

Disponibilização de imagens mamográficas via Internet: uma forma de contribuir para a integração regional de pesquisas em diagnóstico auxiliado por computador

**Fátima L. S. Nunes¹, Homero Schiabel², J. Alaor Oliveira Jr¹, Claudio E. Goes²,
Rodrigo H. Benatti²**

¹Programa de Pós-Graduação em Ciência da Computação
Centro Universitário Eurípides de Marília, Marília (SP), Brasil

²Laboratório de Processamento de Imagens Médicas e Odontológicas
Departamento Engenharia Elétrica – EESC/USP, São Carlos, SP, Brasil.
fatima@fundanet.br

Abstract. *This work presents a tool to manage a mammographic images basis by Internet. The goal is become available a big set of digitized mammographic images and their information, in order to aid in mammography's CAD (computer-aided diagnosis) schemes evaluation. The system was built with open-source technology, allowing fast and efficient access, without additional cost for the user. The images basis can also be used as a didactic tool in order to aid education of subjects related to mammography.*

Key-words: *database images, image processing, mammography, computer-aided diagnosis.*

Resumo. *Este trabalho apresenta uma ferramenta para gerenciar uma base de imagens mamográficas através da Internet. O objetivo é disponibilizar um grande conjunto de imagens mamográficas digitalizadas e suas respectivas informações, a fim de que seu uso possa contribuir na avaliação de esquemas CAD (Computer-Aided Diagnosis) em mamografia. O sistema foi implementado com tecnologia gratuita, permitindo acesso rápido, eficiente e sem custos adicionais ao usuário. A base de imagens disponibilizada pode ser usada também como ferramenta didática para o ensino de tópicos relacionados à mamografia.*

Palavras-chave: *base de imagens, processamento de imagens, mamografia, diagnóstico auxiliado por computador.*

1. Introdução

As aplicações da computação na área de medicina são cada vez mais freqüentes. Algumas áreas da medicina já não sobrevivem sem os aplicativos computacionais que controlam desde dados de pacientes até unidades inteiras de exames que necessitam de imagens médicas. Com o advento da Internet, uma nova classe de aplicações tem surgido, vencendo barreiras de distâncias e propiciando a troca de experiências entre desenvolvedores de aplicações.

O diagnóstico auxiliado por computador (*CAD – computer-aided diagnosis*) é uma área de pesquisa que tem o objetivo de fornecer aos profissionais de medicina uma segunda opinião no diagnóstico dos mais diversos tipos de doenças [1] e, muitas vezes, contribuir para a detecção precoce de anomalias. Na área de mamografia, vários esquemas CAD têm sido desenvolvidos por grupos de pesquisas. Esquemas completos fornecem a identificação, localização e classificação dessas estruturas, podendo, dessa forma, contribuir para a descoberta precoce da doença, inclusive em pacientes assintomáticas. A partir do desenvolvimento de aplicações de processamento de imagens para esta área [2,3] notou-se a grande dificuldade apresentada em relação à obtenção de imagens para execução de testes e avaliação dos procedimentos implementados. Os resultados obtidos pela aplicação de uma técnica podem variar de acordo com o conjunto de imagens utilizado [4]. Por isso, medir a eficiência de tais esquemas é uma tarefa que demanda testes com um vasto conjunto de imagens. Seria recomendável, porém, que tal conjunto apresentasse algumas características peculiares a fim de possibilitar uma avaliação consistente dos esquemas de processamento, tais como: (a) imagens provenientes de locais diversos a fim de que se possa avaliar o desempenho do sistema com características de aquisição variadas; (b) devem também conter as estruturas de interesse procuradas em um diagnóstico; (c) estar associadas aos laudos médicos correspondentes a cada caso.

A obtenção dessas imagens envolve alguns passos: (1) estabelecimento de convênios entre hospitais e clínicas que tenham interesse em participar do projeto em curso; (2) pesquisa para obtenção dos filmes radiográficos e dos respectivos laudos médicos que atendam aos requisitos estabelecidos; (3) digitalização das imagens com características de resolução espacial e de contraste suficientes para os testes necessários e (4) catalogação eficiente que permita uma recuperação rápida e precisa das imagens de acordo com as suas características.

Todavia, ainda que todos os passos mencionados tenham sido realizados, o conjunto de imagens obtido beneficiaria somente o grupo de pesquisa que o formou, uma vez que estaria disponibilizada somente localmente. Mesmo dentro do mesmo grupo de pesquisa é necessário também desenvolver uma interface apropriada para que o acesso às imagens e dados seja possível para qualquer integrante que não esteja diretamente envolvido com a elaboração da base, além do que a disponibilização local dificulta o acesso aos membros da equipe que, por algum motivo, estão geograficamente distantes. A estrutura da Internet pode auxiliar nesta limitação, uma vez que aplicativos com tecnologia para World Wide Web (WWW) podem ser construídos a fim de disponibilizar o material a todos os interessados, de forma a permitir acesso às imagens de qualquer lugar onde exista uma conexão à WWW. Alguns centros de pesquisa têm oferecido *sites* que disponibilizam, entre outras informações, imagens utilizadas em diagnóstico na área de mamografia [5-9]. Percebe-se, entretanto, que geralmente as bases

de imagens disponibilizadas apresentam limitações em relação à quantidade de imagens oferecidas, características das imagens em relação ao diagnóstico, informações necessárias para confrontar laudos médicos com resultados dos esquemas CAD e formatos proprietários, que demandam árduo trabalho de conversão de dados.

Visando contribuir para o desenvolvimento de esquemas CAD, este trabalho apresenta um sistema desenvolvido para permitir o armazenamento e recuperação de imagens mamográficas via Internet. Assim, foi disponibilizada uma base de imagens formada ao longo de anos de pesquisa nesta área [10]. O armazenamento e recuperação de imagens são feitos através de parâmetros fornecidos pelos usuários, considerando os atributos de um modelo de Banco de Dados definido para armazenar os dados constantes nos laudos médicos. A finalidade é contribuir com o desenvolvimento das pesquisas em sistema de auxílio ao diagnóstico, aproveitando a Internet para promover a integração e fomentar a troca de experiências entre grupos que trabalham nesta área.

2. Materiais e Métodos

Depois de estudar as características dos usuários potenciais e dos dados que seriam armazenados, foram definidos alguns requisitos em relação a *hardware*, *software* e interface. Primeiramente foram definidos os aspectos de *hardware* para o servidor, observando-se que deveria apresentar como características importantes a velocidade do processador, o espaço em memória principal e o espaço de armazenamento em disco. Esses aspectos exercem influência no desempenho do sistema, considerando a sua abordagem multiusuária e as características das imagens. A Figura 1 apresenta um esquema do funcionamento da ferramenta, considerando os usuários definidos. O restante desta seção apresenta uma visão mais detalhada da implementação.

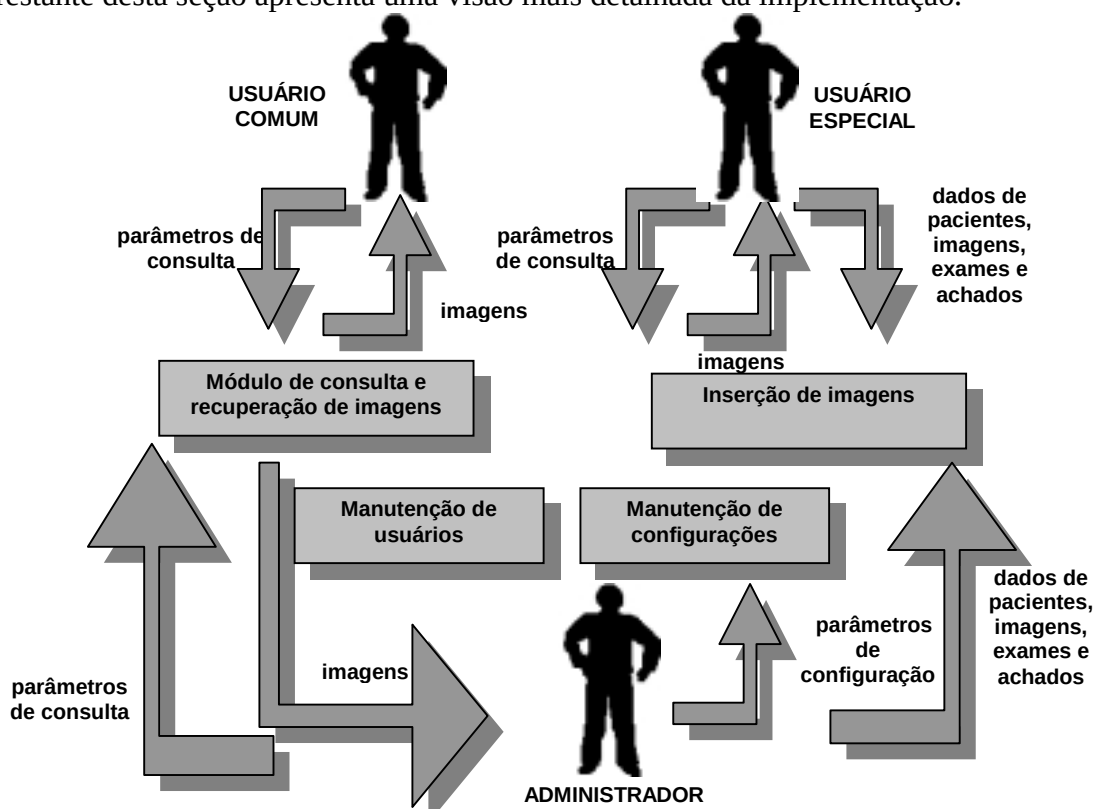


Figura 1 – Estrutura geral do sistema.

A análise de requisitos levou ao estudo de diversas opções de tecnologias de *software* que atendessem a quatro aspectos básicos: custo de aquisição e manutenção, desempenho, construção de interfaces eficientes e facilidade de implementação. A partir desses elementos decidiu-se pelo desenvolvimento de um projeto sem custos adicionais de aquisição e manutenção de *software*. Optou-se, então, pela utilização do sistema operacional Linux, utilizando-se PHP e MySQL, ferramentas *open-source* que atendem aos requisitos estabelecidos. Atualmente o Linux possui compatibilidade com praticamente todos os grandes pacotes gratuitos de *software* [11]. O MySQL é considerado um Sistema Gerenciador de Banco de Dados (SGBD) Relacional, desenvolvido para diversas plataformas, sendo um *software* de livre distribuição para várias plataformas que o utilizam em um servidor Web [12]. A linguagem PHP permite criar *sites* dinâmicos, possibilitando uma interação com o usuário através de formulários, parâmetros da URL (*Uniform Resource Locator*) e *links* [13,14].

Foi realizado um estudo aprofundado dos aspectos das imagens, equipamentos de digitalização, pacientes e exames mamográficos a fim de estabelecer um modelo relacional para armazenar os dados necessários de forma consistente e sem redundância. A Figura 2 mostra as onze tabelas que compuseram o modelo. Sete delas são denominadas tabelas de configuração (destacadas em cor mais escura), com o objetivo de padronizar valores relacionados a hospitais e clínicas que forneceram as imagens (ORIGEM), achados mamográficos (TIPO), posicionamento da imagem (POSICAO), digitalizadores utilizados (SCANNER) e classificação BIRADS - *Breast Imaging Reporting and Data System* (que indica, entre outros aspectos, o nível de malignidade de uma lesão) [15]. As demais armazenam os dados principais, relacionados a pacientes, suas patologias, exames, imagens e achados mamográficos. Como o cadastro do usuário é obrigatório para utilizar o sistema, uma última tabela é usada para armazenar dados dos usuários.

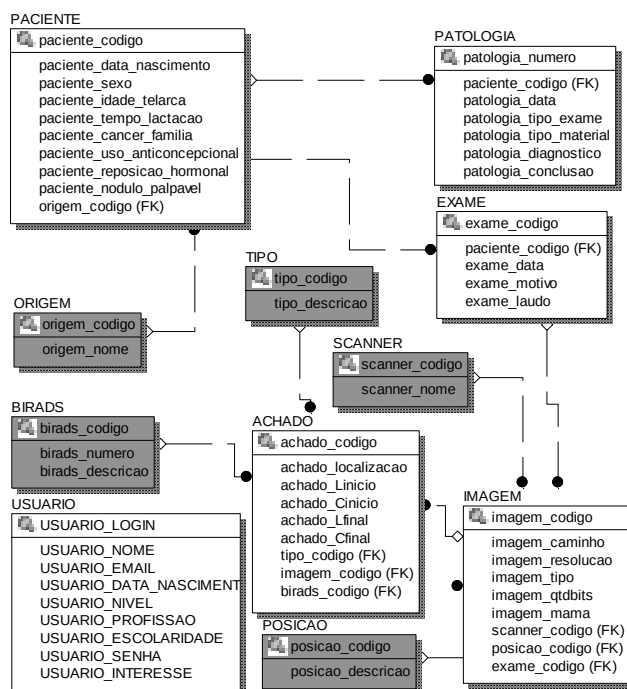


Figura 2 – Modelo de dados utilizado no sistema

A partir da verificação de que usuários potenciais poderiam ser os desenvolvedores de projetos de CAD, professores e alunos de graduação e pós-graduação de áreas relacionadas ao projeto, foram estabelecidos, então, três níveis de usuários: comuns (que podem consultar e copiar imagens); especiais (que podem consultar, copiar e também incluir novas imagens) e administradores (que, além das tarefas dos demais usuários, podem autorizar novos usuários, incluir, alterar e eliminar informações de qualquer parte do Banco de Dados). Um usuário deve ser registrado para obter autorização de acesso ao sistema. Todo usuário é inicialmente incluído com o nível mais baixo de acesso (usuário comum). Ao fornecer suas informações, deve também solicitar autorização para incluir imagens, caso deseje realizar esta tarefa.

2.1. Interfaces e funcionamento do sistema

A partir do modelo de dados, os programas necessários foram construídos, respeitando as regras de integridade e as restrições dos tipos de usuários. As funções disponibilizadas permitem a manutenção em todas as tabelas da base de dados criada, respeitando os seus relacionamentos, recuperação fácil e rápida de informações e manutenção de usuários. Um usuário é incluído no sistema a partir da tela de cadastro. O gerenciamento de usuários é uma opção exclusiva para administradores do sistema. Ao acessar esta função torna-se disponível uma relação de todos os usuários cadastrados. Na alteração, é possível alterar qualquer dado, incluindo o nível do usuário, que é o atributo utilizado para a configuração das permissões de acesso. Além das funções já apresentadas, o administrador tem ainda o controle total das tabelas de configuração do sistema.

Novos dados de pacientes e suas respectivas imagens podem ser incluídos por usuários avançados ou administradores. O administrador pode executar essas opções sem restrições. O usuário avançado também pode inserir novos casos, mas somente podem alterar ou excluir casos que foram cadastrados por ele. A integridade referencial define as ações que devem ser controladas quando operações de inclusão, alteração e exclusão são realizadas em uma tabela que tem relacionamento com outra. Assim, no sistema implementado, o conteúdo de uma chave estrangeira em uma tabela só pode ser incluído se tiver valor nulo ou se houver valor correspondente na tabela referenciada. Analogamente, a alteração desse conteúdo na tabela-filha também obriga a existência de uma ocorrência pré-determinada na tabela principal.

Para respeitar a integridade, a ordem dos procedimentos de inclusão são estabelecidas pelo sistema: primeiro, devem ser incluídos os dados de paciente e suas patologias; depois, os dados dos seus exames e, somente então, são incluídos os dados das imagens e eventuais achados que estas contenham. Ao cadastrar uma paciente o usuário deverá informar os dados apresentados na Figura 3. A inclusão de exames é vinculada a uma paciente já cadastrada. Uma imagem deve referir-se a um exame e, para ser cadastrada, devem primeiramente ser fornecidos os dados a respeito do digitalizador e do seu posicionamento e depois deve ser executado o processo de *upload* da imagem. Após inserir as imagens, o usuário pode inserir os seus achados, fornecendo o tipo (por exemplo: nódulo ou microcalcificações) e a classificação BIRADS. Esses atributos devem ser selecionados, sendo provenientes das tabelas de configuração respectivas. Após a inserção desses dados deve ser fornecida a localização do achado. Na ferramenta não são utilizados padrões de codificação de doenças porque os laudos

médicos não continham tal informação. No diagnóstico do câncer mamário o padrão utilizado é o BIRADS, o qual foi explicado anteriormente.

A consulta genérica é o módulo que executa a principal função do sistema: a recuperação de informações e imagens a partir das necessidades do usuário. Conforme mostra a Figura 4a, a busca pode ser realizada de acordo com as características armazenadas na base de dados a respeito de pacientes, exames, imagens e achados. O resultado é mostrado em HTML, sendo que todo o processamento da ferramenta PHP é feito no servidor, ficando transparente ao usuário. Para a realização dessa consulta foi implementada uma visão - estratégia de SGBDs para selecionar dados de acordo com parâmetros fornecidos - que possibilita a pesquisa das imagens em toda a base de dados. A fim de atender à requisição do usuário, o programa monta dinamicamente o comando SQL (*Structured Query Language*) de acordo com as opções selecionadas. Após o retorno do resultado da consulta, o usuário tem a possibilidade de visualizar a imagem no formato JPEG (*Joint Photographic Experts Group*), ou fazer o *download* da imagem no formato TIFF (*Tagged Image File Format*), o qual representa um armazenamento de imagens sem perda de informação e com 16 bits de resolução de contraste, gerando, por isso, um grande volume de dados a ser copiado. A opção pela conversão das imagens no formato JPEG teve a finalidade de diminuir, portanto, o volume de dados, facilitando a visualização da imagem antes *do download*. Um exemplo de visualização de imagens neste formato é mostrado na Figura 4b.

ID	Nome	Sexo	Data de Nascimento	Idade	Estado Civil	Anticoncepcional	Repetição hormonal	Modo de parto
10001	JOÃO SILVA	M	10/10/1980	28	Casado	Não	Não	Não
10002	MARIA OLIVEIRA	F	15/05/1985	33	Solteira	Não	Sim	Não
10003	PEDRO ALVES	M	20/03/1990	30	Casado	Sim	Não	Não
10004	ANNA COSTA	F	05/08/1992	27	Casada	Não	Sim	Não
10005	CARLOS SOUZA	M	12/01/1995	25	Casado	Não	Não	Sim

Nome do paciente		Data de Nascimento		Idade	
Sexo		Estado Civil		Tempo de gestação	
Câncer na família?		Usa anticoncepcional?		Repetição hormonal?	
Modo de parto?		Modo de parto?		Modo de parto?	

Figura 3 – Interface para inclusão do paciente.

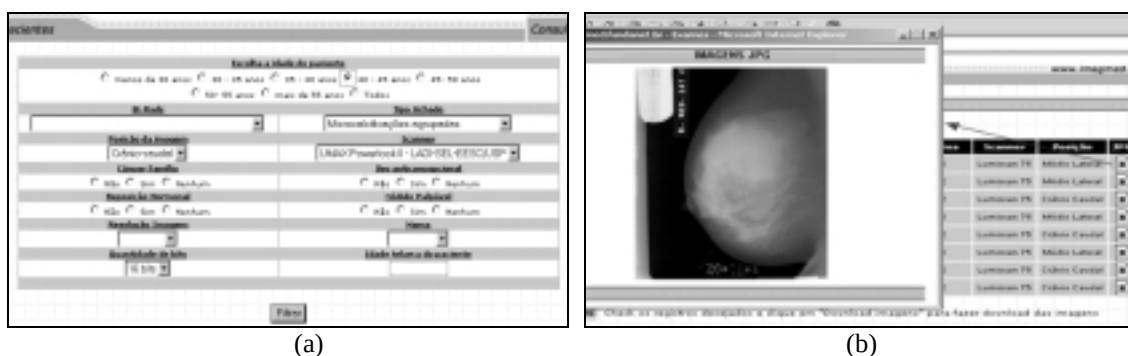


Figura 4– Consulta genérica: (a) Interface; (b) visualização de imagem JPEG.

Finalmente, para o usuário fazer o *download* de um conjunto de imagens, deve selecionar os registros desejados. Será criado um arquivo compactado que engloba todas as imagens selecionadas.

2.2. Formação da base de dados

Uma das tarefas mais árduas na composição de uma ferramenta como a descrita aqui é a formação da base de imagens, que envolve a aquisição dos mamogramas, a digitalização e o armazenamento correto das informações relativas a eles. Para que o teste de esquemas CADs seja efetivo, é desejável que o conjunto de imagens tenha grande quantidade de casos com as estruturas procuradas. O Laboratório de Análise e Processamento de Imagens Médicas e Odontológicas – Departamento de Eng. Elétrica da EESC/USP – vem formando a sua base há alguns anos e, por isso, uma grande quantidade de imagens já estava armazenada em meios digitais antes do início do desenvolvimento do presente sistema. As imagens são provenientes de diferentes hospitais: Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (HC-FMRP/USP), Santa Casa de Misericórdia de São Carlos, Hospital São Paulo (EPM/UNIFESP), Hospital Pérola Byington (São Paulo), Clínica Segalla (Bauru-SP), Hospital Amaral Carvalho (Jaú-SP) e Hospital das Clínicas de Botucatu-SP (UNESP). Além do grande número de imagens, deve-se mencionar que, obviamente em função de provirem de diferentes fontes, as imagens foram obtidas em equipamentos com características, na maioria, bastante diferentes entre si. Em particular, em relação a um grupo específico de imagens provenientes do HC-FMRP/USP, obtidas algumas há mais de uma década em mamógrafos já substituídos por outro equipamento de melhor qualidade em uso atualmente. As imagens disponibilizadas foram editadas com a finalidade de excluir qualquer informação que pudesse identificar pacientes, garantindo, assim, a privacidade dos mesmos.

Na digitalização das imagens foram utilizados dois digitalizadores a laser, ambos da marca Lumisys (Lumiscan 50 e Lumiscan 75) e que, segundo o fabricante, possibilitam obter imagens com até 12 bits de resolução de contraste (4096 níveis de cinza). Uma interface já havia sido confeccionada anteriormente [16,17] para gerenciar localmente as imagens existentes. No entanto, os dados antes armazenados não estavam obedecendo aos paradigmas do modelo relacional de Banco de Dados, possibilitando a ocorrência de erros decorrentes de inconsistências e redundâncias. O SGBD utilizado inicialmente havia sido o Paradox, disponibilizado junto com a linguagem Borland Delphi. A fim de adequar os dados ao novo modelo proposto, foi necessário idealizar algumas alterações assumindo-se alguns valores que ainda não eram armazenados no modelo anterior.

3. Resultados e Discussões

A ferramenta apresentada está disponível no endereço: <http://143.107.235.15/lapimo/bancoweb>. Estão disponíveis atualmente em torno de 3200 imagens, referentes a quase 700 pacientes. A maior parte destas imagens (90%) apresenta os respectivos laudos radiológicos acompanhados dos laudos anátomo-patológicos e citológicos (quando realizados). O restante das imagens (aproximadamente 10%) encontra-se sem laudo devido a razões tais como perda/extravio do documento dentro do próprio hospital ou por óbito da paciente, o que impossibilitou o acesso às informações desejadas.

A Figura 5a mostra de forma gráfica a distribuição das imagens de acordo com as estruturas patológicas visualizadas na imagem, confirmando que a base de imagens formada retrata com satisfatória fidelidade a realidade clínica observada nos hospitais

estudados, uma vez que praticamente um terço das imagens incorporadas apresentam laudo sem qualquer tipo de anormalidade. Estes exames correspondem à grande parte da população feminina que se submete à mamografia preventiva ou mesmo de acompanhamento e controle clínico. Portanto, é de se esperar que sejam retratados em maior número dentro deste estudo. Além disso, as imagens consideradas “não-patológicas” têm grande utilidade na avaliação das detecções falsas-positivas dos diversos algoritmos de processamento de imagens dentro de um esquema CAD. Dentre as imagens com algum tipo de patologia, pode-se notar uma distribuição praticamente uniforme (13%, 15% e 17%) entre os três principais tipos de achados mamográficos. Este fato também retrata a realidade clínica analisada e apresenta a vantagem de contemplar de forma igualitária as estruturas de interesse para o processamento. Isto mostra que o banco de imagens encontra-se preparado para ser utilizado na busca/recuperação de imagens independentemente de qual seja o algoritmo em teste.

Outro fato importante que deve ser levado em conta no processamento de imagens mamográficas diz respeito ao intervalo de idade das pacientes estudadas. A Figura 5b apresenta graficamente a distribuição das imagens de acordo com a idade das pacientes. Neste gráfico estão representadas todas as imagens, não importando se os laudos estão ou não cadastrados no Banco de Dados. O gráfico apresentado mostra que grande parte dos exames (aproximadamente 48%) refere-se a pacientes com idades entre 40 e 60 anos. Este fato é consequência direta de fatores como a recomendação de se realizar mamografias de base em mulheres a partir de 40 anos. Outro fator que leva a esta distribuição é a crescente busca por tratamentos de reposição hormonal, o que leva à realização de mamografias de controle, geralmente realizadas também a partir dos 45 anos. Um terceiro fator que leva à concentração de exames nesta faixa etária é o nível de risco de desenvolvimento de doenças da mama a partir desse período. Mulheres mais jovens geralmente apresentam mamas com maior densidade, o que dificulta a detecção de possíveis estruturas patológicas pela mamografia. Adicionalmente, estatísticas médicas mostram que pacientes com mais de 60 anos que não apresentam histórico de doenças mamárias, dificilmente irão desenvolver qualquer tipo de patologia da mama após essa idade, diminuindo assim a recomendação da realização de mamografias em mulheres com tal característica.

A escolha dos casos que são utilizados em cada teste pode influenciar decisivamente na análise de desempenho do CAD em questão. Dessa forma, quaisquer comparações envolvendo diferentes processamentos são questionáveis se nenhuma informação adicional for fornecida, uma vez que a performance depende da “dificuldade” associada aos casos analisados. Conjuntos que possuem imagens com estruturas de interesse evidenciadas irão ocasionar processamentos mais precisos e, obviamente, taxas menores de erros [4]. Atualmente diferentes grupos de pesquisa desenvolvem novas técnicas de processamento e algoritmos de otimização de desempenho de esquemas de auxílio ao diagnóstico em mamografia e, para isso, desenvolvem também bancos de imagens para a realização de seus testes [5-9,18-20]. Desta forma, a validação dos resultados obtidos com essas novas técnicas é feita basicamente empregando-se as imagens disponíveis dentro do próprio grupo. Dado que esses mamogramas são obtidos em mamógrafos diversos e digitalizados em diferentes equipamentos, as imagens correspondentes podem possuir características diferentes quanto à resolução espacial, resolução de contraste, formato de armazenamento do arquivo digital, como efeito de diferentes características dos equipamentos de aquisição

e digitalização, gerando imagens com qualidades diferentes entre si. Assim, estas técnicas de processamento poderiam ser melhor avaliadas e rapidamente descartadas caso apresentassem baixo desempenho.

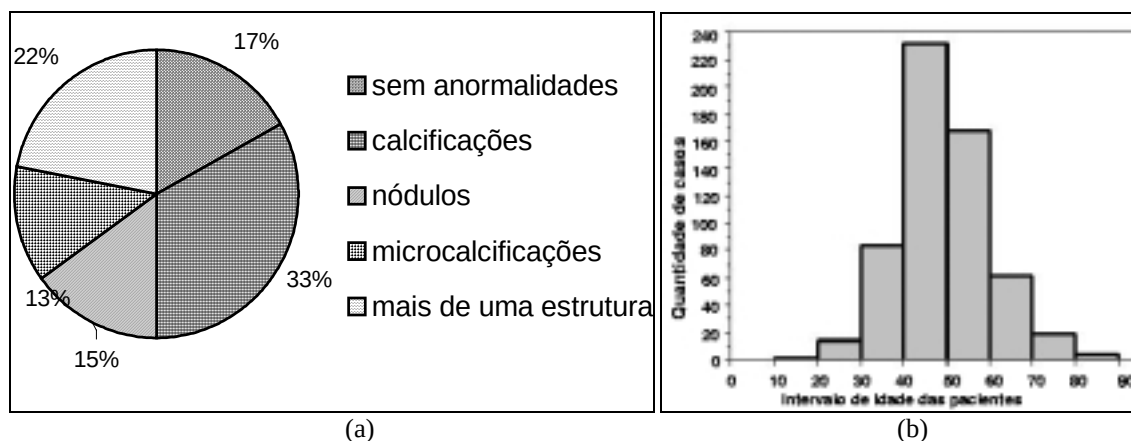


Figura 5 – Composição da base de imagens. (a) Distribuição de acordo com as estruturas de interesse; (b) Distribuição das imagens de acordo com o intervalo de idades.

Deve-se ainda considerar que a distribuição de patologias mamárias em determinada região pode diferir muito da distribuição observada em outros lugares, sendo que o percentual de imagens consideradas normais e patológicas presentes no banco de imagens também pode, e deve, variar de acordo com o local onde está sendo desenvolvido. Este procedimento não invalida o algoritmo utilizado, mas o índice de acertos/erros obtido deve ser analisado considerando os fatores mencionados anteriormente e também a qualidade das imagens processadas. Na base atual estão inseridas imagens somente do Estado de São Paulo. A identificação do local de origem é feita com base na tabela apropriada para isso, que consta no modelo de dados. A facilidade de utilização da ferramenta via Internet permite que inclusão de novos casos seja realizada remotamente, permitindo que imagens de outras regiões do Brasil e do mundo venham compor a base e possibilitando uma representatividade mais homogênea de regiões diferentes.

Estudos recentes mostram que, no caso de imagens mamográficas digitalizadas para fins de processamento, deve ser utilizada resolução de contraste de, no mínimo, 10 bits [19,20]. Assim, é possível diferenciar de forma adequada o maior intervalo possível de densidades óticas (graus de escurecimento) presentes no filme que, depois de digitalizadas, são expressas pelos valores dos níveis de cinza da imagem.

Atualmente existe um consenso também no que diz respeito à resolução espacial que deve ser utilizada no caso de imagens digitalizadas empregadas em processamento. Por exemplo, nível de resolução com pixels menores do que 0,1 mm não contribui significativamente para o aumento das taxas de detecções verdadeiras-positivas, mas aumenta de forma inconveniente as detecções falso-positivas [21,22]. O fato de aumentar o número de falsos-positivos contribui para a deterioração do desempenho dos esquemas computacionais de processamento e pode, inclusive, falsear diagnósticos que eventualmente sejam feitos de forma visual.

Uma possível solução para esse problema seria a utilização do mesmo conjunto de imagens na avaliação dos diversos CADs e das novas técnicas desenvolvidas. Isso demandaria a criação de uma base de imagens comum, de uso público, que pudesse ser

utilizada sempre que necessário. Durante a criação de uma base de imagens comum seria necessária a realização de estudos quanto às taxas de ocorrência das patologias mamárias de acordo com a região considerada, uma vez que estas patologias estão relacionadas à população submetida ao exame. Somente assim a nova base poderia ser considerada representativa da realidade clínica existente, justificando seu uso comum durante os testes de processamento [18].

É importante ressaltar que os equipamentos utilizados para digitalizar as imagens da base aqui considerada, possibilitam obter imagens com até 12 bits de resolução de contraste (4096 níveis de cinza). Testes realizados utilizando medições de densidades óticas em ambos os aparelhos mostraram que a resolução de contraste real alcançada é de aproximadamente 3600 níveis de cinza [22]. Comprovou-se, no entanto, que estes 3600 níveis de cinza são suficientes para um bom desempenho do processamento. Este fato pode ser melhor entendido quando se nota que as principais estruturas de interesse para o processamento de imagens mamográficas têm característica de maior densidade, apresentando, portanto, imagens mais claras. Os níveis de cinza representados por estas estruturas estão localizados em sua maior parte nos valores mais próximos ao início da escala, não tendo sua visibilidade alterada significativamente pela diferença de amostragem citada. A principal diferença entre os digitalizadores utilizados diz respeito à questão da resolução espacial. Enquanto o modelo Lumiscan 50 tem a capacidade de digitalizar as imagens com tamanho de pixel de 0,150 mm, o modelo Lumiscan 75 pode fornecer imagens com pixels de até 0,075 mm. Esta diferença no tamanho dos pixels pode alterar o processamento, conforme citado anteriormente, mas não afeta a visualização normal da imagem no banco de imagens, uma vez que a maior limitação, neste caso, está na resolução máxima que pode ser expressa pelo equipamento de exibição utilizado, no caso, o monitor de computador. A única alteração significativa que pode ser notada quando as imagens são comparadas é o efeito de aumento no tamanho dos arquivos gerados por ambos os digitalizadores. Deve-se salientar que uma diminuição de 50% no tamanho de cada pixel da imagem acarreta um aumento de quatro vezes na quantidade de pontos expressos na imagem por unidade de área. Desse modo, o tamanho do arquivo também é quadruplicado, acentuando ainda mais o problema da escolha da melhor forma de armazenamento das imagens.

3.1. Comparação com outras bases de imagens

No geral, os bancos de imagens já desenvolvidos ou em processo de elaboração apresentam grandes diferenças entre si. Esta diferenciação deve-se principalmente à grande quantidade de diagnósticos médicos que levam em consideração imagens geradas pelos diversos exames existentes. Dessa forma, os bancos são, geralmente, compartimentados de acordo com o exame e a natureza da imagem de interesse a ser posteriormente processada e/ou visualizada.

No caso das imagens mamográficas, que são geradas por aparelhos de raios-X, existem atualmente cinco bases de imagens já completas ou em desenvolvimento, sendo que três delas estão disponíveis para uso de outros grupos de pesquisa e duas ainda continuam restritas ao grupo de pesquisa proprietário.

A Tabela 1 mostra um quadro comparativo entre essas bases e o banco de imagens desenvolvido neste trabalho. É possível constatar a compatibilidade e

proximidade com os demais sistemas desenvolvidos por outros grupos de pesquisas com a mesma área de interesse. Deve-se salientar que a base aqui apresentada continua em expansão, através da incorporação de novas imagens/laudos provenientes dos hospitais com os quais mantemos convênio ou que venham a fazê-lo futuramente. A partir da sua disponibilização efetiva, espera-se favorecer um aprimoramento, tanto do ponto de vista de inclusão de novas imagens quanto de inclusão de novos atributos que se forem tornando necessários ao longo do tempo. A seguir, é apresentada uma breve discussão sobre as características de cada uma das bases mencionadas.

Tabela 1- Comparação entre bancos de imagens existentes (* indica informações não disponibilizadas)

Características	Bancos de imagens					
	MIAS	DDSM	UCSF/LLNL	CALMa	Nijmegen	Nossa base
Origem	Grã-Bretanha	Estados Unidos	Estados Unidos	Itália	Holanda	Brasil
Quantidade de imagens	320	2620	198	3000	40	3300
Resolução de contraste (bits/pixel)	8	12/16	12	12	12	12
Resolução espacial (μm)	50	42/50	35	85	100	100
Quantidade de hospitais	*	4	*	5	*	7
Formato da imagem armazenada	PGM	JPEG sem perdas	*	*	*	TIFF/JPEG

O banco de imagens proposto pelo MIAS [5] é o mais utilizado hoje em dia por grupos de pesquisa espalhados pelo mundo que realizam os mais diversos testes de processamento em imagens mamográficas. Esse fato deve-se principalmente por ter sido o primeiro banco de imagens a ser efetivamente concluído (1994) e também por estar disponível através da Internet há um longo tempo. De acordo com os próprios criadores, a base possui atualmente 57 cópias completas sendo utilizadas em 11 países. As imagens incluídas neste banco foram obtidas através do Programa Britânico de Acompanhamento Médico (*United Kingdom National Screening Programme*), sendo referentes a pacientes com idades entre 50 e 65 anos. Todas as imagens disponíveis são de vista médio-lateral e foram digitalizadas com um equipamento da marca JOYCE-LOEBL (SCANDIG3) que fornece imagens com tamanho de pixel de até 50 μm (resolução espacial) e resolução de contraste de 8 bits/pixel (256 níveis de cinza). O banco conta com 320 filmes digitalizados (aproximadamente 160 pacientes) e está armazenado no formato de compressão TAR, utilizado pelo sistema operacional UNIX. Existe ainda uma segunda versão deste banco, chamada de “MIAS MiniDatabase”, onde são disponibilizados 322 filmes digitalizados com 8 bits/pixel e armazenamento no formato PGM. Esta versão apresenta a grande desvantagem de ter sido elaborada com tamanho de pixel de 200 μm , o que limita sua aplicação prática em testes de processamento. Juntamente com esta versão, são disponibilizadas informações sobre regiões de interesse para eventuais processamentos na forma de diagnóstico acompanhado de coordenadas do centro da região e do raio da mesma.

O DDSM (*Digital Database for Screening Mammography*) foi desenvolvido pela *University of South Florida*, tendo sido encerrado em 1999 [6]. Conta com 2620

filmes digitalizados, sendo 4 filmes referentes a cada exame (4 vistas mamográficas usuais), totalizando aproximadamente 650 pacientes. As imagens foram obtidas junto a 4 hospitais e digitalizadas em 4 equipamentos diferentes, cujas características são apresentadas na Tabela 2. Juntamente com as imagens presentes neste banco, constam também informações consideradas relevantes para o acompanhamento da paciente e para eventuais processamentos futuros, tais como: idade da paciente, data da digitalização, data da realização do exame e informação referente à densidade da mama fornecida por radiologista, segundo critério adotado pelo ACR (*American College of Radiology*). Com o objetivo de preservar a identidade da paciente, as imagens mostradas ao usuário final tiveram as partes em que constavam informações sobre ela removidas. Também para facilitar a visualização, o fundo da imagem foi removido de forma automática, evitando a presença de artefatos indesejados na imagem. O armazenamento das imagens foi feito utilizando-se algoritmo de compressão JPEG sem perdas. O banco de imagens inclui também um *software* para detecção de massas mamográficas. Outro pacote é utilizado para a conversão de formatos de imagens em 8 e 16 bits, redimensionamento em outras resoluções espaciais (diferentes da aquisição original) e gravação em 2 formatos: PGM (*Portable Gray Map*) e TIFF.

Tabela 2 - Características de digitalizadores do DDSM

SCANNER \ CARACTERÍSTICA	RESOLUÇÃO ESPACIAL	RESOLUÇÃO DE CONTRASTE
DBA M2100 ImageClear	42 μm	16 bits/pixel
HOWTEK 960	43 μm	12 bits/pixel
LUMYSIS 200 LASER	50 μm	12 bits/pixel
HOWTEK MULTIRAD850	43 μm	12 bits/pixel

O UCSF/LLNL *Digital Mammogram Library* é um banco de imagens criado através da parceria entre a *University of California at San Francisco* e o *Lawrence Livermore National Laboratory* [7]. Vem sendo largamente utilizado, principalmente dentro dos Estados Unidos, devido a duas características: está disponível em CD-ROM, ocupando pouco espaço em disco, e utiliza compressão sem perdas com redução entre 10 e 30 vezes no tamanho do arquivo original, o que facilita a transmissão entre pontos remotos através de redes de comunicação. Um aspecto desvantajoso desta base é o baixo número de imagens (e conseqüentemente pacientes) disponível. Até o momento, estão cadastradas 50 pacientes do programa de acompanhamento da *University of California* que deram origem a 198 filmes digitalizados, sendo quatro filmes de cada uma das 49 pacientes correspondendo às quatro vistas mamográficas e mais duas imagens de uma paciente que foi submetida a mastectomia total. As imagens foram digitalizadas utilizando-se um *scanner* DuPont (Wilmington, DE) que gera imagens com 12 bits de resolução de contraste e tamanho de pixel de 35 μm .

O projeto CALMa (*Computer Assisted Library for Mammography*) é resultante da colaboração entre o *Istituto Nazionale di Física Nucleare* (Itália) e alguns centros médicos italianos com o objetivo de desenvolver um esquema CAD baseado em redes neurais [8]. Integrada a este projeto foi desenvolvida também uma base de imagens para ser utilizada nos testes de processamento e na validação/treinamento das redes neurais. As imagens obtidas junto aos hospitais encontram-se hoje armazenadas em 240 CD-ROM's, mas ainda não estão disponíveis para uso externo. Todas as 3000 imagens, referentes a aproximadamente 1400 pacientes, foram digitalizadas com resolução de

contraste de 12 bits e 85 μm de resolução espacial. Cerca de um terço das imagens correspondem a casos considerados patológicos, o que possibilita o treinamento e o teste das redes neurais com uma grande diversidade de opções.

O *Nijmegen Digital Mammogram Database* [9] é a base de imagens menos utilizada atualmente devido ao seu reduzido número de casos/pacientes e ao formato proprietário utilizado no armazenamento. Ainda é utilizada em alguns estudos, mas somente com o propósito de comparações entre desempenhos de esquemas de processamento. A base foi desenvolvida por um grupo de pesquisa sediado na Holanda e conta com apenas 40 imagens provenientes de 21 pacientes. As imagens foram digitalizadas com 12 bits/pixel e resolução espacial de 100 μm .

Como pode ser percebido através da análise da Tabela 1, a base aqui apresentada contempla os mesmos requisitos das bases disponibilizadas em relação a alguns aspectos e é superior em relação a outras características. A quantidade de imagens disponibilizada é superior a todas as outras, o que pode ser afirmado como um ponto forte, pois o volume de imagens utilizadas nos testes é um fator importante para ratificar a eficiência de um CAD. Em relação à resolução espacial e resolução de contraste, nossa base é igual àquelas que oferecem resoluções adequadas para o processamento, conforme já discutido. Isso acontece também em relação ao formato. Por apresentar formatos conhecidos, que podem ser facilmente decodificados por programas disponíveis comercialmente, nossa base apresenta-se igual àquelas que também preenchem este requisito. O fornecimento de formatos proprietários pode dificultar ou inviabilizar a utilização do material disponibilizado. Por fim, é importante destacar que a quantidade de locais de obtenção das imagens considerados na nossa ferramenta é maior que o informado nas demais bases. Esta característica auxilia na diversidade de imagens em relação às características de aquisição e representatividade de anomalias.

4. Conclusões

O processo de formação de uma base de imagens médicas é longo e detalhista. Além de contemplar os requisitos de variedade das imagens, este processo pode ainda esbarrar em duas dificuldades concretas: disponibilização dos exames pelos hospitais e clínicas (muitas vezes as imagens ficam em poder do paciente) e o curto tempo disponível para que as imagens e seus dados sejam armazenados no banco, uma vez que os filmes originais não podem permanecer por longo tempo fora dos arquivos hospitalares. Assim, uma primeira conclusão que podemos tirar do trabalho apresentado é que através dele essas dificuldades podem estar superadas para aqueles que necessitam de imagens mamográficas.

Um outro ponto a ser ressaltado é que a modelagem relacional do sistema aqui proposto visa também à expansão da base a partir da inclusão de outros tipos de imagens, sejam de imagens de mama provenientes de outras modalidades (ultra-som e ressonância magnética, por exemplo) ou de imagens de outros órgãos que podem ser úteis à comunidade de pesquisadores na área. Aproveitando a estrutura já construída, é possível acrescentar novas tabelas ao modelo e elaborar consultas adequadas a tais imagens.

Por fim, é importante ressaltar a contribuição que sistemas deste tipo podem trazer para a telemedicina, facilitando a integração entre instituições e pesquisadores

envolvidos em projetos nesta área. Com as dimensões de um país como o Brasil, sistemas gratuitos disponibilizados pela Internet podem efetivamente auxiliar no desenvolvimento de pesquisas mesmo em regiões nas quais é difícil conseguir matérias-primas para a execução de projetos – no nosso caso, imagens digitalizadas de mamogramas reais. A facilidade de utilização de uma base de imagens via Internet com as características aqui apresentadas é clara: interessados de qualquer parte do mundo que tenha um ponto de Internet poderão acessá-la e copiar as imagens que lhe interessam. A arquitetura proposta no sistema aqui apresentado facilita tanto a recuperação dessas imagens quanto a inclusão de novos casos, podendo enriquecer, assim, o conteúdo da base de imagens em construção. Esse aspecto pode fazer com o sistema seja útil também como ferramenta educacional, visto que as imagens disponibilizadas apresentam grande variedade de achados, podendo constituir materiais didáticos complementares em cursos relacionados a essa área de conhecimento. A ferramenta apresentada permite que imagens de diversas localidades e com características diferentes de aquisição sejam armazenadas. Assim, pesquisadores da área podem compartilhar este material e comparar os resultados de sistemas desenvolvidos para auxílio ao diagnóstico em mamografia, considerando-se o mesmo conjunto de dados de teste. Isto pode contribuir para uma avaliação mais adequada do CAD, assim como fomentar a colaboração entre grupos que têm este objeto de pesquisa.

Desta forma, acreditamos estar contribuindo para o desenvolvimento de pesquisas na área de esquemas de auxílio ao diagnóstico, através da minimização de um desafio a ser vencido, que é a disponibilização de material para avaliação de tais sistemas.

Agradecimentos

À Fapesp – Fundação de Amparo à Pesquisa do Estado de São Paulo, pelo apoio financeiro, e aos hospitais e clínicas que cederam imagens.

Referências bibliográficas

1. DOI, K. Computer-aided diagnosis: present and future. *A new horizon on medical physics and biomedical engineering*. H. Abe, K. Atsumi, T. Iiuma, M. Saito, M. Inoue, eds. Elsevier Science Publishers B.V., p. 59-66, 1991.
2. SCHIABEL, H.; NUNES, F. L. S.; AZEVEDO MARQUES, P. M.; FRÈRE, A. F. A computerized scheme for detection of clusters of microcalcifications by mammograms image processing. *Medical & Biological Engineering & Computing*, v.35 (Supplement Part 2), p. 705, 1997.
3. NUNES, F. L. S.; SCHIABEL, H. Detecção de agrupamentos de microcalcificações mamárias através de processamento de imagens mamográficas com detecção semi-automática de regiões de interesse. *Revista Brasileira de Engenharia Biomédica*, v.16 (3), p.139-151, 2000.
4. NISHIKAWA, R. M.; JIANG, Y.; GIGER, M. L.; SCHMIDT, R. A.; VYBORNÝ, C. J.; ZHANG, W.; PAPAIOANNOU, J.; BICK, U.; NAGEL, R. DOI, K. Performance of automated CAD schemes for the detection and classification of clustered microcalcifications. *Digital mammography*. A. G. Gale et al., eds. Elsevier Science Publishers B.V., p. 13-20, 1994.
5. SUCKLING, J.; PARKER, J.; DANCE, D.; ASTLEY, S.; HUTT, I.; BOGGIS, C.; RICKETTS, I.; STAMATAKIS, E.; CERNEAZ, N.; KOK, S.; TAYLOR, P.; BETAL, D.; SAVAGE, J.; (1994) "The mammographic images analysis society mammogram database" -

- Excerpta Medica. International Congress Series 1069: 375-378 (1994)
mias@sv1.smb.man.ac.uk
6. HEATH, M.; BOWYER K.; KOPANS D.; MOORE R.; KEGELMEYER, JR., P. (2000)- "The Digital Database for Screening Mammography", Proceedings of the 5th International Workshop on Digital Mammography (Toronto, Canada, June 2000), Medical Physics Publishing
 7. MASCILO, L. N.; FRANKEL, S. D.; HERNANDEZ, J. M.; LOGAN, C.M.(1996)- "Building the LLNL/UCSF Digital Mammogram Library with Image Groundtruth" - Proceedings of the 3rd International Workshop on Digital Mammography, Chicago, June 1996 (Doi K ed.) Elsevier Science BV: Amsterdam 1996.
 8. AMENDOLIA, S.R.; BISOGNI, M.G.; BOTTIGLI, U.; CECCOPIERI, A.; DELOGU, P.; DIPASQUALE, G.; FANTACCI, M.E.; LORENZINI, E.; MARCHI, A.; MARZULLI, V.M.; OLIVA, P.; PALMIERO, R.; REGGIANI, M.; ROSSO, V.; STEFANINI, A. ; STUMBO, S.; TANGARO, S.; VENIER, O. (2001) – "The CALMA Project" - Nuclear Instruments & Methods in Physics Research Section A-Accelerators Spectrometers Detectors and Associated Equipment, Vol. 461 (1-3) pag. 428-429, 2001.
 9. KARSSEMEIJER, N. A common database for research in mammographic image analysis, *SPIE 93: Biomedical Image Processing and Biomedical Visualization*, v. 1905, p. 542-543, 1993.
 10. SCHIABEL, H.; NUNES, F.L.S.; ESCARPINATI, M.C.; GÓES, C.E. Investigations on the effect of different characteristics of images sets on the performance of a processing scheme for microcalcifications detection in digital mammograms. *Journal of Digital Imaging*, v.14 (2), p.224-225, 2001.
 11. JAMIL, G. L. *Linux para principiantes*, Axcel Books, Rio de Janeiro, 1999.
 12. YAGER, R. Y. et al. *MySQL & mSQL*, O'Reilly, USA, 1999.
 13. BARRETO, M. V. S. *Tutorial de Linguagem PHP*, <http://www.vivas.com.br>, 06/2000, 1999.
 14. FISHER, H. G. *PHP Guia de Consulta Rápida*, Novatec, São Paulo, 1999.
 15. AMERICAN COLLEGE OF RADIOLOGY (ACR), *Breast imaging reporting and data system (BI-RADS®)*, 3ed. Reston (VA): American College of Radiology, 1998.
 16. BENATTI, R. H.; SCHIABEL, H.; NUNES, F. L. S. Construção de uma base de mamogramas digitalizados para utilização em sistema de diagnóstico auxiliado por computador, *XXIV Encontro Nacional de Física da Matéria Condensada*, São Lourenço - MG, 15 a 19 de maio de 2001.
 17. BENATTI, R. H.; SCHIABEL, H.; Nunes, F.L.S. A mammographic images database for computer-aided diagnosis schemes, *World Congress on Medical Physics and Biomedical Engineering*, 25-31 August 2003, Sidney, Australia.
 18. NISHIKAWA, R. M.; YARUSSO, L. M. Variations in measured performance of CAD schemes due to database composition and scoring protocol, *SPIE 99: Medical Imaging*, v. 3338, p. 840-844 – 1998.
 19. DANCE, D. R. Design of a common database for research in mammogram image analysis, *SPIE 93: Biomedical Image Processing and Biomedical Visualization*, v. 1905, p. 538-539 - 1993.
 20. RANGAYYAN, R.; PARANJAPE, R.; SHEN, L.; DESAUTELS, J. E. L. A database for mammographic image research, *SPIE 93: Biomedical Image Processing and Biomedical Visualization*, v. 1905, p. 550-551, 1993.
 21. NUNES, F. L. S.; SCHIABEL, H.; ESCARPINATI, M. C.; GOES, C. E. Comparisons of different contrast resolutions effects on a computer-aided detection system intended to cluster microcalcifications detection in dense breasts images, *Journal of Digital Imaging*, v. 14, n. 2, suppl. 1, p. 217-219, 2001.
 22. ESCARPINATI, M.C. *Investigação de formatos e compressão de imagens digitais para processamento de imagens mamográficas de mamas densas*. Dissertação (Mestrado), EESC/USP, São Carlos (SP), 2002.