expressos como o resultado do mapeamento de uma imagem em um conjunto específico de polinômios complexos, chamados polinômios de Zernike, que formam uma base ortogonal dentro da esfera unitária. Devido ao fato desses polinômios serem ortogonais entre sí, os momentos gerados desta forma conseguem representar as propriedades da imagem sem redundância ou sobreposição de informação (TAHMASBI; SAKI; SHOKOUHI, 2010; NOLL, 1975).

O conjunto de polinômios pode ser expresso em sua forma complexa, utilizando para isto a função exponencial complexa como base, ou então separando a parte imaginária da parte real em sua forma par e ímpar com base em cossenos e senos. Esta segunda representação é conforme as equações (1a) e (1b), onde esta primeira representa os termos pares e a segunda os ímpares:

$$Z_n^m(\rho, \phi) = R_n^m(\rho)\ cos(m\phi) \tag{1a}$$

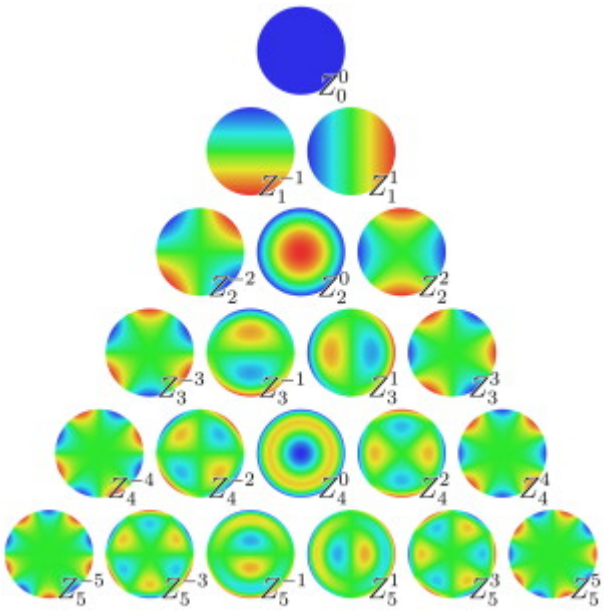
$$Z_n^{-m}(\rho, \phi) = R_n^m(\rho)\ sen(m\phi) \tag{1b}$$

Z_n^m representa o polinômio, os índices m e n são inteiros não negativos, onde $n \geq m$, ϕ representa o ângulo da projeção com relação ao plano de referência, ρ é a distância radial e deve estar entre o intervalo $0 \leq \rho \leq 1$ sendo definido para a imagens conforme a equação (2). O termo R_n^m é o polinômio radial e está definido conforme equação (3) para o caso de $n - m$ ser par e é igual a 0 caso contrário.

$$\rho(x, y) = \sqrt{\frac{(2x - X_{max})^2 + (2y - Y_{max})^2}{(2x - X_{max})(2y - Y_{max})}} \tag{2}$$

Onde $0 \leq x < X_{max}$, sendo X_{max} o valor da resolução da imagem na dimensão x. De forma semelhante $0 \leq y < Y_{max}$, sendo Y_{max} o valor da resolução da imagem na

Figura 9 – Magnitude de polinômios de Zernike de baixa ordem como função do raio e ângulo azimutal dentro do disco unitário



Fonte: (TAHMASBI; SAKI; SHOKOUHI, 2011)

dimensão y. Posições com valores de ρ maiores do que 1 são ignorados na realização do produto interno com os polinômios de Zernike.

$$R_n^m(\rho) = \sum_{k=0}^{\frac{n-m}{2}} \frac{(-1)^k (n-k)!}{k!(\frac{n+m}{2}-k)!(\frac{n-m}{2}-k)!} \rho^{n-2k} \tag{3}$$

A figura 9 representa a topografia de alguns polinômios de Zernike de baixa ordem dentro do disco unitário.

Para calcular os valores dos momentos de Zernike de uma imagem, a mesma deve ser tratada como uma função $f(x, y)$, onde x e y são as coordenadas espaciais e f representa o valor intensidade do nível de cinza naquele ponto. O procedimento é descrito conforme o algoritmo 1.

4.4 Máquinas de Aprendizado

Máquinas de aprendizado são sistemas artificiais com a capacidade de extrair informações e similaridades em conjuntos de dados, de forma a aprender padrões de

Algoritmo 1 Algoritmo para cálculo dos momentos de Zernike de uma imagem

```
1: procedure CalculaMomentosZernike(ImagemI, Inteiron, Inteirom)
2:     SumZp ← 0
3:     SumZi ← 0
4:     cont ← 0
5:     for ρ ← 0 até 1 do                          ▷ Cobrindo todo espaço radial
6:         for φ ← −π até π do
7:             SumZp ← SumZp + I(ρ, φ) × Zp(n, m, ρ, φ)        ▷ Polinômio par
8:             SumZi ← SumZi + I(ρ, φ) × Zi(n, m, ρ, φ)        ▷ Polinômio ímpar
9:             con ← cont + 1
10:    Retorna (n + 1) × √(SumZp² + SumZi²) / cont      ▷ Retornando valor normalizado
```

informações relevantes dentro desses dados. Sistemas classificadores são tipos de máquinas de aprendizado que possuem a capacidade de inferir, a partir das características ou atributos de uma dada instância, a qual grupo ou classe a referida instância pertence, a partir de critérios internos à arquitetura de tal classificador. Esses classificadores podem ser bio-inspirados, como no caso das redes neurais artificiais, onde normalmente são máquinas de aprendizado estatístico que realizam uma etapa de treinamento (aprendizado) de forma a alterarem suas características internas, moldando-se às propriedades relevantes de grupos de instâncias já classificadas. Depois de treinados, esses classificadores devem ser capazes de extrapolar o "conhecimento" adquirido e com isso, classificar com certa acurácia um indivíduo desconhecido, mas que pertença à uma das classes do conjunto de treinamento.

4.4.1 Redes Neurais Artificiais

O trabalho em redes neurais artificiais, usualmente denominadas "redes neurais", tem sido motivado desde o começo pelo reconhecimento de que o cérebro humano processa informações de uma forma inteiramente diferente do computador digital (HAYKIN, 2001). Tendo isto em vista, foram propostos modelos bio-inspirados, ou seja, na tentativa de emular o raciocínio humano através de máquinas, principalmente o computador digital, modelos matemáticos inspirados nas estruturas biológicas do próprio cérebro humano foram desenvolvidos para tentar promover o processamento de informações de forma semelhante à capacidade cognitiva humana.

O cérebro é um computador (sistema de processamento de informação) altamente complexo, não-linear e paralelo. Ele tem a capacidade de organizar seus constituintes

estruturais, conhecidos por neurônios, de forma a realizar certos processamentos (p.ex., reconhecimento de padrões, percepção e controle motor) muito mais rapidamente que o mais rápido computador digital hoje existente (HAYKIN, 2001).

As redes neurais artificiais são estruturas computacionais que tentam emular o comportamento do cérebro humano. Para tal, essas redes são formadas pela interligação de estruturas computacionais chamadas neurônios. Esses neurônios são processadores de dados que tentam se comportar como neurônios naturais, recebendo estímulos nas entradas, processando esses estímulos e através de uma função de ativação exibem uma saída, que para o caso mais simples (o perceptron de camada única) é a simples ativação, ou não, da saída, em valor alto ou baixo dependendo do valor das entradas e de um limiar (*threshold*). Nesses modelos, as conexões sinápticas são modeladas por pesos que multiplicam o valor das entradas e é justamente nesses pesos que está codificado o conhecimento adquirido pela rede.

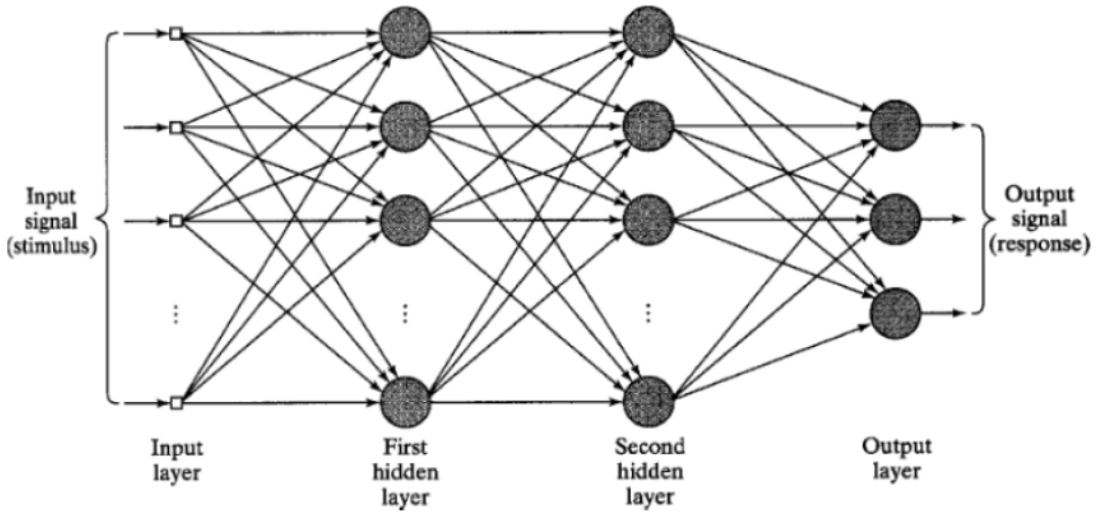
As redes neurais artificiais também emulam uma característica muito importante do raciocínio humano, que é o aprendizado. Existem basicamente dois tipos de aprendizado para redes neurais, supervisionado e não-supervisionado.

Para o aprendizado supervisionado, temos um conjunto de treinamento com pares formados por informações de entrada e sua respectiva informação de saída (ou padrão desejado) que já é conhecida, com isso, depois de treinada a rede, outras entradas com saídas não conhecidas são apresentadas à mesma e ela é capaz de oferecer respostas a esses estímulos, onde o sucesso ou não dessa extrapolação dependerá de fatores como o tipo da rede para determinado problema, o dimensionamento adequado da rede, e o próprio treinamento considerando o algoritmo e o conjunto de dados. O treinamento da rede se dá de forma iterativa, onde os padrões de entrada são apresentados à mesma e, com a medida dos erros em cada camada da rede, é utilizado um algoritmo (sendo o de retro-propagação o mais comum) para a atualização sucessiva dos pesos desse mesma rede.

A figura 10 apresenta o modelo de rede neural conhecido como perceptron de múltiplas camadas que é um modelo que utiliza normalmente funções sigmoide ou tangente hiperbólica como funções de ativação não lineares e além das camadas de entrada e de saída, apresentam pelo menos uma camada escondida e com isso conseguem se tornar apro-

ximadores universais de funções, com aplicações em regressão, classificação, reconhecimento de padrões e outras.

Figura 10 – Rede Neural Artificial Perceptron de Múltiplas Camadas



Fonte: (HAYKIN, 2001)

Para as redes neurais de aprendizado não supervisionado, não existe padrão desejado para o conjunto de treinamento. Isso quer dizer que a rede deve ser capaz de se auto-organizar com o propósito de agrupar os dados de entrada de forma totalmente inerente à própria arquitetura interna da rede. Para isso são usados métodos de treinamento competitivo entre os neurônios, onde essas redes ou mapas de neurônios vão recebendo os dados de entrada e o neurônio vencedor (e por vezes sua vizinhança) vão atualizando seus pesos. Este paradigma de aprendizado tenta emular o mecanismo de inibição lateral das redes neuronais do próprio cérebro humano, dessa forma, um neurônio vencedor, ou um padrão de respostas para esses neurônios vencedores correspondentes à saída da rede são responsáveis por agrupar esses dados de entrada (ou formar *clusters*) de forma totalmente autônoma, sem a necessidade de um "professor" ou padrões desejados de saída.

4.4.2 Máquinas de Vetor de Suporte

Um outro tipo de rede *feed-forward* (semelhante às redes neurais artificiais) com características de aproximação universal são as máquinas de vetor de suporte (CORTES; VAPNIK, 1995).

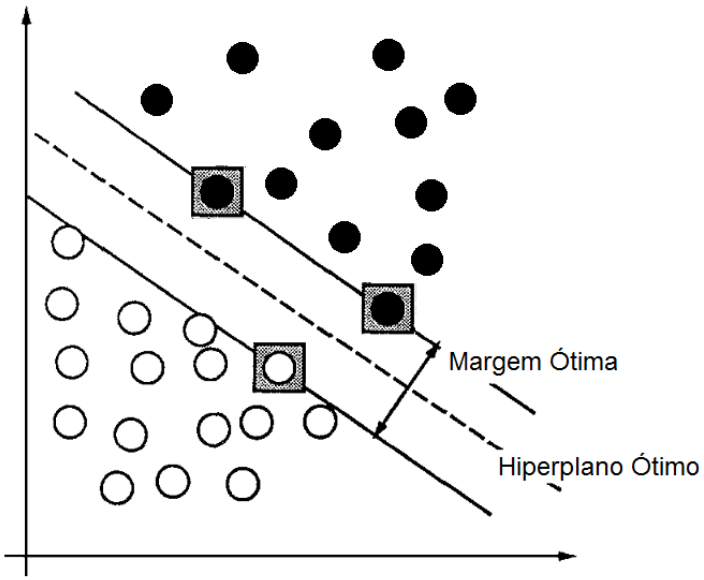
As máquinas de vetor de suporte são basicamente máquinas lineares cujo objetivo consiste em construir um hiperplano como superfície de decisão de tal forma que a margem de separação entre exemplos positivos e negativos seja máxima (HAYKIN, 2001). Dado um conjunto de exemplos de treinamento, com cada exemplo marcado como pertencente a uma de duas categorias, o algoritmo de treinamento das máquinas de vetor de suporte constrói um modelo que associa novos exemplos a uma categoria ou à outra, se comportando, portanto, como um classificador binário linear e não-probabilístico. Um modelo é uma representação dos exemplos ou instâncias como pontos no espaço multi-dimensional, mapeados de tal maneira que os mesmos estejam separados (levando em consideração a categoria à qual pertencem) por um hiperplano ótimo, ou seja, que este hiperplano gere um espaço de separação que seja o mais largo possível. Desta maneira, novas instâncias são mapeadas para este espaço e categorizadas de acordo com a região na qual a mesma está situada.

Para ser capaz de realizar tarefas de classificação com uma abordagem não-linear, é utilizado o chamado truque do *kernel*, de forma a mapear implicitamente as entradas em um espaço de atributos de maior dimensionalidade, onde espera-se que dessa forma, instâncias de diferentes classes sejam mapeadas para regiões disjuntas deste novo hiper-espaço. Normalmente, os dados não são linearmente separáveis, por isso, o truque do *kernel* deve ser considerado.

Na figura 11 está apresentado um exemplo bidimensional para a classificação pela máquina de vetor de suporte. Observa-se as duas diferentes classes, o separador ótimo é a reta apresentada nesta figura, pois a margem (distância entre a reta e a instância mais próxima de cada classe) é máxima, o que possibilita o máximo distanciamento entre ambas as classes, de forma a permitir que uma nova instância desconhecida tenda a ser classificada corretamente, ficando do lado correto da reta.

O treinamento de máquinas de vetor de suporte necessita da solução da otimização de complexos problemas de programação quadrática. Redes de otimização sequencial mínima (*Sequential Minimal Optimization* - SMO) são algoritmos utilizados para resolução de problemas deste tipo, desenvolvido por (PLATT, 1998), e é utilizado para o treinamento das máquinas de vetor de suporte. Basicamente, as redes de otimização sequencial mínima permitem a resolução de problemas de programação quadrática complexos e grandes com

Figura 11 – Separador Ótimo de Padrões pelas Máquinas de Vetor de Suporte



computação razoável através da partição desses problemas em outros menores, que são resolvidos analiticamente.

4.5 Análise de Componentes Principais

Análise de componentes principais (PEARSON, 1901) é uma técnica estatística que através de uma transformação ortogonal é capaz de converter o espaço de observações com variáveis linearmente dependentes em um outro espaço de variáveis independentes linearmente que são chamados de componentes principais.

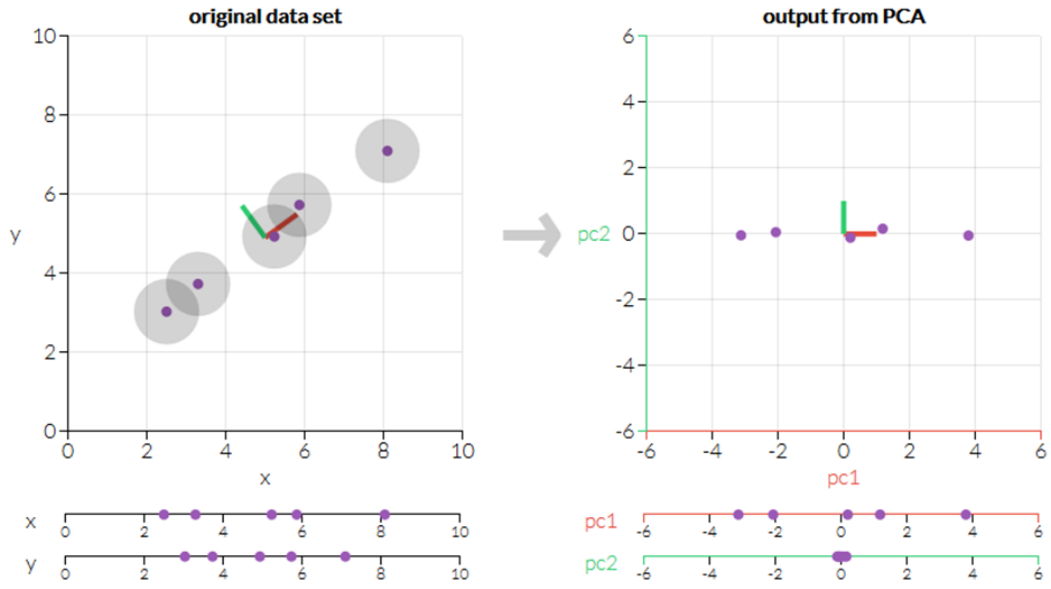
Basicamente, o que a análise de componentes principais faz é gerar uma matriz de covariância, que é simétrica, e a partir dos auto-vetores são obtidas as componentes principais, sempre ordenadas de tal maneira que a primeira componente principal é aquela que possui a maior variância e assim por diante, com a restrição de que todas as componentes principais devem ser ortogonais entre si. O resultado é um novo espaço de variáveis que pode ser menor ou igual em tamanho ao espaço original, onde essas novas variáveis são combinações lineares das originais.

O processo de análise costuma iniciar com uma subtração da média dos dados, com o intuito de deslocar o "centro de massa" para a origem do espaço, para depois ser calculada a referida matriz de covariância, calcula-se os auto-valores e seus correspondente

auto-vetores, com a consequente ortogonalização deste conjunto de auto-vetores. Em seguida, é realizada uma etapa de normalização (mudança de escala) dos autovetores. Ao final, a contribuição de cada auto-vetor é dada pela divisão do auto-valor associado pelo somatório de todos esses auto-valores.

Na figura 12 pode-se observar uma ilustração exemplificando o processo de análise de componentes principais para duas variáveis (representando um espaço bidimensional). Observa-se no gráfico à esquerda desta mesma figura, que visivelmente os dados estão dispostos ao longo de uma reta inclinada e percebe-se que, pela projeção dos dados ao longo de ambos os eixos x e y, que a informação presente em ambas as variáveis apresenta grande variância. O que a análise de componentes principais realiza na prática é modificar esses eixos (variáveis) através da combinação linear de ambos, onde a primeira componente principal apresenta a maior variabilidade dos dados (observando no gráfico à direta da figura 12) enquanto que a segunda componente principal deve ser a com maior variância depois da primeira, com a restrição de ser ortogonal a esta. Como é possível observar neste exemplo, a segunda componente agrega o mínimo de informação, de maneira que se essa for eliminada, haveria uma perda de informação que poderia ser considerada desprezível.

Figura 12 – Ilustração Sobre o Processo de Análise de Componentes Principais



Fonte: (POWELL; LEHE, 2015)

A análise de componentes principais pode ser utilizada para a redução de atributos. Tendo em mente que esta técnica destaca os "pontos-de-vista" dos dados onde há maior variância e que o sistema classificador se beneficia deste fato (é bastante razoável imaginar isto), então remover componentes que estão no final da lista, ou seja, as componentes principais que possuem o mínimo de (ou nenhuma) variância acarretaria em uma redução de atributos de entrada com pouco impacto na perda de informação para a classificação. O resultado seria um vetor de entrada bastante reduzido (a depender da conformação dos dados no espaço) e com o menor impacto possível ao desempenho de classificação (a depender da quantidade de componentes removidas e do percentual de variância o qual essas contribuíam).

4.6 Ganho de Informação

O conceito de ganho de informação está atrelado à área de teoria da informação. Dentro do campo de aprendizado de máquina, o ganho de informação pode ser utilizado como uma métrica para avaliar a evolução da "pureza" de um conjunto de observações a partir da divisão desse grupo através do valor de uma dada variável (atributo) do mesmo.

Dentro do campo da teoria da informação existe uma métrica conhecida como entropia (mais especificamente a entropia de Shannon (SHANNON, 1948)) que representa o valor esperado da informação contida em uma dada mensagem. A expressão da entropia de Shannon está presente na equação 4. De modo grosseiro, esta entropia da informação representa uma medida da "impureza" dos dados dentro de um conjunto, ou uma "quantificação da novidade" que uma seleção aleatória dentro de um grupo pode ter ao ser selecionada uma instância de um tipo ou de outro.

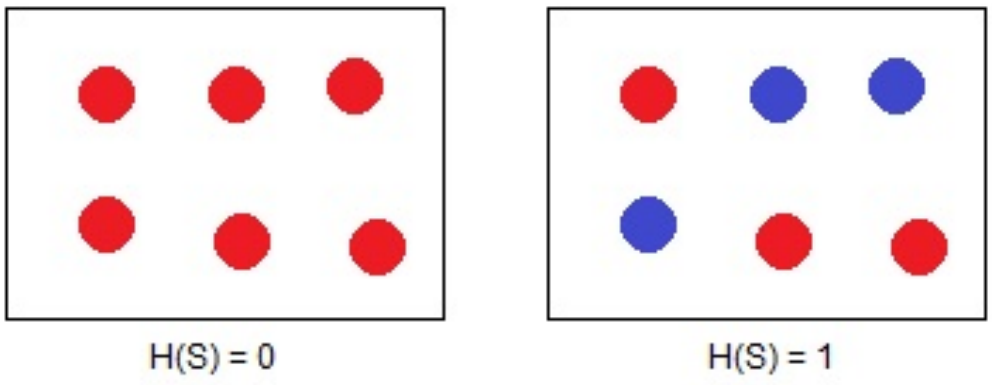
$$H(S) = -\sum_{s_i \in S} p(s_i) log_2 p(s_i) \tag{4}$$

Onde H representa a entropia, S representa o conjunto e p a probabilidade.

Um exemplo visual está presente na figura 13. Observa-se que no conjunto de seis círculos vermelhos à esquerda, a probabilidade de selecionar, dentre eles, aleatoriamente um círculo vermelho é igual a um, portanto utilizando a equação 4 chega-se a um valor de

entropia igual a zero, ou seja, não há novidade alguma na informação de que foi selecionado um círculo vermelho de um grupo de círculos vermelhos. De forma semelhante, no grupo da direita, há uma probabilidade de 0,5 de se selecionar um círculo vermelho e a mesma probabilidade para selecionar um círculo azul, de maneira que utilizando a mesma equação chega-se a um valor de entropia igual a um, o que indica que o mesmo não é um conjunto puro (homogêneo).

Figura 13 – Ilustração Sobre a Medida de Entropia da Informação



Utilizando a entropia da informação, pode-se definir o conceito de ganho de informação como uma métrica para avaliar a variação de entropia provocada após a divisão do conjunto baseado no valor relativo a algum atributo das instâncias deste conjunto. O ganho de informação pode ser calculado segundo a equação 5.

$$\begin{aligned} IG(a,S) =& H(S) - \sum_{t \in T} p(t)H(t) \\ S =& \bigcup_{t \in T} t \end{aligned} \tag{5}$$

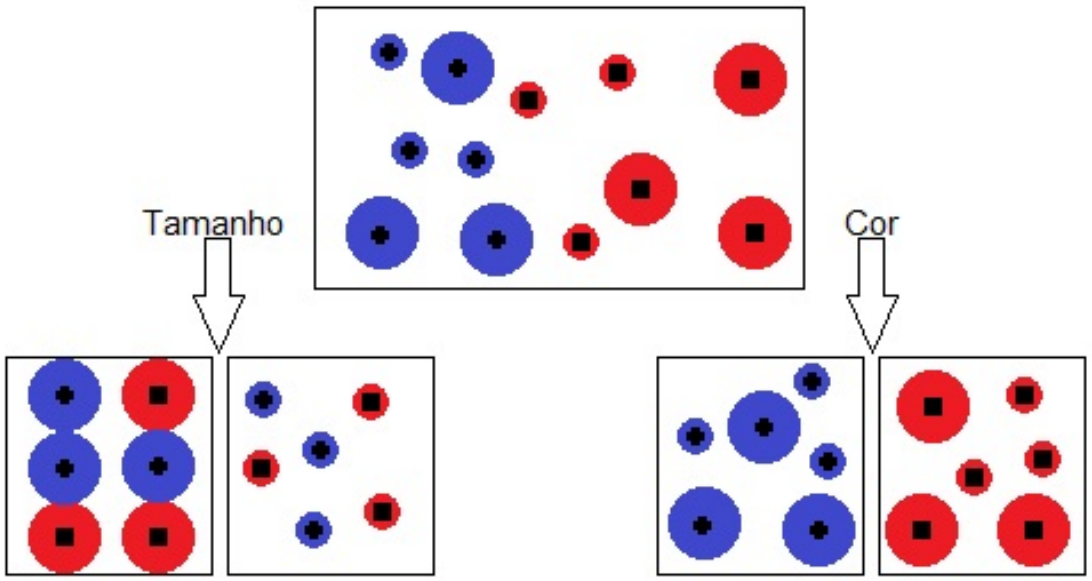
Onde IG representa o ganho de informação, a o atributo para segmentar o conjunto (representado por S), H representa a entropia, p a probabilidade, T são os conjuntos resultantes da separação.

Este processo pode ser ilustrado conforme a figura 14. Imaginando-se que os círculos apresentam os atributos cor, tamanho e símbolo, onde este último seria o atributo desejado (ou classe) e será utilizado para medir a entropia dos conjuntos. Antes da separação, a

entropia do conjunto é igual a um (0,5 de probabilidade de selecionar um círculo com o simbolo quadrado e igualmente para símbolo de cruz), após a divisão do conjunto original utilizando o atributo tamanho, percebe-se que ambos os conjuntos resultantes continuam com suas entropias iguais a um, nesta situação o ganho de informação foi nulo. Porém, no caso do atributo cor, a entropia de ambos os conjuntos é igual a zero e, segundo a equação 5 o ganho de informação é igual a um (valor máximo).

O exemplo abordado na figura 14 pode ser utilizado para entender como seria realizada a seleção de atributos baseado nesta técnica. Seria calculado o ganho de informação referente a todos os atributos do conjunto, em seguida os mesmos seriam ordenados e os de maior valor selecionados.

Figura 14 – Ilustração Sobre o Ganho de Informação de um Conjunto Mediante a Divisão pelos Atributos Cor e Tamanho



4.7 Seleção de Atributos Baseados em Correlação

Um problema de grande relevância para a área de aprendizado de máquina é o que concerne à seleção de um conjunto de atributos que seja extremamente representativo com o objetivo de construir um modelo de classificação para uma tarefa específica.

Existem basicamente duas abordagens diferentes para tratar de seleção de atributos, nomeadamente *Wrapper* e *Filter*. Esta primeira abordagem trata essencialmente do uso do próprio sistema classificador como métrica para avaliar o desempenho de diversos subconjuntos de atributos. Esses subconjuntos são, em problemas práticos de complexidade razoável, selecionados por algoritmos de busca utilizando heurísticas, tendo em vista que para um conjunto com n atributos, então existe um total de n^2 subconjuntos possíveis a serem explorados. A abordagem baseada em *Wrapper* demonstra ser muitas vezes inviável na prática pelo fato de utilizar o modelo classificador a cada avaliação de um subconjunto e mesmo com o uso de ferramentas de busca heurística para não percorrer todo o espaço de busca, ainda assim o esforço computacional é muito grande, os modelos de classificação estão ficando cada vez mais complexos. A vantagem desses tipo de abordagem é que a solução de subconjunto encontrada tende a ser otimizada para o classificador utilizado na seleção, mas isso também pode ser visto como uma desvantagem, pois no caso de uma mudança de sistema classificador, o subconjunto antes selecionado pode não ser mais o ótimo e ser necessária a realização de seleção mais uma vez com o novo classificador, tarefa essa bastante custosa.

Como uma alternativa de contornar as limitações apresentadas existe a abordagem *Filter*. Nesse tipo de abordagem, são utilizadas métricas heurísticas para tentar encontrar um subconjunto que satisfaça essas métricas na esperança de que a mesma seja de certa maneira compatível com o sistema classificador, o que não é absurdo de se supor. O subconjunto encontrado por esse tipo de técnica normalmente não é o ótimo para o classificador, mas tem a vantagem de ser independente do mesmo, ou seja, o subconjunto selecionado para ser usado em um classificador é o mesmo para ser utilizado com outro diferente (a seleção é totalmente desacoplada do sistema alvo). A maior vantagem desse tipo de abordagem é o fato de que essas métricas heurísticas são muito mais simples computacionalmente falando do que gerar um modelo de classificador para cada subconjunto avaliado. São exemplos de sistemas seleção do tipo *Filter*: Sistema baseado em análise de componentes principais, onde as componentes mais relevantes de acordo com a variância são selecionadas; sistemas baseados em ganho de informação, onde os atributos são ordenados em uma lista conforme o valor do ganho de informação que cada um desses atributos tem quando divide o conjunto de instâncias em relação à classe; e outro exemplos seria a técnica

de seleção de atributos baseada em correlação (*Correlation-based Feature Selection* - CFS) (HALL, 1999).

A técnica CFS é baseada na avaliação de subconjuntos (*CFS Subset Evaluation*), ou seja, ela trabalha em conjunto com uma técnica de busca heurística para não explorar exaustivamente todo o espaço de busca, com a ideia de utilizar alguma métrica de correlação (por vezes o coeficiente de correlação de Pearson (HALL, 1999)) para quantificar tanto a dependência atributo-classe quanto a dependência atributo-atributo. Essa abordagem é utilizada, pois atributos que são fracamente correlacionados com a classe tendem a ser irrelevantes para a classificação (lembrando que o classificador não é um sistema linear, portanto relações não-lineares que não são mensuradas por métricas lineares como a correlação podem passar despercebidos), do mesmo modo que atributos que são fortemente correlacionados entre si tendem a ser redundantes.

O maior objetivo das técnicas de redução de atributos é justamente remover os irrelevantes e redundantes, pois os mesmos somente contribuem para o crescimento desnecessário do classificador e por vezes podem até dificultar todo o processo de classificação trazendo informações que dificultam a construção do modelo do classificador.

A equação 6 descreve uma métrica possível para esta técnica (HALL, 1999):

$$r_{zc} = \frac{k\bar{r}_{zi}}{\sqrt{k + k(k-1)\bar{r}_{ii}}} \tag{6}$$

Onde r_{zc} representa o coeficiente de correlação do subconjunto selecionado, k é a quantidade de atributos, $\bar{r}_{zi}$ é a média das correlações entre atributos e a classe, $\bar{r}_{ii}$ é a média de correlação intra-atributos. Observa-se que conforme a média da correlação dos atributos com a classe então o coeficiente final cresce, ao contrário do que acontece quando a média da correlação intra-atributos é elevada.

Utilizando métricas semelhantes à da equação 6 como função objetivo para os algoritmos de busca, então espera-se que ao final do processo seja encontrado um subconjunto de atributos que seja representativo das instâncias e ao mesmo tempo satisfatoriamente reduzido com a remoção de atributos considerados irrelevantes ou redundantes.

4.8 Algoritmos de Busca

Dentro da ciência da computação, algoritmos de busca são algoritmos utilizados para procurar, dentro de um espaço de busca definido, por instâncias que possuam propriedades ou atributos desejados. Normalmente as buscas utilizam heurísticas, que são abordagens para resolução de problemas que aplicam métodos práticos sem garantia de resultados ótimos ou perfeitamente corretos, porém obtendo soluções suficientemente boas para o dado problema. Muitas técnicas de busca são inspiradas em processos totalmente não relacionados com o problema ao qual são utilizadas para resolver. Algumas delas são inspiradas na teoria da evolução de Darwin (GOLDBERG, 1989), outras são inspiradas em comportamentos de sociedades humanas e filosofia (SANTOS; ASSIS, 2013) e outras são inspiradas em comportamento coletivo de bandos de animais (KENNEDY; EBERHART, 1995). Algoritmos de busca com heurísticas são bastante utilizados, pois realizar uma busca cega exaustiva na maioria dos problemas práticos torna-se uma tarefa inviável ou até mesmo impossível.

4.8.1 Computação Evolucionária

Dentro da área de inteligência computacional estão inseridos os chamados algoritmos de computação evolucionária. Estes algoritmos (na verdade uma família de algoritmos) são ditos evolucionários pois têm como maior inspiração a teoria da evolução de Darwin (DARWIN, 1859) e também as propriedades de mutação, cruzamento e seleção natural da genética. Esses algoritmos tentam emular a característica de sobrevivência do mais apto, onde uma solução mais adaptada a uma problema leva vantagem contra uma menos adaptada, sobre a penalidade desta última ser extinta enquanto a primeira permanece e tem ainda a possibilidade de perpetuar seus "genes" através do cruzamento (combinação) com outro indivíduo também apto.

Existem diversas técnicas de busca e otimização baseados no conceito da computação evolucionária, são exemplos a programação genética, evolução diferencial, programação evolucionária, busca evolucionária, porém o mais conhecido são os algoritmos genéticos.

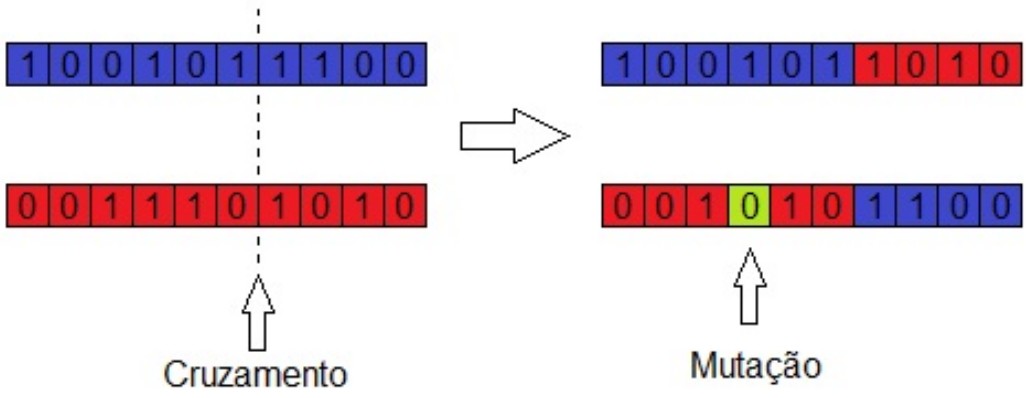
Os algoritmos genéticos (AG) representam uma das abordagens da computação evolucionária, onde são assimilados os conceitos de fenótipo, genótipo, cromossomos,

cruzamento, seleção natural, mutação e outros. os AG consistem de processos iterativos para solução de problemas de busca e otimização utilizando heurística. As candidatas a soluções são modeladas como cromossomos, que normalmente são vetores binários. Inicialmente a população é gerada aleatoriamente. A cada iteração, a aptidão dos indivíduos é avaliada (utilizando a função objetivo) de forma que parte do conjunto que se demonstra menos apto (pior valor de avaliação da função objetivo) é eliminado. Do conjunto restante de indivíduos aptos são gerados novos através de um operador de cruzamento, que normalmente é um valor escolhido aleatoriamente para indicar o ponto de *cross-over*, que é o ponto onde dois indivíduos pais realizam a troca de parte da sua informação genética, gerando outros dois filhos com parte da informação de cada pai. Em seguida há a etapa de mutação que refere-se à possibilidade (normalmente uma probabilidade baixa de ocorrer) da alteração de um bit no vetor de indivíduos, este operador existe para gerar diversidade nas soluções como uma forma de tentar evitar que as soluções fiquem presas em mínimos locais. Em alguns casos, é utilizado o conceito de elitismo, onde indivíduos de alto valor de aptidão são levados diretamente para a próxima geração, não passando pela etapa de seleção natural (que tem um certo fator de aleatoriedade).

Muitos outros aspectos diferenciam as diversas técnicas de computação evolucionária, inclusive algumas alterações mínimas nas técnicas originais também são praticas comuns, fazendo com que haja uma variedade grande de algoritmos de computação evolucionária.

Na figura 15 observa-se um exemplo de operadores de cruzamento e mutação em algoritmos de computação evolucionária. O ponto de cruzamento ou *cross-over* é selecionado aleatoriamente e a partir de dois indivíduos geradores, são obtidos novos indivíduos resultantes da combinação das informações contidas nos cromossomos desses "pais". O operador de mutação é representado por uma alteração aleatória no valor de algum gene (neste caso um *bit-flip*) cuja probabilidade de ocorrência é normalmente muito baixa, este tipo de operador existe para gerar diversidade nos indivíduos que percorrem o espaço de busca. Esses e outros operadores tentam emular, de forma muito simplificada, o comportamento natural de evolução das espécies.

Figura 15 – Exemplo de Operadores de Cruzamento e Mutação em Algoritmos de Computação Evolucionária



4.8.2 Otimização por Enxames de Partículas

Otimização por enxames de partículas é um algoritmo de busca desenvolvido por (KENNEDY; EBERHART, 1995) e modela de maneira simplificada o comportamento coletivo de um bando pássaros ou cardumes de peixes.

Cada indivíduo da população, neste algoritmo representa uma possível solução para o problema, em um espaço multi-dimensional cada indivíduo é um vetor cujo tamanho é a dimensão do espaço. Os indivíduos iniciam aleatoriamente dentro do espaço de busca e possuem informação de posição e velocidade e através disso, vão se movendo pelo espaço de busca levando em consideração ambos os comportamentos individuais quanto coletivos.

O problema pode ser de minimização ou de maximização da função objetivo e esta é uma função que é utilizada para medir a aptidão atual de cada um dos indivíduos (candidatos a solução do problema). Cada um dos indivíduos tende a procurar retornar à posição onde obteve a melhor avaliação da função objetivo, mas também procura pela melhor posição global de todo o bando e nesse conflito de interesses, o bando tende a progredir enquanto realiza a busca pelo espaço, tendendo a procurar máximos (ou mínimos) globais. A cada iteração do algoritmo, os vetores de posição e velocidade de cada indivíduo é atualizado conforme essas buscas pela melhor posição coletiva (busca global) e individual (busca local). Embora que devido ao fato de ser utilizada uma heurística, a solução encontrada muitas vezes não é a ótima, bem como existe a possibilidade de

divergência e não ser encontrada solução, por isso, os parâmetros da busca devem ser cuidadosamente selecionados.

São parâmetros da otimização por enxames de partículas, o número de indivíduos no enxame, o fator e inércia, e as constantes de constrição, também conhecidas como coeficientes de aceleração.

As regras de busca do algoritmo de otimização por enxames de partículas está expresso na equação 7.

$$\begin{aligned} \mathbf{x}_i(n+1) &= \mathbf{x}_i(n) + \mathbf{v}_i(n+1) \\ \mathbf{v}_i(n+1) &= w\mathbf{v}_i(n) + c_1 r_1(n)(\mathbf{p}_i(n) - \mathbf{x}_i(n)) + c_2 r_2(n)(\mathbf{p}_g(n) - \mathbf{x}_i(n)) \end{aligned} \tag{7}$$

Onde:

1. m: O número de indivíduos no enxame
2. w: O fator de inércia
3. $r_1(n)$ e $r_2(n)$: Números aleatórios uniformemente distribuídos entre zero e um
4. c_1 e c_2: Constantes de constrição ($c_1 + c_2 = 4$)
5. $\mathbf{x}_i$: Posição
6. $\mathbf{v}_i$: Velocidade
7. $\mathbf{p}_g$: Melhor posição global
8. $\mathbf{p}_i$: Melhor posição individual

O algoritmo 2 descreve o procedimento desta técnica de busca.

4.8.3 *Best First*

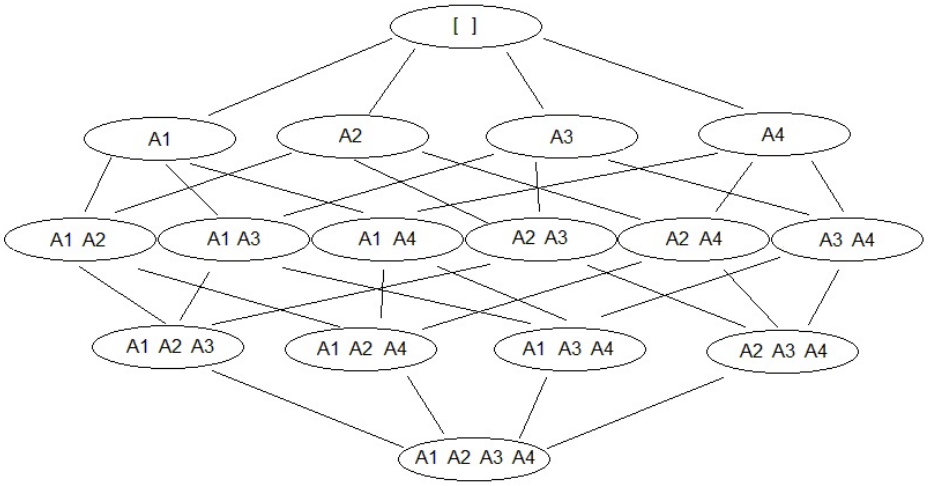
Best First é uma estratégia de busca que explora o grafo (que representa o espaço de busca) a partir de uma abordagem gulosa, sempre buscando o vértice vizinho mais promissor, mas permitindo a possibilidade de retroceder (*backtracking*) caso o caminho tomado não seja promissor.

Algoritmo 2 Algoritmo para realização da otimização por enxames de partículas

```
1:  procedure PSO
2:      Inicialização
3:      Gera indivíduos aleatoriamente
4:      Procura as melhores posições individuais
5:      Procura a melhor posição global
6:      Ajusta velocidades
7:      Ajusta posições
8:      if Critério de parada alcançado then
9:          Retorna A melhor posição geral
10:     else
11:         GoTo : 4
```

Este algoritmo usa uma representação do espaço de busca que estratifica os vértices em níveis onde as arestas conectam apenas vizinhos que variam em um único elemento do vetor. De acordo com a figura 16, pode ser observado que há duas possibilidades de ponto inicial, que são o vértice com o vetor vazio (acima) e o vértice com todos os quatro atributos no vetor (embaixo). A1, A2, A3 e A4 estão representando todos os quatro atributos a serem selecionados e cada um dos vértices contém um vetor com diferentes combinações da presença ou ausência desses atributos. Percebe-se ainda, de acordo com a mesma figura, que vértices adjacentes são aqueles que acrescem ou removem apenas um atributo.

Figura 16 – Grafo representando o espaço de busca para o algoritmo *Best First*



No algoritmo 3 está explicitado o procedimento de busca. Como já mencionado, existem duas possibilidades de estado inicial (totalmente vazio ou totalmente completo) correspondente a um parâmetro do algoritmo de busca. Após selecionado o estado inicial,

são inicializadas as lista que auxiliam no percorrimento do grafo, a primeira lista (lista A) representa a lista com os vértices candidatos a serem explorados e ela deve ser inicializada com o vértice inicial. Já a lista B (que representa os vértices já visitados) deve ser inicializada vazia. O laço principal do procedimento se inicia escolhendo o vértice presente na lista A que possua o maior índice de avaliação (segundo a heurística utilizada no problema de busca), em seguida, este vértice é movido para a lista de vértices já visitados e sua avaliação de aptidão é comparada com o vértice que atualmente é considerado o melhor e caso seja superado, este novo vértice tomará este lugar, em sequência, cada um dos vértices não-explorados adjacentes ao último vértice visitado é adicionado à lista A para possibilitar que sejam explorados. Por fim, é checada a condição de parada, que também é um parâmetro, e caso o vértice de melhor aptidão tenha sido alterado dentro das n últimas iterações, então permanece o laço em busca de outros vértices, caso contrário a busca é encerrada, retornando o atual vértice com melhor aptidão.

Algoritmo 3 Algoritmo para realização da busca pelo *Best First*

```
1:  procedure BestFirst
2:      Vértice V ← Estado Inicial
3:      Lista A ← Insere V                    ▷ Lista com candidatos a serem avaliados
4:      Lista B ← Vazia                       ▷ Lista com Vértices já visitados
5:      Vértice C ← Vértice com maior avaliação da lista A
6:      Remove C da lista A
7:      Insere C na lista B
8:      if φ(C) ≥ φ(V) then      ▷ φ(X) : Valor da Avaliação da aptidão do Vértice
9:          V ← C
10:     for Cada Vértice T adjacente de C do          ▷ Realizando a expansão
11:         if C não pertence a A ou B then
12:             Insere T em A
13:     if V foi alterado nas últimas n expansões then          ▷ Critério de parada
14:         GoTo : 5
15:     Retorna V                                  ▷ Retornando melhor Vértice
```

5 Metodologia Proposta

Pode-se definir pesquisa como o procedimento racional e sistemático que tem como objetivo proporcionar respostas aos problemas que são propostos (GIL, 2002). Sob esta ótica, tem-se que a pesquisa deve ser um procedimento sistemático e a maior implicação disto está na repetibilidade dos experimento e abordagens, em outras palavras, não possui grande importância a ciência (mais precisamente os experimentos) que não pode ser reproduzida por outros experimentadores quando submetidos às mesmas condições, ou condições similares. Isto é, quando aplicando as mesmas abordagens e condições para a execução de um dado experimento, as mesmas conclusões devem ser obtidas (ciência deve ser independente das influências de quem executa os experimentos). Para que isto ocorra mais facilmente e de maneira mais imparcial possível, toda a metodologia e protocolo experimental de uma pesquisa deve ser exposto, qualquer etapa que seja obscura ou não esteja bem explicitada pelo pesquisador está fadada a não contribuir para a universalidade ou democratização da ciência e assim, a contribuição trazida por um projeto de pesquisa com falta de clareza no processo metodológico acaba perdendo um pouco da sua confiabilidade.

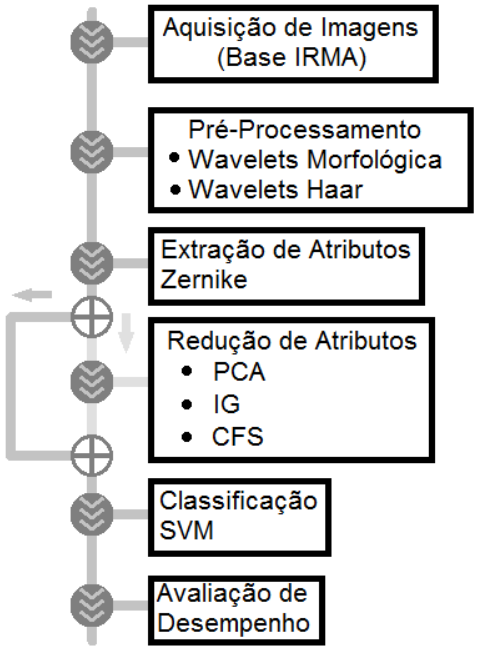
Para corroborar com a assertiva anteriormente descrita, serão relatadas todas as etapas da proposta deste trabalho e, com isto, elucidar através das partes, a abordagem experimental como um todo.

Como já mencionado, este trabalho tem como o principal objetivo, o estudo comparativo de algumas técnicas de seleção de atributos aplicados ao problema de classificação de tumores de mama a partir do processamento de imagens de mamografia, onde o sistema proposto está baseado na abordagem realizada por (FERNANDES, 2015), sendo o atual trabalho uma continuação do citado.

A figura 17 resume o fluxograma do sistema aqui proposto. As imagens que alimentam o sistema são provenientes da base IRMA, onde são selecionadas aquelas pertinentes para cada um dos quatro tipos diferentes de composição tecidual da mama. Para a etapa de pré-processamento das imagens, duas abordagens são utilizadas em todas as combinações possíveis de composição mamária e presença ou ausência de seleção de atributos, essas abordagens são a transformada de wavelets de Haar e a transformada de wavelets mor-

fológica. Os polinômios de Zernike são utilizados para a extração dos atributos baseados nos momentos de Zernike. Em seguida, para o sistema proposto, é possível realizar seleção de atributos ou seguir para a próxima etapa com todos os atributos gerados na etapa anterior, caso seja desejado o uso de seleção de atributos, então uma das três técnicas é utilizada (análise de componentes principais, ganho de informação ou seleção de atributos baseado em correlação). A etapa seguinte realiza o treinamento e posterior classificação com a máquina de vetor de suporte. Por fim, o desempenho do sistema é avaliado através da análise comparativa das taxas de acerto para cada uma das combinações individuais de técnicas.

Figura 17 – Fluxograma do sistema inteligente proposto



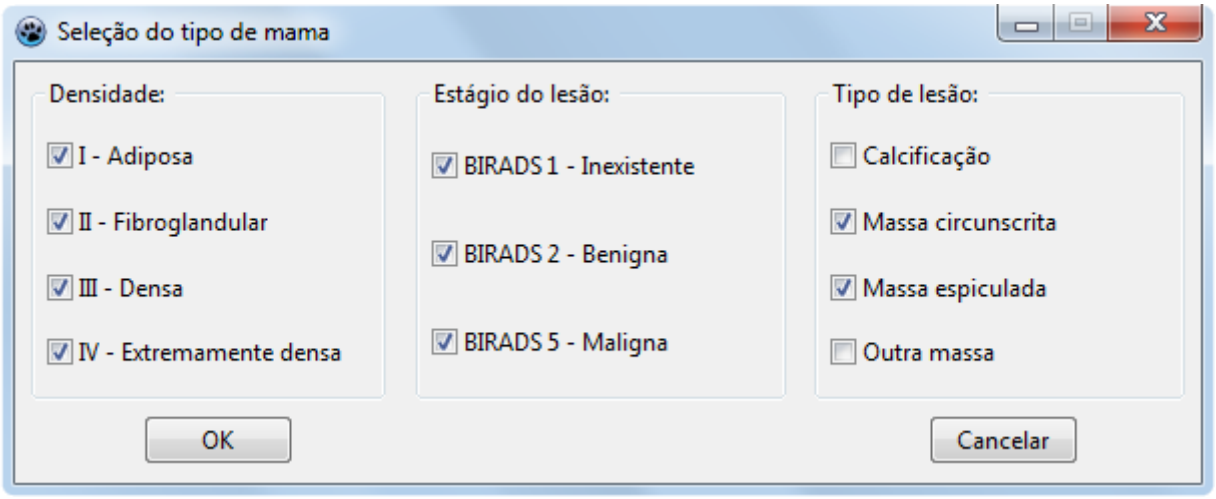
5.1 Preparação e Processamento das imagens

Inicialmente, como descrito anteriormente, todas as imagens mamográficas utilizadas neste trabalho foram obtidas da base de imagens IRMA[1] (OLIVEIRA et al., 2008; DESERNO et al., 2012) . É importante frisar que estas imagens são *patches* da região de interesse (lesões ou tecidos normais), de forma que isto implica que este trabalho está focado na etapa de classificação propriamente dita. Levando em conta a correlação existente entre a forma da massa (circunscrita ou espiculada) e a malignidade ou não da mesma

[1] Cortesia de TM Deserno, Dept. de Informática Médica, RWTH Aachen, Alemanha

(massas mais “comportadas” geometricamente tendem a ser benignas e o oposto tende a ser verdadeiro também), este trabalho explora o uso de um descritor de forma em imagens e com isso, se atém a selecionar somente os tipos de massas explicitados, conforme figura 18, desconsiderando por exemplo os casos de lesões de forma desconhecida ou microcalcificações, tendo em vista que este último tipo de lesão, por exemplo, é classificado utilizando outros métodos, não sendo o aqui proposto capaz de classificá-lo.

Figura 18 – Tipos de mamas, Classificação BI-RADS e tipos de lesão para seleção de imagens no ConvIRMA



A base IRMA possui um total de 2796 imagens mamográficas divididas em 12 subclasses, possuindo 233 imagens cada. Como não foram utilizadas todas os diferentes tipos de lesão neste trabalho, então as quantidades de imagens aqui utilizadas estão como descrito na tabela 1. Pelo fato de a seleção de imagens ter desequilibrado a quantidade original de 233 instâncias por subclasse, então a etapa de balanceamento com inclusão de indivíduos sintéticos a partir dos indivíduos originais se faz mandatória para não tornar o classificador viciado em determinada classe.

Tabela 1 – Descrição das quantidades de imagens utilizadas da base IRMA

Tipo de Tecido	Normal	Benigno		Maligno	
		Circuns.	Espic.	Circuns.	Espic.
Ext. Denso	233	32	7	2	41
Denso	233	45	11	9	63
Fibroglandular	233	86	5	12	95
Adiposo	233	66	6	27	56

Para garantir uma melhor discriminação entre as classes de lesões (maligna, benigna e ausente) é utilizada, neste trabalho, a abordagem de dividir o conjunto de imagens selecionadas de acordo com a composição e densidade do tecido mamário. Na base de imagens IRMA, as mesmas estão subdivididas no que tange à densidade em: Extremamente densas; densas; de composição predominantemente fibroglandular e predominantemente adiposas. Na prática, isto significa que para cada um desses rótulos, será treinado um classificador diferente. É interessante mencionar que em uma situação de utilização prática do sistema, caso a imagem não esteja rotulada previamente pelo profissional radiologista quanto à densidade da mama, a idade da paciente pode se tornar uma variável substituta neste caso, tendo em vista a grande correlação (negativa) entre a densidade da composição da mama com a idade, sendo esta última uma variável clínica extremamente simples e acessível, porém também pode ser proposto o desenvolvimento de um outro sistema de classificação automático (que não faz parte do escopo deste trabalho) que processe a imagem e classifique-a com relação à densidade servindo como mais uma entrada para o sistema aqui proposto.

A primeira etapa na cadeia de processamento das imagens dos conjuntos (*datasets*) é a de pré-processamento das mesmas. Para esta etapa, duas abordagens similares, porém com uma diferença sutil, foram selecionadas, ambas são baseadas na transformada de Wavelets da imagem, a diferença está no tipo dos filtros utilizados, onde a primeira abordagem (Haar) é baseada no uso de filtros digitais lineares passa-baixas e passa-altas nas imagens, onde com o passa-baixas é evidenciada a imagem de "aproximações", enquanto que com o passa-altas é evidencia a imagem de detalhes (destacando as bordas e maiores transições). A segunda abordagem é baseada no uso de filtros de morfologia matemática, realizando erosão e dilatação das imagens. Após testes para ajuste de parâmetros, em ambos os casos foram gerados dois níveis de decomposição Wavelet (fornecendo para cada imagem original, sete imagens decompostas, sendo três de detalhes para o primeiro nível, outras três para o segundo nível e uma de aproximações para o segundo nível). Para cada uma das imagens geradas pela transformada de Wavelets, será gerado um vetor de atributos (explicado mais a frente), todos esses sete vetores são concatenados para formar um único vetor que é o que contém todos os atributos que descrevem a imagem original.

A etapa de descrição de imagens em atributos talvez seja a etapa mais importante de todas em sistemas de visão computacional, esta é a etapa onde informação bruta na forma de dados será extraída da imagem pré-processada. É de fundamental importância que o algoritmo de descrição da imagem seja adequado ao ponto de extrair de forma sintetizada o máximo de informação possível da mesma. Ressalta-se a importância do poder de síntese do algoritmo de extração de atributos, pelo fato de que se procura uma redução da dimensionalidade do objeto a ser classificado, isto é, em outras palavras, deseja-se utilizar um vetor de muito menor dimensão para descrever todas as informações relevantes de uma imagem, possuindo uma semelhança "espacial" entre instâncias de mesma classe, ao passo que deve-se ter uma maior distância entre vetores de classes diferentes (JAIN; DUIN; MAO, 2000). Neste trabalho, foi utilizado um tipo de descritor de forma (como já mencionado, para explorar a tendência de forma de alguns tumores com a existência de câncer), onde este descritor é baseado nos polinômios de Zernike. Este descritor possui uma característica muito interessante e necessária para o tipo de classificação aqui proposto que é a invariância a rotações do objeto na imagem, também podem ser obtidas invariâncias em relação a escala e translações através de transformações geométricas. Os polinômios de Zernike formam uma base ortogonal entre si limitados no interior do círculo unitário. Os momentos de Zernike são as projeções da função que define o nível de intensidade da imagem sobre os polinômios ortogonais de Zernike. Para este trabalho, foram gerados trinta e dois momentos de Zernike conforme tabela 2 para cada uma das sete imagens geradas pela etapa de pré-processamento, resultando em um total de duzentos e vinte e quatro atributos no vetor que descreve a imagem original e é esse vetor que será trabalhado para a posterior etapa de redução de atributos proposta neste trabalho.

Tabela 2 – Combinações dos índices n e m dos polinômios de Zernike para geração dos momentos

n	m
3	1, 3
4	0, 2, 4
5	1, 3, 5
6	0, 2, 4, 6
7	1, 3, 5, 7
8	0, 2, 4, 6, 8
9	1, 3, 5, 7, 9
10	0, 2, 4, 6, 8, 10

Uma etapa de grande importância para melhorar o desempenho do classificador, mas que muitas vezes é negligenciada é a etapa de balanceamento. Um *dataset* desbalanceado pode acabar por viciar o classificador, pois tende a fornecer mais informação específica de uma ou mais classes, enquanto desfavorece outras, portanto uma maneira de se contornar este fato é a inclusão de "indivíduos sintéticos" no *dataset* com o objetivo de equilibrar essas populações. Algumas abordagens comuns de balanceamento são: Gerar um indivíduo sintético como um vetor médio de outros dois pré-existentes, ou realizar uma triangulação de múltiplos vetores, porém no caso aqui proposto, foi utilizada uma abordagem baseada nas estratégias evolucionárias de combinação e mutação utilizadas na técnica de Evolução Diferencial, conforme descrito pelo algoritmo 4:

Algoritmo 4 Algoritmo para geração de novos indivíduos sintéticos no *dataset*

1: **procedure** GeraNovoIndividuo($ConjuntoS$)
2: $i_1 \leftarrow GetIndividuoAleatorio(S)$
3: $i_2 \leftarrow GetIndividuoAleatorio(S)$
4: **if** $i_2 == i_1$ **then**
5: $GoTo : 3$
6: $i_3 \leftarrow GetIndividuoAleatorio(S)$
7: **if** $i_3 == i_1$ or $i_3 == i_2$ **then**
8: $GoTo : 6$
9: **for** $k \leftarrow (1$ até $i_1.tamanho)$ **do**
10: $i_4[k] \leftarrow i_1[k] + Rand(0, 2) \times (i_2[k] - i_3[k])$
11: **Retorna** i_4

Todas as etapas de leitura das imagens, pré-processamento, extração de atributos e balanceamento foram realizadas utilizando o software ConvIRMA.

5.2 Classificação

Dando sequência à cadeia, é a vez da etapa de classificação. Neste primeiro instante, os vetores ainda estão completos com todos os duzentos e vinte e quatro atributos e mais um para representar a classe ao qual o mesmo pertence. Para ser mais preciso, foi utilizado o algoritmo de otimização mínima sequencial (SMO - Sequential Minimal Optimization) de Jonh Platt (PLATT, 1998) para o treinamento da máquina de vetor de suporte (SVM - Support Vector Machine) com o kernel linear, daqui em diante este classificador será chamado apenas de SVM por simplicidade. Foram treinados oito classificadores diferentes, pois para cada um dos quatro rótulos de densidade mamária haviam duas configurações diferentes de pré-processamento Wavelet das imagens, onde para cada uma dessas abordagens foi treinado um classificador SVM diferente, consequentemente todas as métricas de classificação (taxas de acerto, matrizes de confusão, áreas abaixo da curva ROC, especificidades, sensibilidades, índice Kappa, dentre outros) são geradas individualmente para cada uma dessas abordagens.

Para todos os treinamentos, com e sem redução de atributos, foi utilizada a técnica de validação cruzada *10-fold*. Da mesma forma, todos os treinamentos foram repetidos trinta vezes devido ao caráter estocástico das etapas de treinamento (onde os pesos são inicializados aleatoriamente) e validação cruzada (onde cada fold é composto por indivíduos selecionados aleatoriamente).

5.3 Redução de Atributos

Quando se fala em sistemas de apoio ao diagnóstico, é razoável pensar na criticidade do desempenho. Não adianta ter uma ferramenta poderosa, mas que não é utilizada por ser bastante dispendiosa, principalmente quando se fala em bases de imagens cada vez maiores e etapas de classificação ou processamento cada vez mais complexas computacionalmente.

Tendo em vista o fato de que os atributos do vetor não contribuem igualmente para a descrição da imagem e principalmente o fato de que alguns até chegam a ser irrelevantes e não agregam informação para o discernimento por parte do classificador, bem como o fato de muitos também o serem redundantes, sem agregar informações nova, então uma etapa de seleção desses atributos se torna mandatória quando se deseja melhorar o desempenho

em utilização de recursos de memória e processamento computacional (menos atributos implica em uma arquitetura de classificador mais simples). Em alguns casos, a redução de atributos pode até ser positiva no quesito de acerto na classificação, tendo em vista que alguns atributos além de serem irrelevantes à tarefa de classificação, podem, mais do que isto, serem totalmente descorrelacionados e introduzirem "ruído" ao classificador, dificultando este processo e contribuindo negativamente para a etapa de classificação. Outra contribuição negativa do excesso de atributos seria uma maior chance de provocar *overfitting*, fazendo com que a rede "decore" casos mais específicos e perca um pouco a sua preciosa capacidade de extrapolação.

Neste trabalho foram utilizadas seis diferentes abordagens de seleção de atributos. A primeira é baseada no uso da técnica de ganho de informação, a segunda, por sua vez, é baseada na análise de componentes principais e as quatro seguintes são algoritmos de busca diferentes (algoritmos genéticos, otimização por enxames de partícula, busca evolucionária e best first) associados à técnica de avaliação de subconjuntos por seleção de atributos baseados em correlação (CFS Subset Evaluation) (HALL, 1999).

Para os algoritmos de busca mencionados, foram utilizados os seguintes valores para os parâmetros:

1. Algoritmos Genéticos
    a) Probabilidade de *Cross-Over*: 60%
    b) Gerações: 20
    c) Probabilidade de Mutações: 3,3%
    d) Tamanho da População: 20
2. Otimização por Enxames de Partículas
    a) Peso Individual: 0,34
    b) Peso Social: 0,33
    c) Peso Inercial: 0,33
    d) Iterações: 20
    e) Probabilidade de Mutações: 1%
    f) Tamanho da População: 20
3. Busca Evolucionária

a) Probabilidade de *Cross-Over*: 60%

b) Gerações: 20

c) Probabilidade de Mutação: 1%

d) Tamanho da População: 20

4. *Best First*

a) Máximo de *BackTrack*: 5

É importante destacar que o ajuste dos parâmetros das técnicas de busca é realizado de forma empírica, baseado na realização de diversas tentativas com diferentes valores de parâmetros para a escolha de um dado conjunto com melhores resultados e este fato reflete em uma limitação para o uso das referidas técnicas em situações de prática clínica, tendo em vista que o usuário do sistema seria um profissional da área médica e que não deveria estar preocupado com detalhes de funcionamento interno do sistema.

Após realizadas todas as classificações e reduções de atributos, os resultados são analisados estatisticamente, através da estatística descritiva, como no caso da observação das taxas de acerto médias e seus respectivos desvios padrões, observação de quantidade percentual de atributos reduzido, bem como uma análise de inferência estatística utilizando o teste de Wilcoxon para a comparação das métricas. A fim de chegar a uma conclusão de qual técnica ou abordagem é mais conveniente e eficiente para desempenhar o papel proposto para este trabalho, é utilizado uma métrica chamada de índice acerto-atributo descrito conforme equação 8. Este índice foi criado neste trabalho.

$$I_{aa} = \frac{1 - \frac{Q_r}{Q_t}}{T_t(\%) - T_r(\%)} \tag{8}$$

Onde $T_t(\%)$ representa a taxa de acerto da classificação sem redução de atributos, $T_r(\%)$ representa a taxa de acerto com redução de atributos e Q_r refere-se à quantidade de atributos selecionados pela técnica de redução de atributos e Q_t a quantidade absoluta total de atributos, sem redução.

Todas as etapas de treinamento do classificado, classificação e redução dos atributos foram realizados utilizando a biblioteca do software WEKA.

6 Resultados e Discussão

Após a execução de todas as etapas de simulações e experimentos computacionais conforme descrito na metodologia, foram obtidos dados diversos que são aqui explicitados e analisados. Primeiramente, uma das propostas a serem analisadas pelo trabalho em questão é a análise comparativa entre as técnicas de pré-processamento que são ambas baseadas em transformada de Wavelets das imagens, porém com a distinção no tipo de banco de filtros utilizados. A segunda proposta deste trabalho é fazer uma análise comparativa entre algumas técnicas de seleção de atributos, onde será observado o *trade-off* entre a redução na quantidade de atributos em contra-ponto com a redução nas taxas de acerto (que não são desejadas). Para tarefa de comparação final das técnicas de redução de atributos, é utilizado o índice chamado de razão acerto-atributo, que quantifica a redução de atributos realizada, penalizando a diminuição média da taxa de acerto provocada em relação à mesma quantidade sem a utilização da técnica.

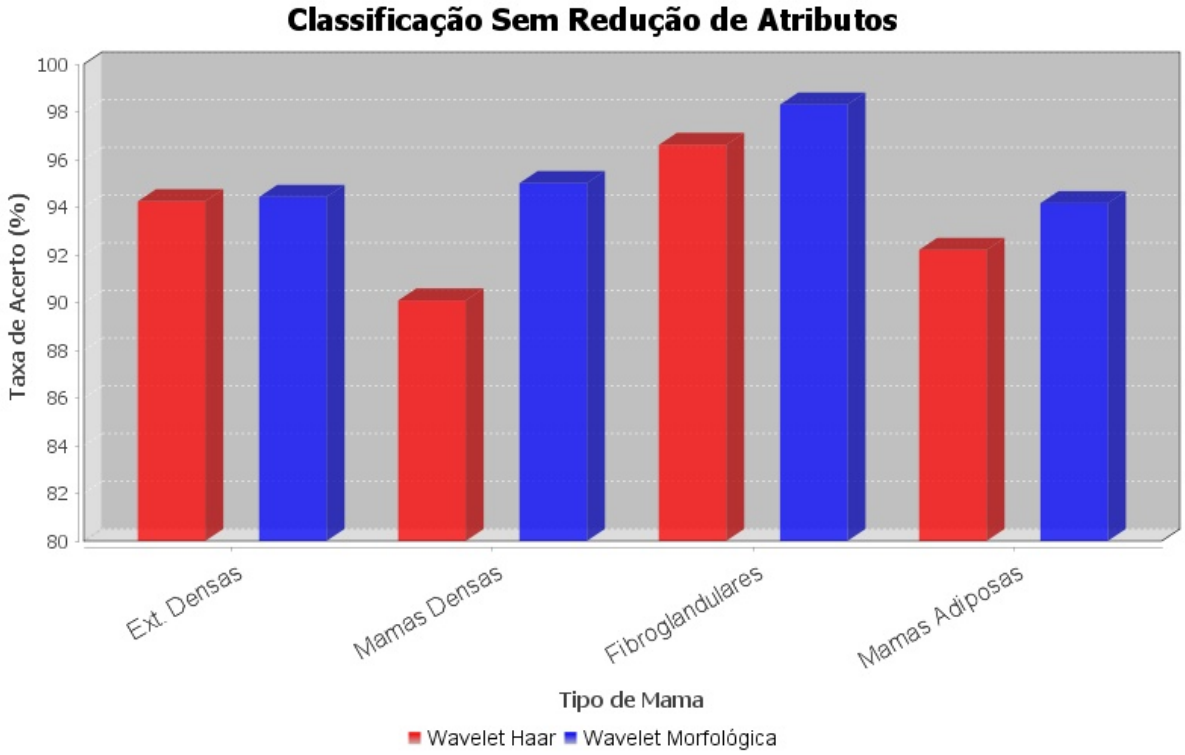
De início, os dados resultantes para a classificação das instâncias sem redução nos duzentos e vinte e quatro atributos do vetor de atributos está descrito conforme na tabela 3 e graficamente na figura 19.

Tabela 3 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias sem redução do vetor de atributos para as duas abordagens de pré-processamento

	Wavelet Haar		Wavelet Morf.	
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)
Ext. Densa	94,25	± 0,45	94,43	± 0,49
Densa	90,09	± 0,51	95,01	± 0,36
Fibroglandular	96,61	± 0,35	98,30	± 0,19
Adiposa	92,21	± 0,45	94,17	± 0,43

De princípio já é possível perceber uma tendência para a superioridade do método de pré-processamento baseado na transformada de Wavelets utilizando filtros de morfologia matemática. Porém esta tendência deve ser ainda confirmada com o uso de testes estatísticos tanto nestes casos, quantos nos casos de maior interesse que são os casos onde há redução

Figura 19 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações sem reduções de atributos, considerando ambos os casos de pré-processamento



de atributos. É possível notar na tabela 3 que para cada caso individual de tipo de mama, a abordagem com Wavelets morfológica obteve melhores resultados e com menores variações em torno do valor médio.

Realizando o teste estatístico de Wilcoxon para cada um dos pares de técnicas dos quatro diferentes tipos de mamas, são obtidos os dados presentes na tabela 4. Como pode-se notar, o *p-value* para os testes nos tipos de mama densa, fibroglandular e adiposa são virtualmente nulos, o que a um nível de significância de 5% representa que a Hipótese nula (ambos os conjuntos de amostras pertencem à mesma população) é rejeitada implicando na afirmação que a técnica de pré-processamento por transformada de Wavelets nesses casos foi superior à abordagem de Haar. Apenas no caso das mamas extremamente densas, que resultou em um *p-value* de aproximadamente 0,23, implica na aceitação da Hipótese nula ($\mathcal{H}_0$) para o mesmo nível de significância, levando à conclusão de que para este caso específico, ambas as abordagens são equivalentes.

Tabela 4 – Resultados para os testes estatísticos de Wilcoxon comparando os pares de diferentes técnicas de pré-processamento

Tipo de mama	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	0,2265	Não ●
Densa	$2,77 \times 10^{-11}$	Sim ●
Fibroglandular	$2,51 \times 10^{-11}$	Sim ●
Adiposa	$2,90 \times 10^{-11}$	Sim ●

6.1 Redução de Atributos

Em sequência, foram aplicadas as técnicas de redução de atributos, onde serão abordadas na ordem:

1. CFS com Algoritmos Genéticos
2. CFS com Otimização por Enxames de Partículas
3. CFS com Busca Evolucionária
4. CFS com *Best First*
5. Ganho de Informação com todas as suas variações de quantidades de atributos
6. Análise de Componentes Principais com todas as suas variações de quantidades de atributos

6.1.1 *CFS Subset Evaluation*

As quatro técnicas descritas a seguir foram utilizadas como suporte à técnica de avaliação de subconjuntos por seleção de atributos baseado na correlação (*CFS Subset Evaluation*), onde as mesmas são algoritmos de busca para a geração dos subconjuntos que serão avaliados pela técnica do algoritmo de seleção em questão.

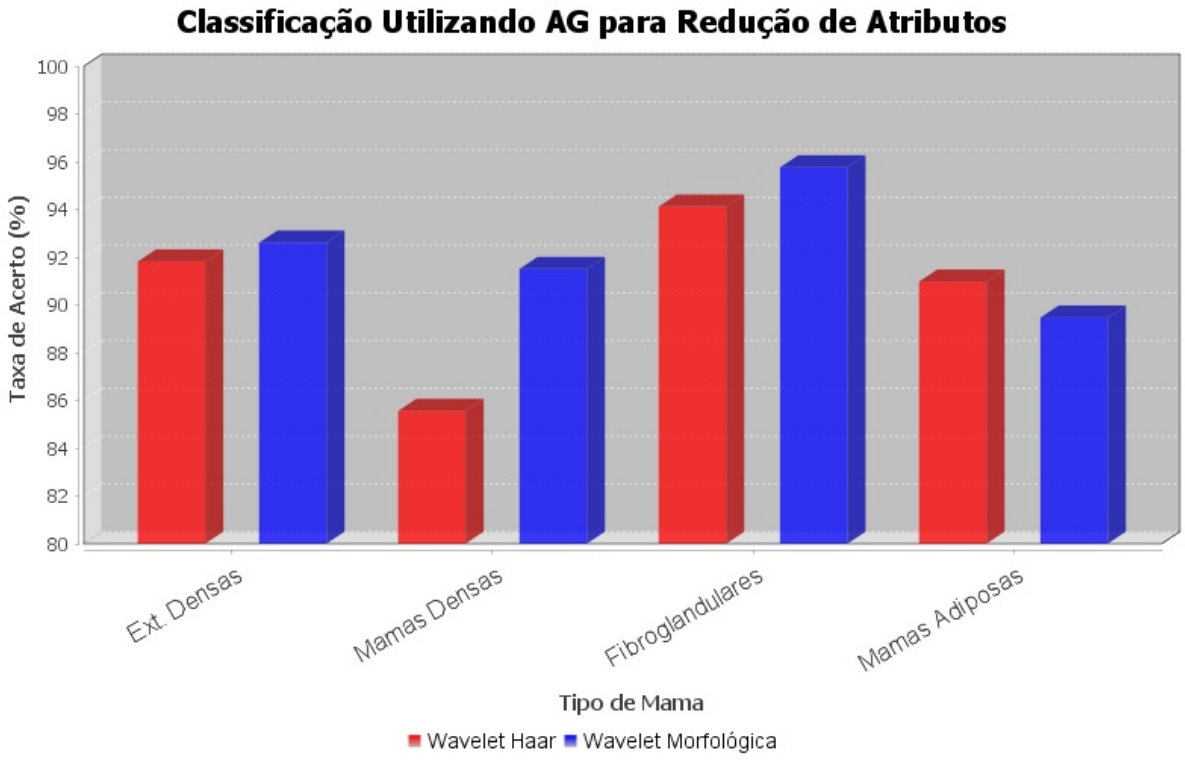
6.1.1.1 Algoritmos Genéticos

Quando executado o algoritmo CFS (*Correlation-based Feature Selection*) em conjunto com o Algoritmo Genético (utilizando os parâmetros definidos anteriormente) para a seleção dos atributos, foram obtidos os resultados presentes na tabela 5 e graficamente são exibidos os valores das taxas de classificação médias na figura 20.

Tabela 5 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de CFS com Algoritmos Genéticos para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	91,82	$\pm$ 1,02	92,61	$\pm$ 0,85	0,0026	Sim ●
Densa	85,56	$\pm$ 1,80	91,51	$\pm$ 1,23	$2,94 \times 10^{-11}$	Sim ●
Fibroglandular	94,12	$\pm$ 0,75	95,76	$\pm$ 0,80	$6,78 \times 10^{-8}$	Sim ●
Adiposa	90,98	$\pm$ 0,74	89,47	$\pm$ 1,32	$3,85 \times 10^{-6}$	Sim ●

Figura 20 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de CFS com Algoritmos Genéticos, considerando ambos os casos de pré-processamento



Percebe-se redução em todas as taxas de acerto médias, onde algumas foram mais penalizadas do que outras. Analisando os pares através do teste estatístico, obtemos os seguintes dados presentes também na tabela 5. Nota-se que em todos os quatro testes,

a hipótese nula foi rejeitada para uma nível de significância de 5%, o que implica que para os casos de mamas extremamente densas, mamas densas e mamas de composição fibroglandular, a transformada de Wavelets Morfológica obteve resultados melhores do que a Wavelets de Haar, ao passo que unicamente no caso de mamas adiposas, a situação foi inversa na superioridade de uma técnica sobre a outra.

Do ponto de vista da quantidade de atributos após a redução, tem-se o disposto na tabela 6, onde, de forma semelhante, estão dispostas as médias e desvios em pontos percentuais com relação ao conjunto completo (100% equivale a 224 atributos no vetor), especificados por tipo de mama e por tipo de pré-processamento. Nota-se que houve uma redução de mais da metade do conjunto para todas as abordagens, o que induz a pensar que mais da metade dos atributos do vetor completo é irrelevante, ou por ser completamente descorrelacionado com as classes, ou por serem redundantes dentro do vetor.

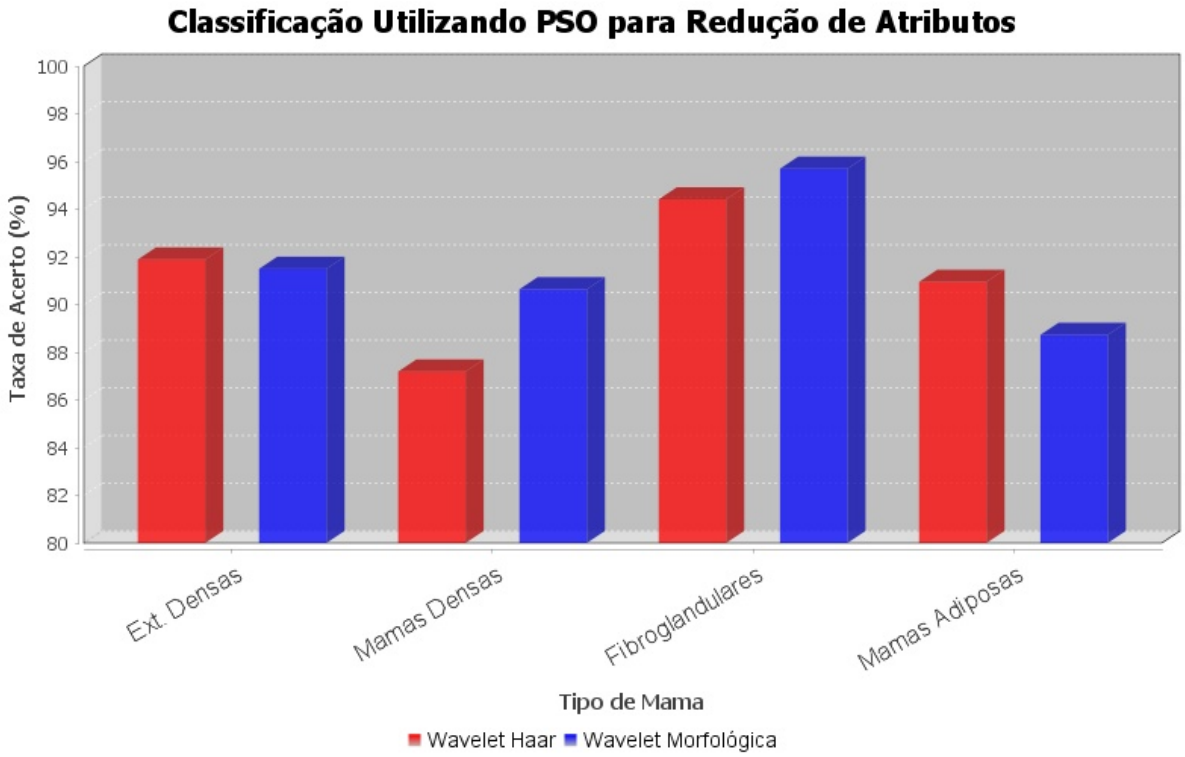
Tabela 6 – Média (μ) e desvio padrão (σ) do percentual da quantidade relativa de atributos selecionados para classificação das instâncias pelo uso da técnica CFS com Algoritmos Genéticos para as duas abordagens de pré-processamento

	Wavelet Haar		Wavelet Morf.	
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)
Ext. Densa	39,52	± 4,12	45,16	± 5,35
Densa	27,44	± 6,41	41,68	± 6,18
Fibroglandular	45,95	± 3,88	47,53	± 5,18
Adiposa	56,68	± 2,30	36,77	± 6,55

6.1.1.2 Otimização por Enxames de Partículas

Do mesmo modo que no caso anterior, os atributos foram selecionados e os vetores foram reduzidos, desta vez utilizando o algoritmo de CFS em conjunto com o algoritmo de Otimização por Enxames de Partículas. As taxas médias de acerto discriminadas por tipo de mama e tipo de pré-processamento estão demonstradas na figura 21 e as médias, desvios padrões e resultados dos testes estatísticos discriminados da mesma forma, estão descritos na tabela 7.

Figura 21 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de CFS com Otimização por Enxames de Partículas, considerando ambos os casos de pré-processamento



Pode ser notado que de forma semelhante ao caso com algoritmos genéticos, ocorreram diminuições nas taxas de acerto, principalmente para os tipos de mama densas e nas mamas adiposas, onde a técnica de transformada de Wavelets utilizando filtros morfológicos foi mais afetada.

Estatisticamente, para o caso de mamas extremamente densas, ambos os casos de pré-processamento foram equivalentes, diferentemente dos outros tipos de mama, onde para o caso de mamas densas e fibroglandulares, a transformada morfológica foi superior, contrariamente ao caso das mamas adiposas, onde, neste último caso, a técnica de redução de atributos atenuou bastante a taxa de acerto desta abordagem de pré-processamento.

Com relação à quantidade média da redução dos atributos, tem-se conforme tabela 8. Observações semelhantes ao caso os algoritmos genéticos podem ser feitas.

Tabela 7 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de CFS com Otimização por Enxames de Partículas para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	91,89	± 0,67	91,50	± 1,05	0,1425	Não ●
Densa	87,19	± 1,26	90,64	± 1,31	$2,94 \times 10^{-10}$	Sim ●
Fibroglandular	94,41	± 0,64	95,70	± 0,94	$4,37 \times 10^{-7}$	Sim ●
Adiposa	90,95	± 0,75	88,72	± 1,43	$3,54 \times 10^{-9}$	Sim ●

Tabela 8 – Média (μ) e desvio padrão (σ) do percentual da quantidade relativa de atributos selecionados para classificação das instâncias pelo uso da técnica CFS com Otimização por Enxames de Partículas para as duas abordagens de pré-processamento

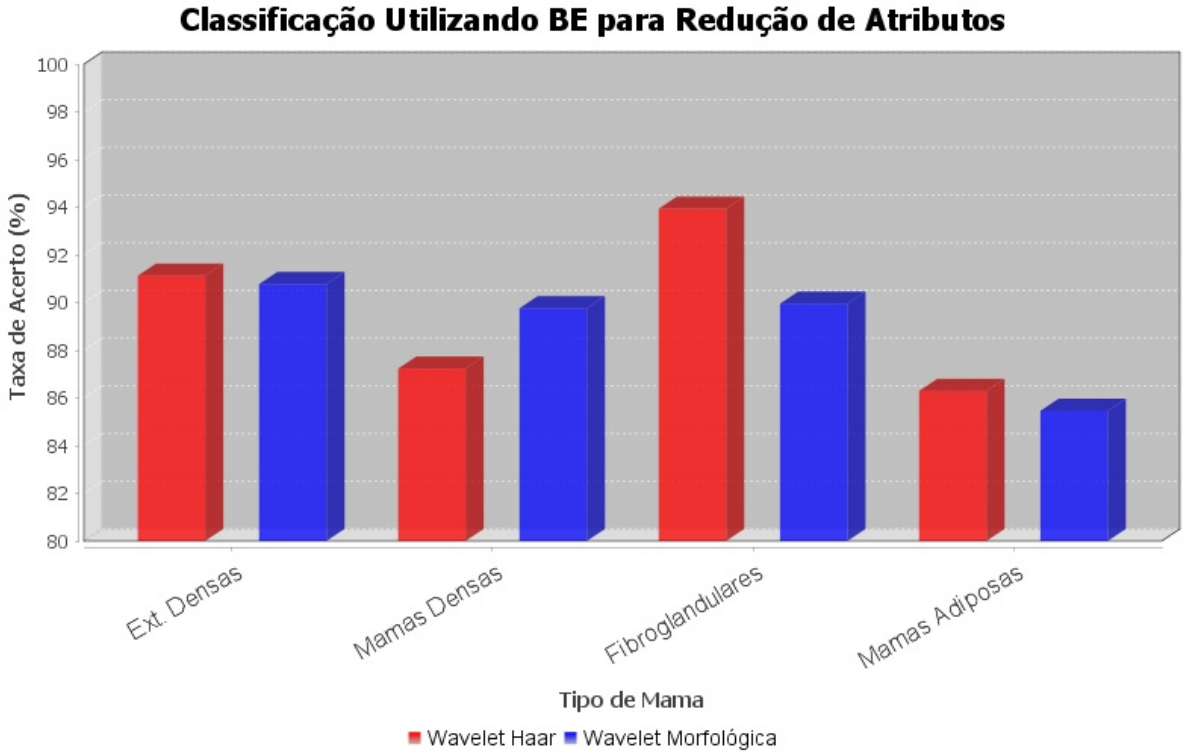
	Wavelet Haar		Wavelet Morf.	
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)
Ext. Densa	37,87	± 4,65	41,83	± 4,76
Densa	28,70	± 4,09	38,07	± 5,58
Fibroglandular	43,53	± 4,87	46,29	± 4,02
Adiposa	56,52	± 4,22	31,70	± 5,03

6.1.1.3 Busca Evolucionária

A terceira técnica baseada em heurística de busca utilizada neste trabalho foi a busca evolucionária, em conjunto com o algoritmo de CFS. De forma idêntica ao protocole realizado para os dois casos anteriores, tem-se na figura 22 as médias das taxas de acerto para a classificação das instâncias após a redução do vetor de atributos a partir da seleção realizada pela técnica em questão.

Na tabela 9 estão expostas as estatísticas descritivas para as taxas de acerto, bem como os resultados dos testes estatísticos realizados para comparação de pares com o objetivo de comparar as técnicas de pré-processamento.

Figura 22 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de CFS com Busca Evolucionária, considerando ambos os casos de pré-processamento



Percebe-se, no contexto geral, um desempenho inferior em taxas de acerto desta abordagem, quando comparado às técnicas de redução de atributos anteriores, de ante-mão já é possível imaginar que esta técnica não seria uma boa candidata a ser selecionada como a melhor, tendo em vista que mesmo que, hipoteticamente, esta técnica seja mais robusta para a redução da quantidade de atributos, existe um limite para o ganho de processamento computacional em detrimento da capacidade de acerto da rede. Em outras palavras, a acurácia do sistema é o fator crítico mais importante em análise.

Conforme tabela 10, observa-se a redução média percentual da quantidade de atributos nos vetores para os diferentes casos.

Nota-se que realmente houve uma redução de atributos maior no geral do que ambas as técnicas anteriores e que em quatro casos houve uma estabilização da quantidade de atributos ($\sigma = 0$), isto é, o algoritmo de busca em questão convergiu para um valor fixo em todas as repetições, porém, em comparação com as técnicas anteriores, esses foram os de menor capacidade de classificação. É possível observa, também, uma tendência onde as abordagens baseadas em transformada morfológica são mais sensíveis às reduções de

Tabela 9 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de CFS com Busca Evolucionária para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	91,13	± 1,47	90,78	± 0,96	0,056	Não ●
Densa	87,23	± 0,69	89,76	± 0,44	$2,97 \times 10^{-11}$	Sim ●
Fibroglandular	93,94	± 0,62	89,95	± 0,41	$2,81 \times 10^{-11}$	Sim ●
Adiposa	86,29	± 0,63	85,46	± 0,65	$1,53 \times 10^{-5}$	Sim ●

atributos, ou pelo menos a redução de atributos abaixo de algum limiar, induzindo a possibilidade de que essa abordagem gera uma quantidade maior de atributos relevantes (ou no mínimo não redundantes) para a classificação e que em algumas situações das técnicas até agora utilizadas não foram capazes de identificar esta nuance.

Tabela 10 – Média (μ) e desvio padrão (σ) do percentual da quantidade relativa de atributos selecionados para classificação das instâncias pelo uso da técnica CFS com Busca Evolucionária para as duas abordagens de pré-processamento

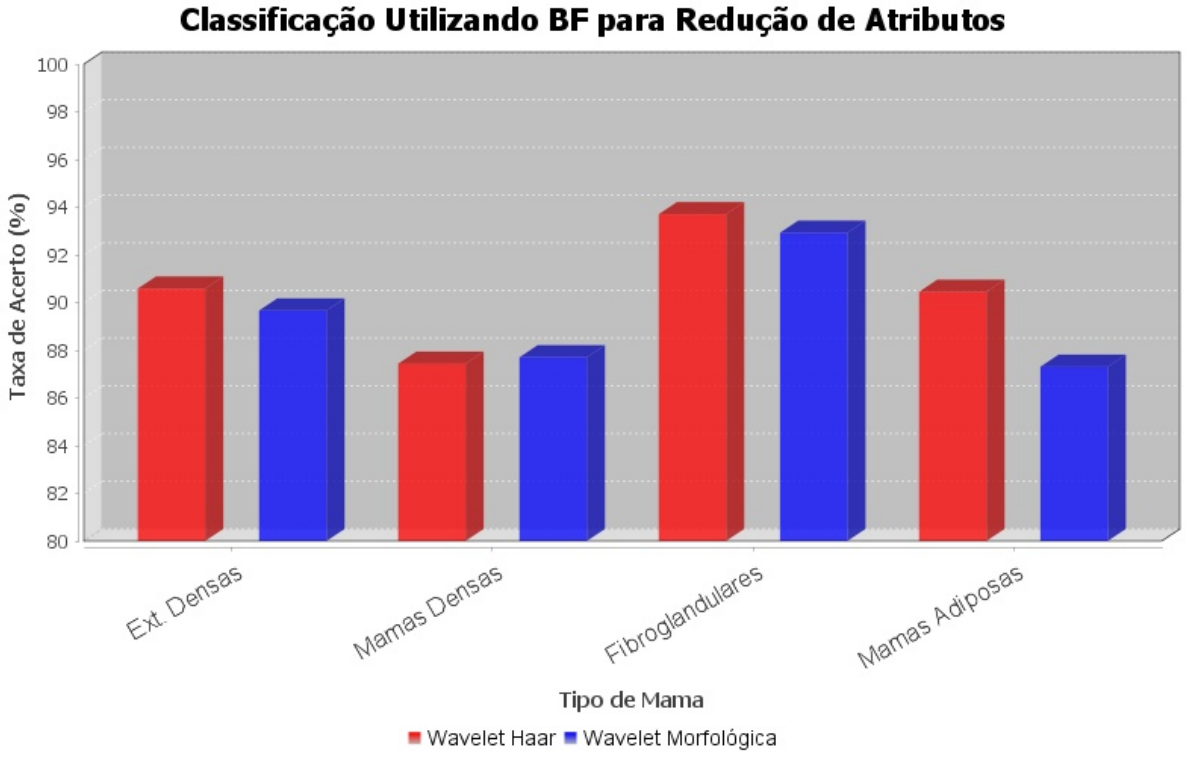
	Wavelet Haar		Wavelet Morf.	
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)
Ext. Densa	38,57	± 3,11	35,94	± 2,51
Densa	33,93	± 0,00	33,93	± 0,00
Fibroglandular	40,13	± 3,03	28,57	± 0,00
Adiposa	28,57	± 0,00	28,38	± 1,06

6.1.1.4 *Best First*

Esta é única técnica, dentre as quatro utilizadas em conjunto com a CFS, que possui uma característica não estocástica. O que significa que para todas as repetições sempre serão utilizadas as mesmas quantidades de atributos no vetor, cabendo apenas às

etapas de classificação e validação o caráter não determinístico das taxas de acerto, que estão representadas pelos seu valores médios na figura 23.

Figura 23 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de CFS com *Best First*, considerando ambos os casos de pré-processamento



Na tabela 11 são observados todos os dados estatísticos descritivos das taxas de acerto e dos testes de Wilcoxon para esta técnica.

Com relação à superioridade das técnicas de pré-processamento, para os casos de mama extremamente densa, mama de predominância fibroglandular e mamas adiposas, a técnica de transformada de Haar se mostrou superior, ao passo que no caso de mamas densas não houve significância estatística suficiente e seus resultado são considerados equivalentes.

Por fim, como mencionado, esta técnica resultou na mesma quantidade de atributos selecionados para todas as repetições, o que pode ser confirmado na tabela 12. Observa-se nesta tabela que esta técnica, de forma semelhante à baseada em busca evolucionária, selecionou menos de um terço dos atributos (salvo o caso para mamas adiposas com Wavelets de Haar) e que para todos os casos, foram selecionados menos atributos para as instâncias pré-processadas com Wavelets morfológica do que para sua contrapartida em

Tabela 11 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de CFS com *Best First* para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	90,59	± 0,49	89,67	± 0,53	$1,37 \times 10^{-7}$	Sim ●
Densa	87,44	± 0,67	87,71	± 0,39	0,1126	Não ●
Fibroglandular	93,71	± 0,50	92,93	± 0,44	$3,35 \times 10^{-7}$	Sim ●
Adiposa	90,45	± 0,62	87,31	± 0,48	$2,87 \times 10^{-11}$	Sim ●

Haar. Isto acabou resultando em uma visível queda nas taxas de acerto, principalmente nos casos com maiores reduções de atributos, onde o melhor resultado está para os tipos de mama fibroglandulares.

Tabela 12 – Média (μ) e desvio padrão (σ) do percentual da quantidade relativa de atributos selecionados para classificação das instâncias pelo uso da técnica CFS com *Best First* para as duas abordagens de pré-processamento

	Wavelet Haar		Wavelet Morf.	
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)
Ext. Densa	35,27	± 0,00	32,14	± 0,00
Densa	26,34	± 0,00	24,11	± 0,00
Fibroglandular	35,71	± 0,00	27,23	± 0,00
Adiposa	46,43	± 0,00	20,98	± 0,00

6.1.2 *Ranker*

As técnicas exploradas a seguir, diferentemente das anteriores, não são utilizadas em conjunto com o algoritmo de CFS (*Correlation-based Feature Selection*), também não são algoritmos de busca. O seu funcionamento é baseado na enumeração (*Ranking*) em ordem decrescente dos atributos que são mais relevantes para a classificação de acordo com a avaliação própria dessas técnicas. Sendo o ganho de informação baseado na comparação

de como o nível de entropia do conjunto se comporta após a categorização por cada atributo individualmente, onde os atributos que promoverem os maiores ganhos neste quesito (são consideradas como as variáveis que melhor distribuem as classes) são enumeradas no topo da classificação. No caso da análise de componentes principais, são procuradas as combinações lineares que descrevem o subconjunto, isto é, levando-se em consideração as interdependências entre as próprias variáveis e a contribuição para as classes.

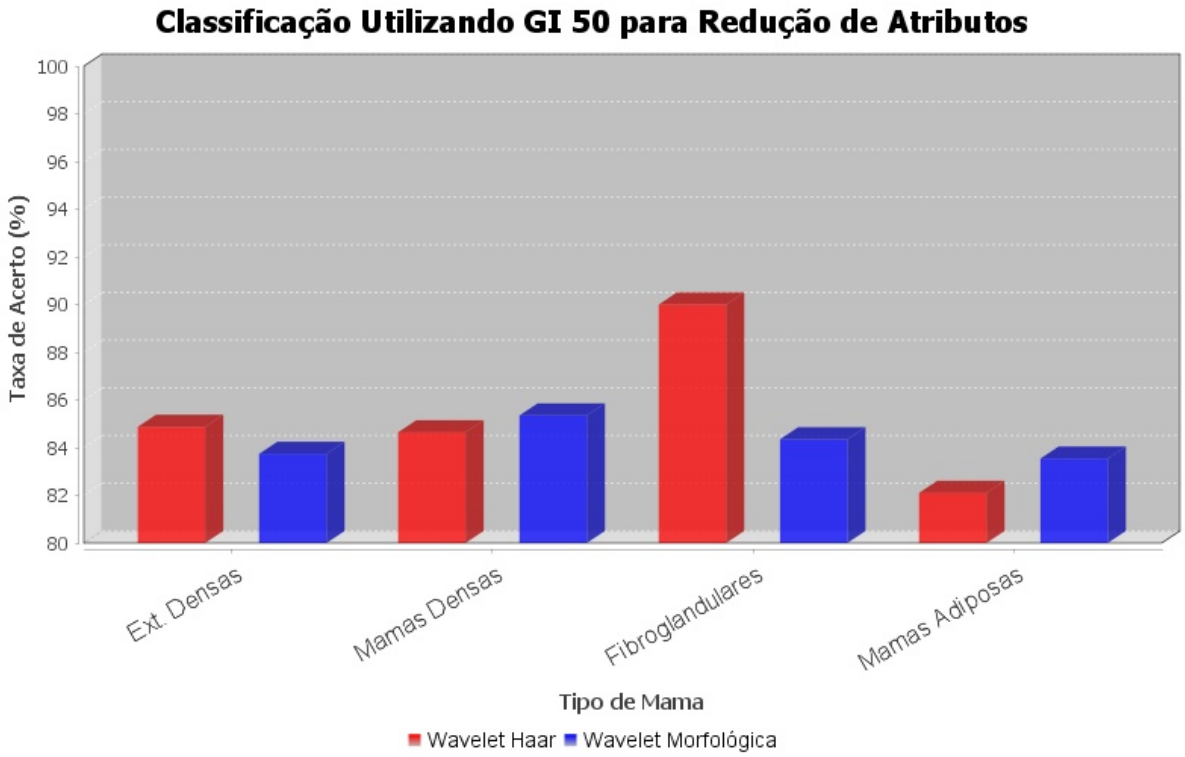
Para estas duas abordagens, foram geradas as enumerações das variáveis em ordem de relevância e para cada uma delas, foram selecionadas as cinquenta, sessenta, setenta, oitenta e noventa atributos mais relevantes e depois realizadas as análises semelhantes às que já foram efetuadas para as técnicas anteriores. Esses valores de atributos foram selecionados manualmente para serem equiparados às técnicas anteriores, porém poderiam ter sido selecionados automaticamente através de alguma métrica como a de significância ou limiares específicos como no caso do ganho de informação, dentre outras.

6.1.2.1 Ganho de Informação

Para o caso do uso da técnica de ganho de informação, onde foram selecionados os cinquenta atributos mais relevantes, foram obtidos os resultados de taxas de acerto para as classificações na figura 24. Observa-se uma queda considerável em todas as taxas de acerto, onde já pode ser evidenciado uma das limitações desta técnica que é apenas considerar a contribuição individual de um atributo para a classificação, deixando de levar em conta as relações inter-atributos, isto é, os que forem selecionados por esta técnica podem até serem relevantes no contexto geral, porém muitos deles possuem alto grau de redundância entre si. Percebe-se que apenas o caso de mamas fibroglandulares com a transformada Haar obtiveram taxas razoáveis.

Estatisticamente, esta abordagem descreve-se conforme tabela 13. Onde pode ser observado que para todas os testes houve a rejeição da Hipótese nula ($\mathcal{H}_0$), de forma que para os tipos de mamas densas, e mamas adiposa, houve uma superioridade da transformada de Wavelets morfológica e para os outros tipos de mama, a situação foi inversa.

Figura 24 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de Ganho de Informação com 50 atributos, considerando ambos os casos de pré-processamento



Com relação à quantidade de atributos, foram selecionados os cinquenta mais relevantes segundo esta técnica, o que representa 22,32% do total sem reduções. Para o caso seguinte foram selecionados sessenta, o que representa um percentual de 26,79%, na figura 25 estão expressos os valores médios de taxas de acerto. Pode ser observada uma óbvia ascensão de todas as taxas de acerto médias em comparação com a abordagem anterior, porém de uma maneira geral, esta não aparenta ser ainda uma boa abordagem.

Na tabela 14 são observadas as análises estatísticas para este caso e de forma muito semelhante as mesmas observações do caso anterior podem ser feitas.

Em seguida, selecionando-se os 70 atributos (equivalente a 31,25% do total) mais relevantes a partir do uso do Ganho de Informação como métrica, temos o exposto na figura 26. É observada a mesma tendência de crescimento geral das taxas de acerto, porém no caso de mamas fibroglandulares na abordagem de Haar é evidenciado uma redução da taxa média de acerto em comparação à anterior. Isso demonstra que uma quantidade grande de atributos, além de ser ineficiente computacionalmente por requerer um sistema

Tabela 13 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de Ganho de Informação com 50 atributos para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	85,87	± 0,43	83,74	± 0,54	$1,23 \times 10^{-9}$	Sim ●
Densa	84,64	± 0,57	85,35	± 0,46	$1,06 \times 10^{-5}$	Sim ●
Fibroglandular	90,00	± 0,41	84,35	± 0,40	$2,73 \times 10^{-11}$	Sim ●
Adiposa	82,10	± 0,64	83,54	± 0,35	$9,09 \times 10^{-11}$	Sim ●

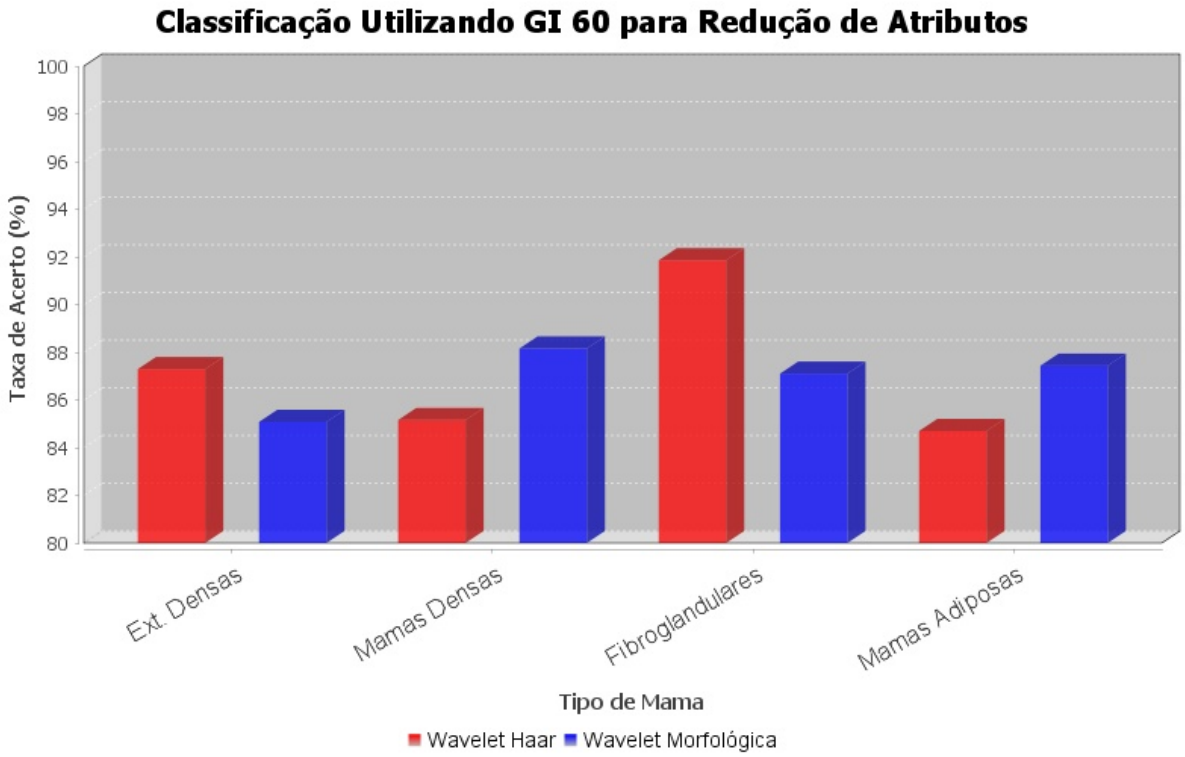
Tabela 14 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de Ganho de Informação com 60 atributos para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	87,29	± 0,49	85,08	± 0,45	$3,01 \times 10^{-11}$	Sim ●
Densa	85,16	± 0,42	88,16	± 0,35	$2,65 \times 10^{-11}$	Sim ●
Fibroglandular	91,85	± 0,35	87,10	± 0,49	$2,75 \times 10^{-11}$	Sim ●
Adiposa	84,70	± 0,41	87,44	± 0,47	$2,78 \times 10^{-11}$	Sim ●

de classificação maior, pode piorar também o desempenho do sistema, induzindo por exemplo efeitos de *over-fitting* ou podem trazer informação não relevante para o contexto que acaba reduzindo a capacidade de extrapolação para instâncias que não pertenceram ao conjunto de treinamento, sendo esses alguns dos problemas a serem combatidos com a seleção de atributos.

Na tabela 15 são analisados estatisticamente esses dados, de maneira que para os tipos de mama extremamente densas, as duas abordagens de Wavelets foram equivalentes, para as mamas densas e adiposas, a transformada morfológica se saiu melhor e o oposto ocorreu para as fibroglandulares.

Figura 25 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de Ganho de Informação com 60 atributos, considerando ambos os casos de pré-processamento



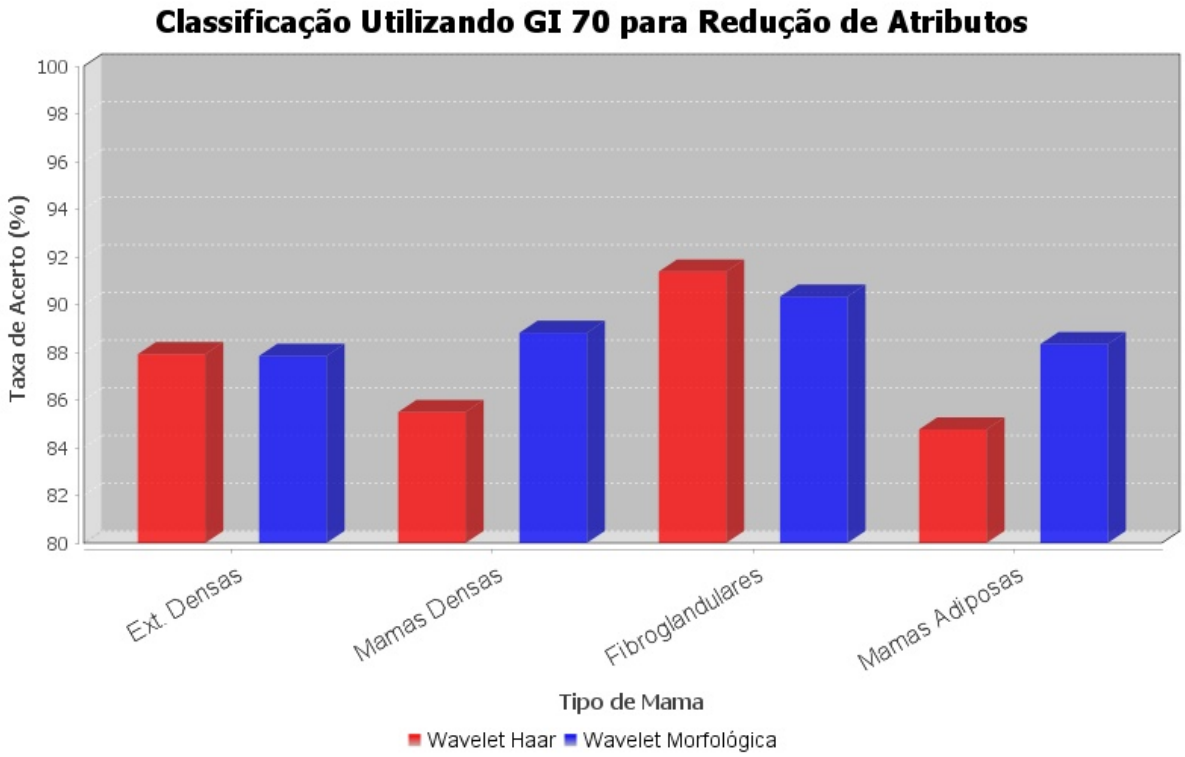
Quando selecionados 80 atributos (35,7%) são obtidas as taxas presentes na figura 27. Novamente, um acréscimo geral nas taxas de acerto, porém esta técnica não demonstra ser robusta o suficiente.

Na tabela 16 analisa-se os dados das taxas de acerto estatisticamente, com um comportamento muito similar ao anterior.

Por último, para esta técnica, são analisados os resultados com a seleção de noventa atributos. O que reflete num percentual de 40,18% do total de atributos. Como esperado, a maioria das taxas de acerto médias cresceu, apenas o caso de mamas densas com a Transformada Haar obteve o oposto desta tendência. Esses resultados já começam a se mostrar um pouco mais satisfatórios, cabendo uma análise comparativa mais geral entre as diferentes técnicas.

Na tabela 17 são expostos os dados estatísticos a respeito dessas taxas de acerto, onde são notados que para o caso de mamas fibroglandulares não houve distinção de

Figura 26 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de Ganho de Informação com 70 atributos, considerando ambos os casos de pré-processamento



técnica, ao passo que para mamas densas e adiposas, a abordagem morfológica foi superior, sendo inferior no caso de mamas extremamente densas.

6.1.2.2 Análise de Componentes Principais

Com relação à técnica de Análise de Componentes Principais, são efetuadas as mesmas análises e observações para cinquenta, sessenta, setenta, oitenta e noventa atributos selecionados, porém com uma diferença a ser considerada. Esta técnica, por efetuar uma combinação linear dos atributos originais, pode resultar em um novo conjunto de atributos que seja numericamente inferior ao original (isto por si só já poderia ser um efeito explorado na redução de atributos) e em algumas vezes, o que pode ocorrer é que o conjunto de componentes principais (o novo vetor de atributos) tenha um tamanho inferior ao requisitado e, com isso, quando for realizada uma seleção de atributos, se a quantidade existente for inferior ao desejado, será utilizada esta quantidade máxima.

Tabela 15 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de Ganho de Informação com 70 atributos para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	87,92	± 0,65	87,85	± 0,41	0,3165	Não ●
Densa	85,49	± 0,44	88,82	± 0,43	$2,69 \times 10^{-11}$	Sim ●
Fibroglandular	91,38	± 0,38	90,33	± 0,33	$9,44 \times 10^{-11}$	Sim ●
Adiposa	84,76	± 0,59	88,34	± 0,57	$2,79 \times 10^{-11}$	Sim ●

Tabela 16 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de Ganho de Informação com 80 atributos para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	89,28	± 0,48	89,39	± 0,62	0,4353	Não ●
Densa	86,76	± 0,58	90,73	± 0,39	$2,81 \times 10^{-11}$	Sim ●
Fibroglandular	91,88	± 0,35	90,92	± 0,44	$1,01 \times 10^{-9}$	Sim ●
Adiposa	86,86	± 0,67	88,99	± 0,58	$1,50 \times 10^{-10}$	Sim ●

Iniciando para o caso de cinquenta componentes principais, temos as médias das taxas de acerto conforme a figura 29. Inicialmente já é possível perceber a superioridade desta técnica quando comparada, por exemplo, com a técnica de Ganho de Informação também com cinquenta atributos. Isto pode ser evidenciado pelo fato de esta técnica fazer uso de uma análise não apenas individual, mas também das dependências entre atributos.

Na tabela 18 são vistos os dados estatísticos para essas taxas de acerto. Para esta abordagem, é observada a superioridade da transformada de Wavelets morfológica em quase todos os casos, salvo no caso de mamas fibroglandulares onde as técnicas foram equivalentes.

Figura 27 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de Ganho de Informação com 80 atributos, considerando ambos os casos de pré-processamento

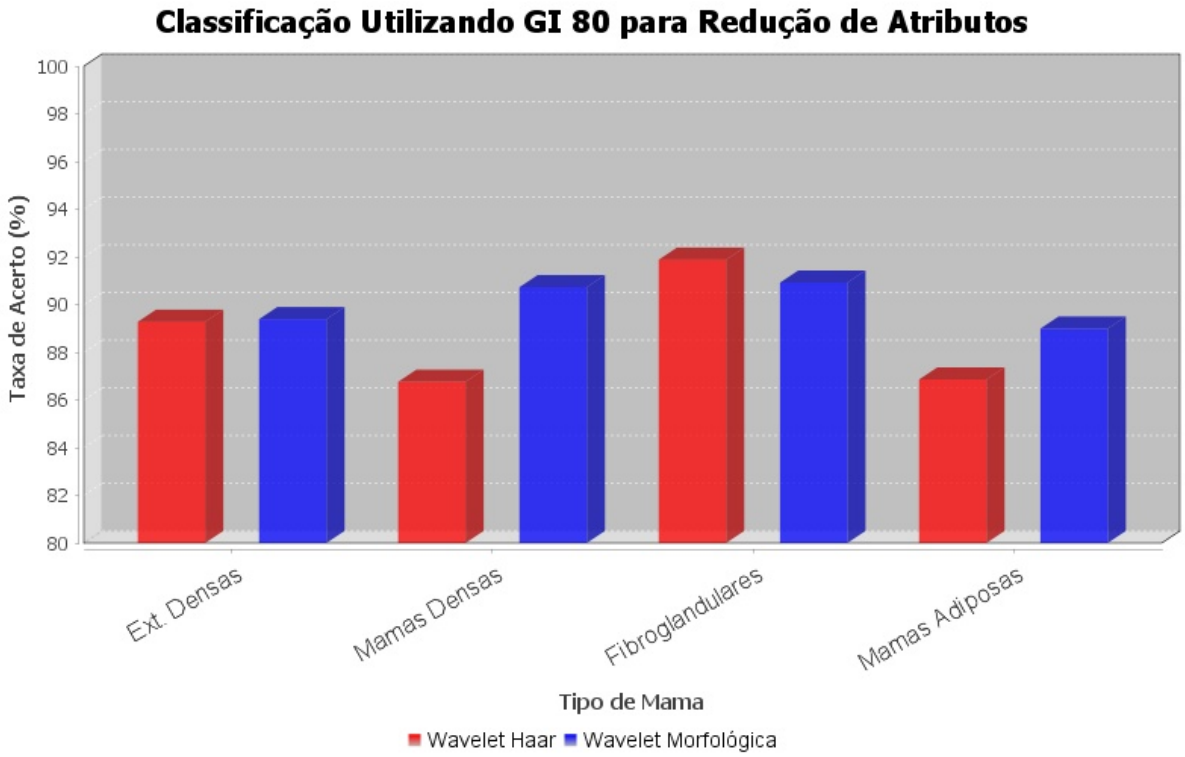
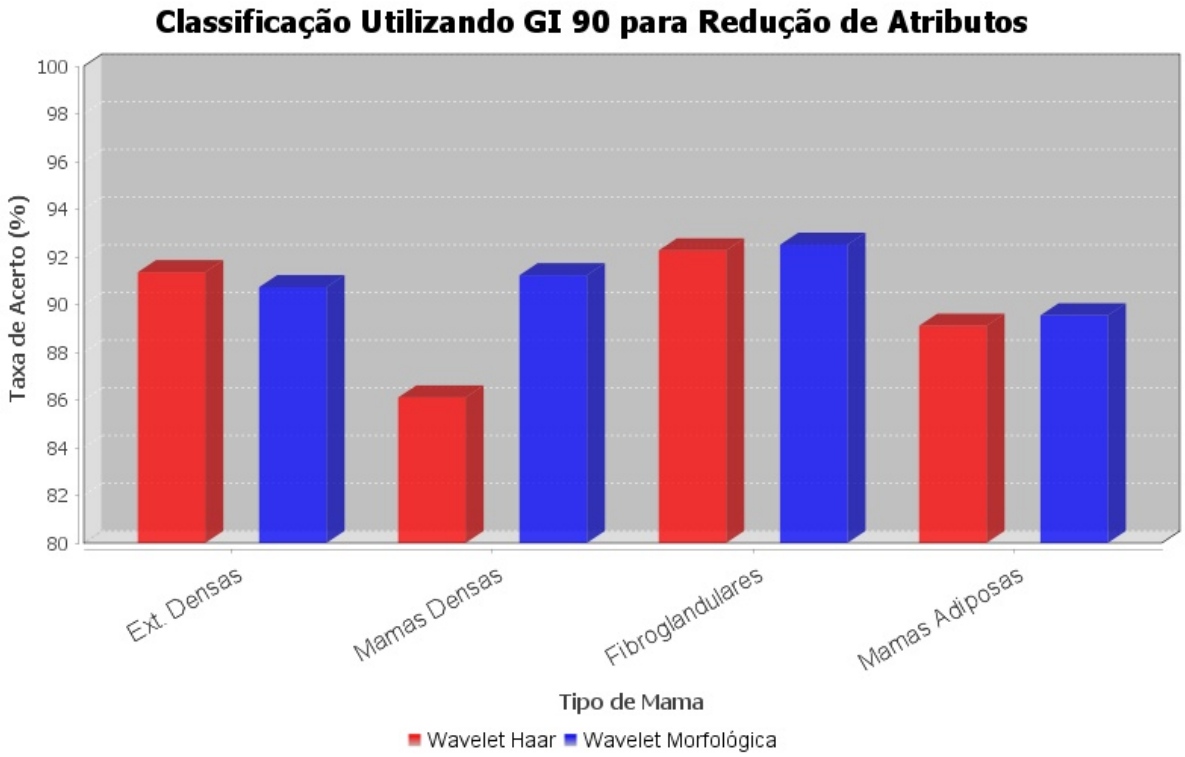


Tabela 17 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de Ganho de Informação com 90 atributos para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	91,36	± 0,38	90,73	± 0,39	$7,32 \times 10^{-7}$	Sim ●
Densa	86,10	± 0,71	91,23	± 0,48	$2,82 \times 10^{-11}$	Sim ●
Fibroglandular	92,27	± 0,51	92,51	± 0,46	0,073	Não ●
Adiposa	89,11	± 0,54	89,55	± 0,42	$6,91 \times 10^{-4}$	Sim ●

Para sessenta atributos selecionados, é tido de acordo com a figura 30. Através desta imagem, pode-se perceber um aumento das taxas de acerto médias para as quatro abordagens com utilização de transformada morfológica, ao passo que para a transformada Haar houve o aumento das taxas de acerto para dois tipos de mama, porém com a redução de outros dois.

Figura 28 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de Ganho de Informação com 90 atributos, considerando ambos os casos de pré-processamento

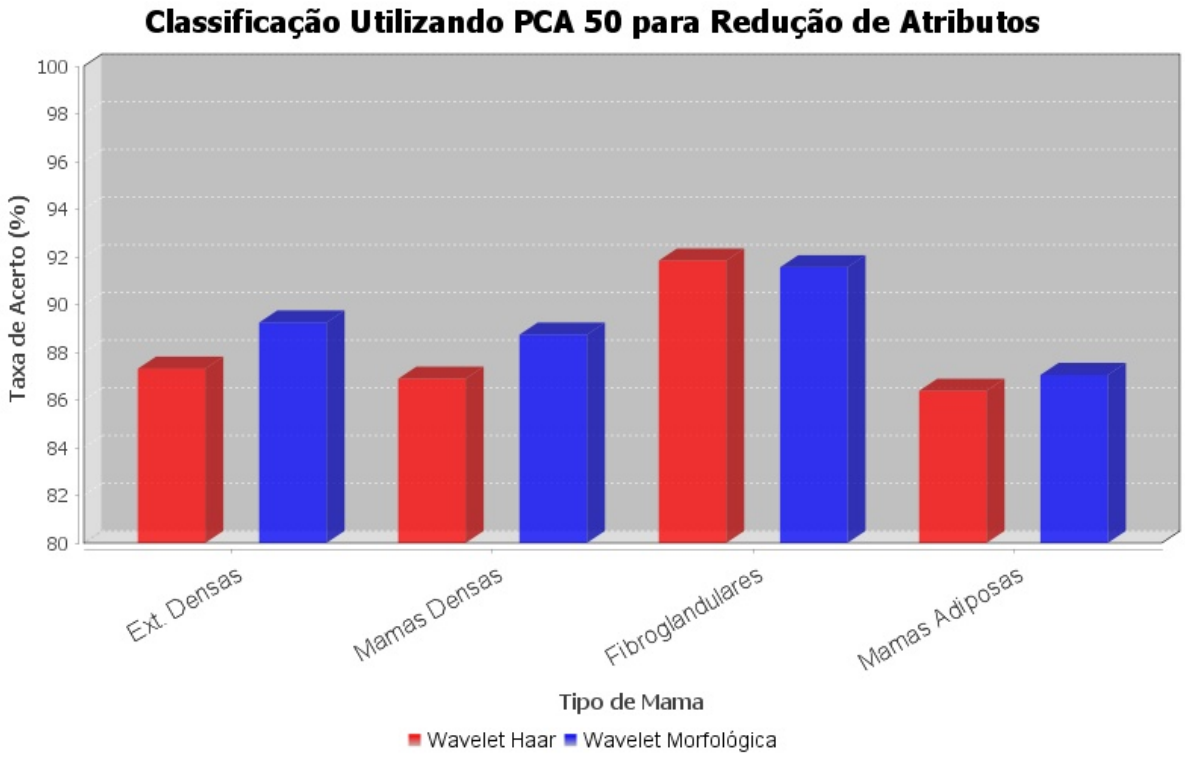


Na tabela 19 são vistos com mais detalhes os dados estatísticos para essas taxas de acerto. Nesta abordagem, é observada a superioridade da transformada de Wavelets morfológica em todos os casos. E, novamente esta abordagem se mostra bastante superior à sua contrapartida com Ganho de Informação.

Com a seleção de setenta atributos, são obtidas as taxas de acerto médias conforme figura 31, onde é observado um aumento geral das taxas de acerto, conforme é de se esperar.

Na tabela 20 são exibidos com mais detalhes os dados estatísticos referentes a essas taxas de acerto. Nesta abordagem, é observada, mais uma vez, a superioridade da transformada de Wavelets morfológica em todos os casos. E, novamente esta abordagem se mostra bastante superior à sua contrapartida com Ganho de Informação. De uma maneira geral, a técnica de redução de atributos por análise de componentes principais apresenta a tendência de despontar como uma forte candidata à melhor técnica para seleção de atributos estudadas neste trabalho.

Figura 29 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de Análise de Componentes Principais com 50 atributos, considerando ambos os casos de pré-processamento



Em sequência, selecionando-se oitenta atributos, são obtidas as taxas de acerto médias conforme figura 32, onde é observado um aumento geral médio das taxas de acerto um pouco mais singelo.

Através da tabela 21 podem ser observados com mais detalhes os dados estatísticos referentes a essas taxas médias de acerto, de forma que para mamas extremamente densas, ambas as abordagens de transformada são equivalentes, porém nos tipos restantes de mama, a abordagem morfológica se mostra superior.

É importante notar que para alguns dos tipos analisados nesta etapa a quantidade de componentes principais obtidas pela técnica foi inferior a oitenta, como está demonstrado na tabela 22. Nos tipos de mama adiposas com uso de transformada morfológica existem um total máximo de setenta atributos selecionados, bem como no caso de mamas fibroglandulares para o mesmo tipo de pré-processamento, há um total de setenta e nove componentes principais.

Tabela 18 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de Análise de Componentes Principais com 50 atributos para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	87,31	± 0,56	89,25	± 0,52	$4,46 \times 10^{-11}$	Sim ●
Densa	86,89	± 0,54	88,74	± 0,61	$5,82 \times 10^{-11}$	Sim ●
Fibroglandular	91,48	± 0,52	91,56	± 0,55	0,4355	Não ●
Adiposa	86,38	± 0,59	87,05	± 0,45	$2,75 \times 10^{-5}$	Sim ●

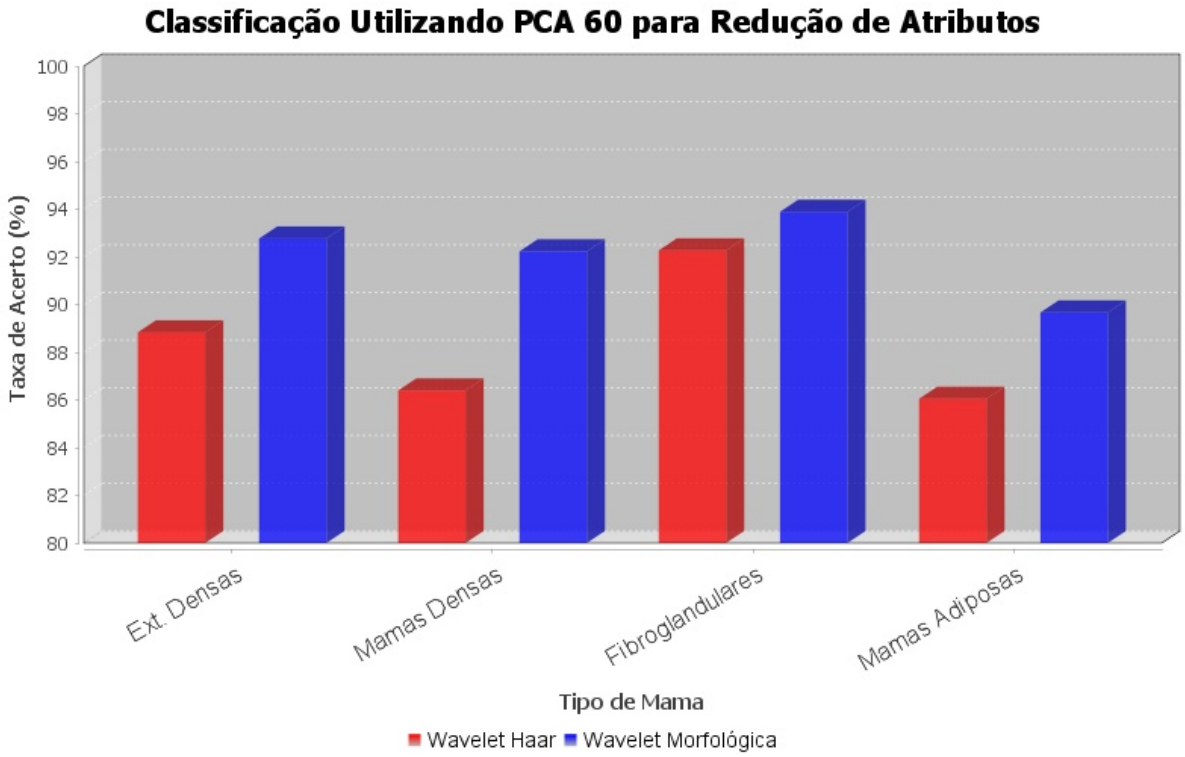
Tabela 19 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de Análise de Componentes Principais com 60 atributos para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	88,83	± 0,49	92,77	± 0,44	$2,81 \times 10^{-11}$	Sim ●
Densa	86,40	± 0,59	92,23	± 0,47	$2,80 \times 10^{-11}$	Sim ●
Fibroglandular	92,27	± 0,48	93,89	± 0,39	$3,42 \times 10^{-11}$	Sim ●
Adiposa	86,05	± 0,51	89,67	± 0,58	$2,87 \times 10^{-11}$	Sim ●

Por fim, para o caso de até noventa atributos selecionados pela técnica de análise de componentes principais, tem-se as taxas de acerto médias conforme figura 33. Pode ser observado que ocorreram apenas modificações singelas, onde apenas houve uma redução um pouco mais aparente nas taxas de acerto para mamas densas com uso da transformada de Haar.

Através da tabela 23 podem ser observados com mais detalhes os dados estatísticos referentes a essas taxas médias de acerto, de forma que para mamas extremamente densas, a abordagem de Haar é mais eficiente, porém nos tipos restantes de mama, a abordagem morfológica se mostra superior.

Figura 30 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de Análise de Componentes Principais com 60 atributos, considerando ambos os casos de pré-processamento

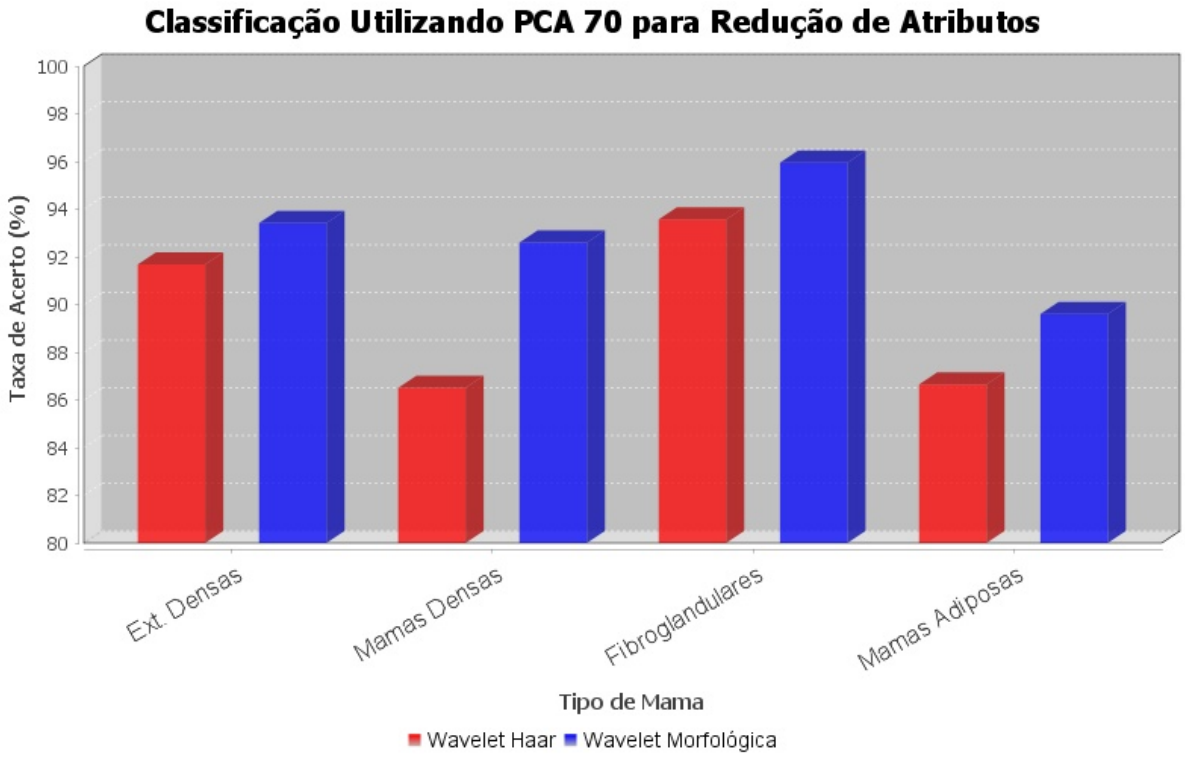


Para este caso, não houve abordagem que possuísse os noventa atributos, na verdade, todos os vetores de componentes principais utilizados para a classificação estavam maximizados de acordo com a técnica de seleção de atributos, cujos valores absolutos e percentuais podem ser observados na tabela 24.

Tabela 24 – Valores absolutos e percentuais da quantidade de atributos selecionados para classificação das instâncias pelo uso da técnica de Análise de Componentes Principais com até 90 atributos para as duas abordagens de pré-processamento

	Wavelet Haar		Wavelet Morf.	
Tipo de mama	Atributos	Percentual (%)	Atributos	Percentual (%)
Ext. Densa	83	37,05	89	39,73
Densa	86	38,39	87	38,84
Fibroglandular	84	37,50	79	35,27
Adiposa	81	36,16	70	31,25

Figura 31 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de Análise de Componentes Principais com 70 atributos, considerando ambos os casos de pré-processamento



6.2 Comparação Geral

Após a análise individual dos dados, foi realizada uma comparação global das técnicas sob uma ótica mais geral, com o objetivo de analisar o desempenho tanto da abordagem de pré-processamento, quanto principalmente para a técnica de seleção de atributos mais adequada para o problema de classificação de tumores de mama a partir do sistema aqui proposto.

Com relação ao comparativo entre as técnicas de pré-processamento, resume-se na tabela 25 os totais dos testes estatísticos realizados, onde estão contabilizadas as vezes em que determinada técnica de pré-processamento foi superior (rejeição de hipótese nula) ou quando ambas obtiveram desempenho idêntico (aceitação da hipótese nula):

Tabela 20 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de Análise de Componentes Principais com 70 atributos para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	91,68	± 0,45	93,42	± 0,46	$5,20 \times 10^{-11}$	Sim ●
Densa	86,51	± 0,62	92,60	± 0,51	$2,78 \times 10^{-11}$	Sim ●
Fibroglandular	93,57	± 0,42	95,95	± 0,36	$2,68 \times 10^{-11}$	Sim ●
Adiposa	86,65	± 0,60	89,60	± 0,44	$2,81 \times 10^{-11}$	Sim ●

Tabela 21 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de Análise de Componentes Principais com 80 atributos para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	92,33	± 0,44	92,43	± 0,46	0,5221	Não ●
Densa	87,75	± 0,67	92,48	± 0,57	$2,89 \times 10^{-11}$	Sim ●
Fibroglandular	94,27	± 0,45	96,54	± 0,42	$2,76 \times 10^{-11}$	Sim ●
Adiposa	87,04	± 0,59	89,60	± 0,44	$2,82 \times 10^{-11}$	Sim ●

Tabela 25 – Resumo comparativo entre as técnicas de pré-processamento sob a ótica dos testes estatísticos, separado pelos tipos de tecido mamário

Tipo de mama	Wavelet Haar	Idênticas	Wavelet Morf.
Ext. Densa	5	6	4
Densa	0	1	14
Fibroglandular	6	2	7
Adiposa	4	0	11
Total	15	9	36

Figura 32 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de Análise de Componentes Principais com 80 atributos, considerando ambos os casos de pré-processamento

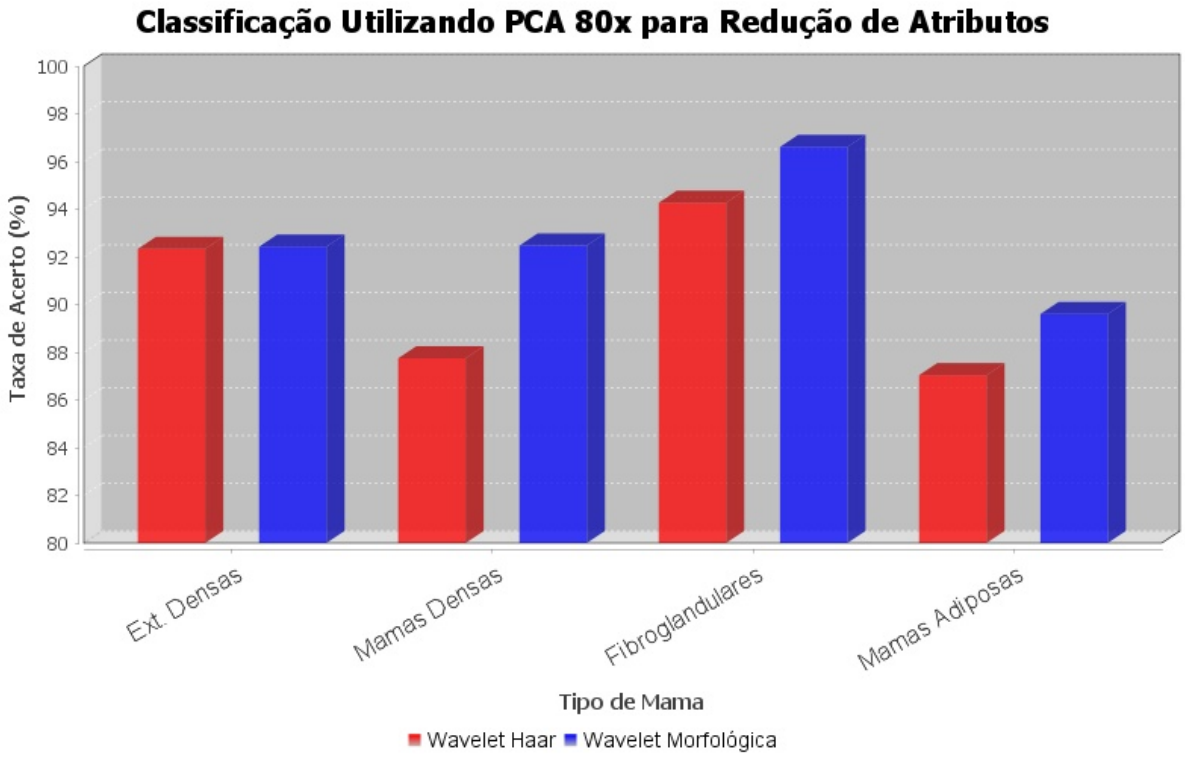
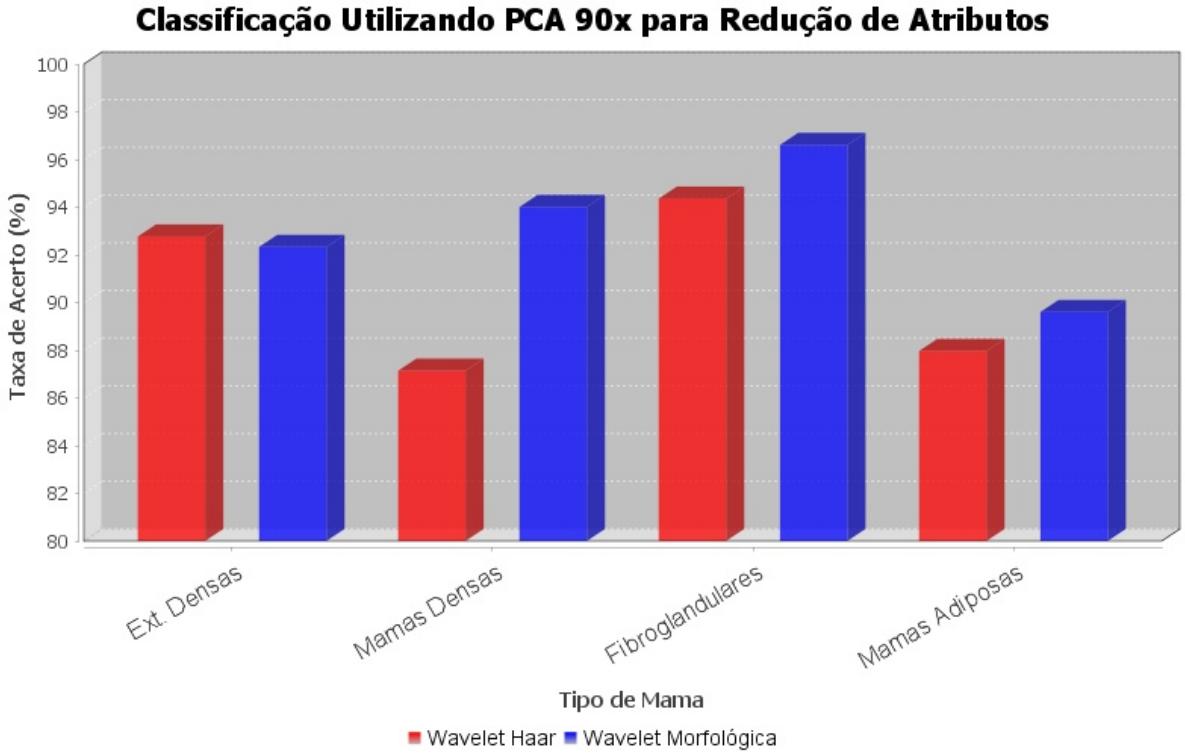


Tabela 22 – Valor do percentual da quantidade relativa de atributos selecionados para classificação das instâncias pelo uso da técnica de Análise de Componentes Principais com até 80 atributos selecionados para as duas abordagens de pré-processamento

Tipo de mama	Wavelet Haar	Wavelet Morf.
Ext. Densa	35,71%	35,71%
Densa	35,71%	35,71%
Fibroglandular	35,71%	35,27%
Adiposa	35,71%	31,25%

De posse destas informações, percebe-se uma forte superioridade do método de pré-processamento baseado em Wavelets morfológicas, porém também é de interesse questionar se este tipo de técnica apresenta os melhores resultados quando levado em consideração o método de seleção de atributos mais apropriado dentre os aqui estudados, pois o ideal seria a melhor combinação de ambos.

Figura 33 – Gráfico de barras das médias das taxas percentuais de acerto para as classificações com redução de atributos pela técnica de Análise de Componentes Principais com até 90 atributos, considerando ambos os casos de pré-processamento



Para chegar à conclusão de qual técnica de seleção de atributos possui o melhor desempenho, foi utilizado um índice para comparar as taxas de redução percentual na quantidade de atributos selecionados em relação ao total de duzentos e vinte e quatro atributos, em contraste com o a redução percentual no nível de acerto da classificação, também em relação à taxa de acerto quando não aplicada a técnica de redução de atributos. Este índice é calculado como a razão entre as duas taxas (em pontos percentuais) citadas.

Na tabela 26 estão presentes todos os índices acerto-atributos discriminados por técnica de seleção de atributos, tipos de composição de mama e técnica de pré-processamento. A técnica de melhor desempenho é aquela que possui o maior valor deste índice. Observa-se que para as abordagens com transformada de Haar, para mamas extremamente densas, a técnica de maior índice é a PCA com até 90 atributos, já para o caso de mamas densas a técnica de CFS com *Best First* foi a melhor, seguida bem de perto pela técnica PCA com 80 atributos. Para o caso de mamas fibroglandulares, a melhor técnica foi a PCA 90 seguida da PCA 80 e CFS com PSO. No caso de mamas adiposas, as técnicas de CFS com AG e CFS com PSO foram melhores. Calculando o valor médios

Tabela 23 – Média (μ) e desvio padrão (σ) do percentual de taxa de acerto para classificação das instâncias com o vetor de atributos reduzido pela técnica de Análise de Componentes Principais com até 90 atributos para as duas abordagens de pré-processamento, seguido do *p-value* para indicação, ou não, da rejeição da hipótese nula ($\mathcal{H}_0$) pelo teste de Wilcoxon a um nível de significância de 5%

	Wavelet Haar		Wavelet Morf.			
Tipo de mama	μ (%)	σ (%)	μ (%)	σ (%)	*p-value*	Rejeição de $\mathcal{H}_0$
Ext. Densa	92,77	$\pm$ 0,52	92,33	$\pm$ 0,59	0,0034	Sim ●
Densa	87,14	$\pm$ 0,63	94,01	$\pm$ 0,50	$2,84 \times 10^{-11}$	Sim ●
Fibroglandular	94,36	$\pm$ 0,50	96,54	$\pm$ 0,42	$2,92 \times 10^{-11}$	Sim ●
Adiposa	87,97	$\pm$ 0,56	89,60	$\pm$ 0,44	$2,74 \times 10^{-11}$	Sim ●

destes índices para os quatro tipos de mama utilizando a transformada de Haar resulta que a técnica com o maior índice médio é a de CFS com PSO com o valor de $0,28$ seguido de PCA 90 com o valor médio de $0,26$.

Tabela 26 – Índices Acerto-Atributos para todos os casos de redução de atributos

	Wavelet Haar				Wavelet Morfológica			
Técnica	E.D.	Den.	Fib.	Adi.	E.D.	Den.	Fib.	Adi.
CFS com AG	0,25	0,16	0,22	0,35	0,30	0,17	0,21	0,13
CFS com PSO	0,26	0,25	0,26	0,35	0,20	0,14	0,21	0,13
CFS com BE	0,20	0,23	0,22	0,12	0,18	0,13	0,09	0,08
CFS com BF	0,18	0,28	0,22	0,30	0,14	0,10	0,14	0,12
GI 50	0,08	0,14	0,12	0,08	0,07	0,08	0,06	0,07
GI 60	0,11	0,15	0,15	0,10	0,08	0,11	0,07	0,11
GI 70	0,11	0,15	0,13	0,09	0,10	0,11	0,09	0,12
GI 80	0,13	0,19	0,14	0,12	0,13	0,15	0,09	0,12
GI 90	0,21	0,15	0,14	0,19	0,16	0,16	0,10	0,13
PCA 50	0,11	0,24	0,15	0,13	0,15	0,12	0,12	0,11
PCA 60	0,14	0,20	0,17	0,12	0,44	0,26	0,17	0,16
PCA 70	0,27	0,19	0,23	0,12	0,68	0,29	0,29	0,15
PCA 80	0,33	0,27	0,27	0,12	0,32	0,25	0,38	0,15
PCA 90	0,43	0,21	0,28	0,12	0,29	0,61	0,38	0,15

Fazendo análise semelhante para o caso com o uso da transformada morfológica, tem-se que para mamas extremamente densas é a PCA com até 70 atributos, para mamas

densas PCA com até 90. Para mamas de predominância fibroglandular, observa-se que as técnicas de PCA 90 e PCA 80 possuem o melhor desempenho, ao passo que para mamas adiposas, PCA 60 se sobressai. Em termos de valores médios, PCA 90 possui um índice de $0,36$, PCA 70 um valor de $0,35$ e o restante das técnicas obteve abaixo de $0,28$. De posse desta informação, também é importante ressaltar que a técnica de PCA com até 90 atributos utilizando transformada morfológica foi a técnica que obteve a maior taxa de acerto média para as classificações dos quatro diferentes tipos de mama, um valor de $93,14\%$.

Desta forma, nota-se uma superioridade no uso da técnica de seleção de atributos baseada na análise de componentes principais, quando selecionados até noventa atributos, na verdade, nenhuma das seleções alcançou o quantitativo de noventa atributos, respondendo por todas as combinações lineares sintetizadas pela referida técnica. Esta técnica apresentou os melhores resultados quando combinada com o uso da transformada de Wavelets morfológica como etapa de pré-processamento das imagens mamográficas e também alcançou taxas de redução de atributos satisfatórias, respondendo por uma redução média percentual de aproximadamente 64% do conjunto total de duzentas e vinte e quatro imagens.

7 Considerações Finais e Conclusões

Câncer de mama é um problema de abrangência mundial. É fato que muitas das técnicas de diagnóstico, acompanhamento e tratamento desta doença não possuem ainda um nível de sofisticação compatível com a tecnologia atual e são, por vezes, arcaicos e um tanto precários, além de serem invasivos e desconfortáveis (como a biópsia para diagnóstico efetivo e a própria mamografia que é um exame bastante doloroso e que expõe a paciente a doses de radiação ionizante que possuem um certo risco de causar outras patologias, bem como o fato deste não ser um exame que preserve bem a dignidade desta paciente). Somado a esses fatos destaca-se também as conhecidas altas taxas de mortalidade associadas a essa patologia, principalmente ligadas ao diagnóstico tardio em estágios de já difícil tratamento e prognóstico desfavorável.

Portanto, o uso de sistemas inteligentes para o auxilio ao diagnóstico se mostra como uma potencial ferramenta de apoio ao profissional, tendo em vista a dificuldade prática de alocar mais de um profissional especialista para a análise de uma mesma imagem mamográfica, bem como por esta ser uma tarefa "enfadonha" por causa do processo bastante repetitivo que pode causar uma fadiga mental ao radiologista, o sistema serviria como um auxilio que destacaria achados para evitar que o médico deixe passar despercebido certos sinais que levem a um diagnóstico mais preciso.

Dentre as lesões procuradas pelos profissionais radiologistas destacam-se as microcalcificações, massas e distorções arquiteturais. Este trabalho esteve focado em lesões do tipo massa ou tumor, levando em consideração principalmente a correlação entre diferentes formas dessas massas com o diagnóstico obtido, de maneira que foi explorado exclusivamente uma técnica de descrição de atributos baseado em forma.

7.1 Conclusões

O primeiro dos objetivos de realização deste trabalho refere-se à comparação de duas técnicas de pré-processamento das imagens mamográficas. Foram utilizadas as transformadas de wavelets utilizando a forma mais simples possível dentre todas as famílias de wavelets (wavelet de Haar) como a primeira abordagem e também foi utilizada a

transformada de wavelets morfológica (utilizando bancos de filtros baseados em morfologia matemática). As duas técnicas foram aplicadas como primeira etapa em todo o fluxo de informação do sistema CAD e foram comparadas para todas as combinações de tipo de tecido mamário e para as diferentes técnicas utilizadas de redução de atributos (e também sem redução alguma) e, através da aplicação de testes estatísticos de Wilcoxon, chegou-se a conclusão que as transformadas morfológicas obtiveram um desempenho superior na grande maioria dos casos (principalmente em mamas do tipo densas e adiposas, sendo de certa equivalência para os outros casos), possivelmente devido ao fato das operações de morfologia matemática em imagens serem de cunho bastante ligado à forma dos objetos presentes nas imagens. Destaca-se, também, o fato da transformada morfológica ter apresentado o melhor desempenho em conjunto com a técnica de redução de atributos que obteve as melhores taxa de redução com menores perdas em acerto.

O segundo e principal objetivo deste trabalho, refere-se à avaliação comparativa entre diversas técnicas de redução de atributos. Foram utilizadas várias dessas técnicas: Análise de componentes principais, ganho de informação e avaliação de subconjuntos para seleção de atributos baseada em correlação, esta última por sua vez utiliza algoritmos de busca para percorrer o espaço de combinações de subconjuntos de atributos, sendo utilizados, para esta tarefa neste trabalho, as técnicas de algoritmos genéticos, otimização por enxames de partículas, busca evolucionária e *best-first*.

Todas as técnicas de seleção/redução de atributos foram testadas em conjunto com as duas diferentes abordagens de pré-processamento de imagens e foram avaliadas a partir do uso de um índice que mensura a quantidade percentual de redução do conjunto original de atributos em relação à redução percentual das taxas de acerto, com o compromisso de penalizar esta última sendo inversamente proporcional, ao passo que é diretamente proporcional à primeira. Chegou-se à conclusão de que a técnica de redução de atributos baseado na análise de componentes principais foi a técnica com os melhores resultados, principalmente pelo fato de levar os atributos original a um novo espaço de representação de maneira a fazer uma "síntese", onde aqueles atributos com maior relevância e representabilidade tentem a se agrupar (através da combinação linear) em novos atributos com grande variabilidade dos dados, como de certa maneira a informação contida em diversos atributos condensasse nas componentes principais, facilitando a classificação por parte das máquinas

de vetor de suporte utilizando um vetor de atributos de entrada consideravelmente menor, sendo a taxa de acerto sensivelmente penalizada. Destaca-se também o fato de que a técnica de análise de componentes principais obteve os melhores desempenhos em conjunto com a técnica de transformada morfológica, sendo ambas consideradas as melhores técnicas comparadas neste trabalho.

Como maiores dificuldades encontradas neste trabalho destaca-se principalmente a dificuldade em ajustar os parâmetros dos algoritmos de busca, principalmente pelo já mencionado caráter empírico de determinação de um conjunto satisfatório de parâmetros. Destaca-se também, que atrelado a isto está o fato de que o sistema CAD deve ser projetado para ser operado pelo profissional radiologista em situações de utilidade prática, o que é totalmente conflitante com a necessidade de ajustes finos em algoritmos que geram dificuldade até mesmo para aqueles que possuem familiaridade com a área. O médico é um profissional da área de saúde que deve preocupar-se o mínimo possível com detalhes de implementação das ferramentas que o mesmo utilizará, portanto, praticidade deve ser um enorme motivador para o desenvolvimento desse tipo de sistema.

Não devem ser medidos esforços para o desenvolvimento de todo tipo de auxilio, como os sistemas CAD, tendo em vista que saúde é o bem de maior importância às pessoas e que doenças como o câncer de mama podem acabar com vidas não somente de quem possui esse tipo de patologia, mas destruir toda uma família. Toda e qualquer forma de contribuição que puder ser dada deve ser encorajada. A engenharia tem muito a contribuir com a área de saúde para que no futuro, não somente esta, mas diversas patologias possam ser diagnosticadas, tratadas e controladas com uma maior facilidade e eficiência, sempre presando o bem estar das pessoas e das sociedades no mundo.

7.2 Sugestões de Trabalhos Futuros

Para trabalhos futuros sugere-se uma melhor exploração dos parâmetros dos algoritmos de busca empregados em conjunto com a técnica de avaliação de subconjunto com seleção de atributos baseado em correlação, também é sugerido o uso de outras técnicas de busca e, mais profundamente, explorar o uso de outras técnicas de seleção/redução de atributos com o intuito de tentar encontras técnicas que usem métricas que possam ser mais compatíveis com o sistema classificador utilizado.

Também é sugerido o uso de outros classificadores que por ventura sejam robustos o suficiente para permitir uma redução mais drástica na quantidade de atributos de entrada sem perder muito da capacidade de classificação.

É sugerido a experimentação de outros descritores de forma, com a intenção de minimizar o tempo de geração desses atributos (utilizando uma técnica menos custosa computacionalmente), bem como o uso de outros tipos de descritores para tentar remover a limitação utilizada neste trabalho de empregar apenas certos tipos de lesão para explorar essas características de forma, permitindo dessa forma a inclusão de imagens de calcificações bem como de massas desconhecidas.

Sugere-se também uma análise da possível ocorrência de atributos que são inerentemente irrelevantes para todos os casos de classificação, implicando na total remoção do mesmo no momento de geração do vetor de atributos, o que seria uma boa maneira de não realizar computação desnecessária.

Para testes e validações, sugere-se a realização de um estudo de caso em pacientes reais, utilizando o sistema proposto empregado em casos de mamografias obtidas em parceria com clínicas ou hospitais locais, obedecendo todo o protocolo ético cabível.

Por fim, sugere-se, também, a aplicação de sistemas CAD em outros tipos de imagens de mama, como as imagens termográficas, de ressonância magnética e principalmente de ultrassonografia da mama, tendo em vista que este é um exame barato, de bastante praticidade, pois existem equipamentos móveis e que permitem uma análise em tempo real da imagem, além de não utilizar radiação ionizante, permitindo o uso mais "indiscriminado" pelas pacientes e permitindo também o uso por pacientes mais jovens, tendo em vista que um dos fatos que mais influencia na não realização de exames de mamografia por pacientes jovens é o fato de não haver uma boa relação de risco/benefício na realização deste tipo de exame, já que são mamas normalmente mais densas (de baixo contraste com possíveis lesões) e que atrelado às baixas taxas de ocorrência de patologias, não compensa a exposição às radiações ionizantes.

Referências[1]

AKAY, M. F. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, v. 36, n. 2, Part 2, p. 3240 – 3247, 2009. ISSN 0957-4174. Disponível em: <http://www-.sciencedirect.com/science/article/pii/S0957417408000912>. Citado na página 22.

BLANKS, R.; WALLIS, M.; MOSS, S. A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the uk national health service breast screening programme. *Journal of Medical Screening*, PubMed, v. 5, p. 195–201, 1998. Citado na página 23.

CHAKRAVARTY, K. et al. Feature selection by differential evolution algorithm - a case study in personnel identification. *IEEE Congress on Evolutionary Computation*, 2013. Citado na página 28.

CHUI, C. K. *An Introduction to Wavelets*. [S.l.]: Academic Press, 1992. Citado na página 41.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, 1995. ISSN 1573-0565. Disponível em: <http://dx.doi.org/10-.1007/BF00994018>. Citado na página 49.

DARWIN, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. [S.l.: s.n.], 1859. Citado na página 58.

DESERNO, T. M. et al. Computer-aided diagnostics of screening mammography using content-based image retrieval. 2012. Citado 2 vezes nas páginas 36 e 65.

FERNANDES, I. M. M. *Sistema de Apoio à Classificação de Lesões Em Mamografias Considerando a Densidade Mamária*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2015. Citado 2 vezes nas páginas 23 e 64.

GANESAN, K. et al. Computer-aided breast cancer detection using mammograms: A review. *Reviews in Biomedical Engineering*, IEEE, v. 6, 2013. Citado 2 vezes nas páginas 21 e 23.

GIL, A. C. *Como Elaborar Projetos de Pesquisa*. 4ª. ed. [S.l.]: São Paulo, Atlas, 2002. Citado na página 64.

GILBERT, F. J. et al. Single reading with computer-aided detection for screening mammography. *New England Journal of Medicine*, v. 359, n. 16, p. 1675–1684, 2008. PMID: 18832239. Disponível em: <http://dx.doi.org/10.1056-/NEJMoa0803545>. Citado na página 22.

GOLDBERG, D. E. *Genetic algorithms in search, optimization and machine learning*. [S.l.]: Addison-Wesley, 1989. Citado na página 58.

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

GONZALEZ; WOODS, R. *Digital Image Processing*. 2ª. ed. [S.l.]: Addison-Wesley, 2002. Citado 3 vezes nas páginas 37, 38 e 39.

GROMET, M. Comparison of computer-aided detection to double reading of screen-ing mammograms: Review of 231,221 mammograms. *American Journal of Radiology*, v. 190, n. 4, 2008. Citado na página 22.

GUYTON, A. C.; HALL, J. E. *Tratado de Fisiologia Médica*. 11ª. ed. [S.l.]: Rio de Janeiro, Elsevier, 2006. Citado na página 33.

HAAR, A. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, v. 69, n. 3, p. 331–371, 1910. ISSN 1432-1807. Disponível em: <http://dx.doi.org/10.1007/BF01456326>. Citado na página 41.

HALL, M. A. *Correlation-based Feature Selection for Machine Learning*. Tese (Doutorado) — University of Waikato, 1999. Citado 2 vezes nas páginas 57 e 71.

HASAN, H.; TAHIR, N. M. Feature selection of breast cancer based on principal component analysis. *6th International Colloquium on Signal Processing and Its Applications (CSPA)*, 2010. Citado na página 29.

HAYKIN, S. *Redes Neurais: Princípios e Práticas*. 2ª. ed. [S.l.]: Porto Alegre, Bookman, 2001. Citado 4 vezes nas páginas 47, 48, 49 e 50.

HELA, B. et al. Breast cancer detection: A review on mammograms analysis techniques. In: *Systems, Signals Devices (SSD), 2013 10th International Multi-Conference on*. [S.l.: s.n.], 2013. p. 1–6. Citado 2 vezes nas páginas 22 e 23.

INCA. *Instituto Nacional do Câncer*. 2015. Disponível em: <http:/-/www2.inca.gov.br/wps/wcm/connect/acoes_programas/site/home/nobrasil/programa_controle_cancer_mama/deteccao_precoce>. Citado na página 22.

INCA. *Instituto Nacional do Câncer*. 2015. Disponível em: <http://www2.inca-.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama/cancer_mama+>. Citado na página 22.

JAIN, A. K.; DUIN, R. P.; MAO, J. Statistical pattern recognition: A review. *Transactions on pattern analysis and machine intelligence*, v. 22, n. 1, 2000. Citado na página 68.

KENNEDY, J.; EBERHART, R. Particle swarm optimization. *Proceedings of IEEE International Conference on Neural Networks*, p. 1942–1948, 1995. Citado 2 vezes nas páginas 58 e 60.

KEYVANFARD, F.; SHOOREHDELI, M. A.; TESHNEHLAB, M. Feature selection and classification of breast mri lesions based on multi classifier. 2011. Citado na página 29.

MAGGIO, C. D. State of the art of current modalities for the diagnosis of breast lesions. *European Journal of Nuclear Medicine and Molecular Imaging*, Springer, v. 31, p. S56–S69, 2004. Citado na página 22.

MUñOZ-MEZA, C.; GóMEZ, W. A feature selection methodology for breast ultrasound classification. *10th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, 2013. Citado na página 30.

NAYEEM, M. A. R. et al. Feature selection for breast cancer detection from ultrasound images. *3rd INTERNATIONAL CONFERENCE ON INFORMATICS, ELECTRONICS and VISION*, 2014. Citado na página 30.

NOLL, R. J. Zernike polynomials and atmospheric turbulence. The Perkin-Elmer Corporation, 1975. Citado na página 45.

OLIVEIRA, J. E. et al. Towards a standard reference database for computer-aided mammography. 2008. Citado 2 vezes nas páginas 36 e 65.

OPPENHEIM, A. V.; SCHAFER, R. W. *Discrete-Time Signal Processing*. 2ª. ed. [S.l.]: Prentice Hall, 1999. Citado na página 41.

OSAREH, A.; SHADGAR, B. Machine learning techniques to diagnose breast cancer. 2009. Citado na página 28.

PEARSON, K. On lines and planes of closest fit to systems of points in space. *Phisiological Magazine*, n. 2, p. 559–572, 1901. Citado na página 51.

PLATT, J. C. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods - Support Vector Learning, MIT Press*, 1998. Citado 2 vezes nas páginas 50 e 70.

POWELL, V.; LEHE, L. *Instituto Nacional do Câncer*. 2015. Disponível em: <http://setosa.io/ev/principal-component-analysis/>. Citado na página 52.

PéREZ, N. et al. Improving the performance of machine learning classifiers for breast cancer diagnosis based on feature selection. *Federated Conference on Computer Science and Information Systems*, 2014. Citado na página 29.

SAMPAT, M. P.; MARKEY, M. K.; BOVIK, A. C. *Computer-Aided Detection and Diagnosis in Mammography, Handbook of Image and Video Processing*. [S.l.]: Elsevier, 2003. Citado na página 22.

SANTOS, W. P. dos; ASSIS, F. M. de. *Algoritmos Dialéticos para inteligência Computacional*. 1ª. ed. [S.l.]: Editora Universitária, 2013. Citado na página 58.

SHANNON, C. E. A mathematical theory of communication. *Bell System Technical Journal*, v. 3, n. 27, p. 379–423, 1948. Citado na página 53.

SUCKLING, J. et al. The mammographic image analysis society digital mammogram database exerpta medica. p. 375–378, 1994. Citado na página 34.

SUN, Y.; BABBS, C. F.; DELP, E. J. A comparison of feature selection methods for the detection of breast cancers in mammograms: Adaptive sequential floating search vs. genetic algorithm. *Engineering in Medicine and Biology 27th Annual Conference*, 2005. Citado na página 30.

TAHMASBI, A.; SAKI, F.; SHOKOUHI, S. An effective breast mass diagnosis system using zernike moments. In: *Biomedical Engineering (ICBME), 2010 17th Iranian Conference of*. [S.l.: s.n.], 2010. p. 1–4. Citado na página 45.

TAHMASBI, A.; SAKI, F.; SHOKOUHI, S. B. Classification of benign and malignant masses based on zernike moments. *Computers in Biology and Medicine*, v. 41, n. 8, p. 726 – 735, 2011. ISSN 0010-4825. Citado na página 46.

THANGAVEL, K.; VELAYUTHAM, C. Rough set based unsupervised feature selection in digital mammogram image using entropy measure. *International Conference on Biomedical Engineering (ICoBE)*, 2012. Citado na página 28.

TORTORA, G. J. *Princípios de Anatomia Humana*. 10ª. ed. [S.l.]: Rio de Janeiro, Guanabara Koogan, 2007. Citado na página 33.

VALENçA, M. J. S. *Fundamentos das Redes Neurais: Exemplos em Java*. 2ª. ed. [S.l.]: Livro Rápido, 2015. Nenhuma citação no texto.