FUNDAÇÃO DE "ENSINO EURÍPIDES SOARES DA ROCHA"
CENTRO UNIVERSITÁRIO "EURÍPIDES DE MARÍLIA" - UNIVEM
PROGRAMA DE MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

MARISA SILVANA DE SOUZA ANDRADE

Utilização de Teste Estrutural para Aperfeiçoar a Avaliação de um Sistema de Auxílio ao Diagnóstico

MARÍLIA/SP
2005

MARISA SILVANA DE SOUZA ANDRADE

Utilização de Teste Estrutural para Aperfeiçoar a Avaliação de um Sistema de Auxílio ao Diagnóstico

Dissertação apresentada ao Programa de Mestrado do Centro Universitário Eurípides de Marília, mantido pela Fundação de Ensino Eurípides Soares da Rocha, para obtenção do título de Mestre em Ciência da Computação

Orientador:
Prof. Dr. Márcio Eduardo Delamaro
Co-orientadora:
Profa. Dra. Fátima L. S. Nunes Marques

MARÍLIA/SP

2005

ANDRADE, Marisa Silvana de Souza.
Utilização de Teste Estrutural para Aperfeiçoar a Avaliação de um Sistema de Auxílio ao Diagnóstico / Marisa Silvana de Souza Andrade;
orientador: Márcio Eduardo Delamaro. Marília/SP:[s.n.], 2005

73 f.

Dissertação (Mestrado em Ciência da Computação) - Centro Universitário Eurípides de Marília - Fundação de Ensino Eurípides Soares da Rocha.

1.CAD. 2.Curva ROC. 3.Critérios de Cobertura Estrutural

CDD: 005.14

MARISA SILVANA DE SOUZA ANDRADE

Utilização de Teste Estrutural para Aperfeiçoar a Avaliação de um Sistema de Auxílio ao Diagnóstico

Banca Examinadora da dissertação apresentada ao Programa de Mestrado em Ciência da Computação do Centro Universitário Eurípides de Marília - UNIVEM, mantido pela Fundação de Ensino Eurípides Soares da Rocha, para obtenção do título de Mestre em Ciência da Computação.

Resultado: **APROVADO**.

Marília/SP, 08 de Agosto de 2005.

Se não fosse pelo amor, amizade, cumplicidade e a dedicação de minha família, a quem muito admiro e tenho orgulho, com certeza este trabalho não seria concluído. Dedico-o a vocês

(meu esposo Nelson, meus filhos André e Thaís, minha mãe Teresinha, minhas irmãs Sandra e Susana, meu irmão Maurício (in memorian)).

AGRADECIMENTOS

A conclusão deste trabalho só foi possível devido ao apoio de várias pessoas, as quais deixo registrado aqui, os meus mais sinceros agradecimentos. Algumas pessoas tornaram-se especiais nesta jornada, seja pelo apoio, pela motivação ou mesmo pelo exemplo dado no dia-a-dia, em especial deixo aqui registrado meu agradecimento:

- A Deus, pela ajuda nos momentos desesperadores da minha vida!
- Fundação de Ensino "Eurípides Soares da Rocha", mantenedora do Centro Universitário Eurípides de Marília, em nome do Magnífico Reitor Dr. Luiz Carlos de Macedo Soares, pelo apoio e incentivo no decorrer deste trabalho.
- Ao prof. Dr. Márcio Eduardo Delamaro, meu orientador, a quem agradeço pela compreensão, pelos ensinamentos. Assim como, também agradeço aos professores da UNIVEM: Fátima, Edmundo, Marcos Mucheroni, Jorge e Beto, pelas constantes palavras de apoio, pelos conselhos e orientações.
- A minha amiga Elizabeth Nardone, pela compreensão e apoio recebido no decorrer deste trabalho.
- Ao coordenador do curso de Mestrado Prof. Dr. Edward David Moreno Ordonez, agradeço pela confiança em mim depositada.
- A minha amiga Lucia Emi Shiraisi Sartori pelo constante apoio e ajuda nas horas mais difíceis.
- Ao meu irmão Mauricio (in memorian), mais que um amigo, um irmão leal. Mais que um irmão, o companheiro presente. Com ele percebi a beleza do ser humano, a ternura e o afeto, independentemente da profissão. A você Mauricio, a minha sempre saudade!
- A minha mamãe Teresinha, que é mensageira de luz e brilho de uma amor incondicional pelos filhos, é permanente benção de Deus, engrandecendo a nossa familia, obrigada mamãe por você existir.
- Ao meu esposo Nelson, meu amigo, companheiro inseparável. Que o nosso viver possa ser transformado continuamente, justificado no significado das nossas ações, de nossos sentimentos. E que as nossas ações sejam sempre guiadas por Deus. Que tenhamos sempre em mente: temos o compromisso do encontro no amanhã... na

responsabilidade do plantio das sementes... os nossos filhos!. O seu caráter tem brilho e valor inestimável e me incentivam a sempre continuar. São essas qualidades que DEUS lhe deu e o faz admirado e amado por todos. E nos faz feliz.

- As minhas irmãs Sandra e Susana, nesta familia, onde as queixas são tão bem acolhidas quanto os elogios... Onde as lágrimas são compartilhadas, igualmente, como as alegrias, sabemos que podemos encontrar sempre um ombro para nos apoiar, mãos para segurar e corações prontos para nos receber com amor... muito obrigada!

- André e Thaís, meus filhos, nossa união e nossa familia, são razões suficientes para estarmos sempre juntos, por toda a vida. São razões para acreditarmos no futuro, obrigada por compreender a minha ausência e me ajudar a continuar esse trabalho.

- Ao meu amigo Leonardo Souza e Silva que mesmo estando a alguns quilômetros, sempre me apoiou.

- Aos meus amigos de trabalho, que sempre me incentivaram no longo desses anos.

“Ando devagar, porque já tive pressa, levo esse sorriso, porque já chorei demais ”

“Hoje me sinto mais forte, mais feliz quem sabe, só levo a certeza”

“De que muito pouco eu sei, eu nada sei...”

Almir Sater

ANDRADE, Marisa Silvana de Souza. **Utilização de Teste Estrutural para Aperfeiçoar a Avaliação de um Sistema de Auxílio ao Diagnóstico**. 2005. 73 f. Dissertação (Mestrado em Ciência da Computação) - Centro Universitário Eurípides de Marília, Fundação de Ensino Eurípides Soares da Rocha, Marília/SP, 2005.

RESUMO

Tradicionalmente, utiliza-se a curva ROC na avaliação de esquemas de diagnóstico auxiliado por computador (CAD). Este método permite que se avalie o esquema através da contraposição do número de resultados falso-positivos contra o número de resultados verdadeiro-positivos, para diferentes configurações do sistema. A escolha dos dados utilizados na geração da curva ROC pode distorcer os resultados obtidos.

Não existe um método padrão para se selecionar os dados de entrada para gerar a curva e isso pode gerar problemas. Dependendo do conjunto de dados de entrada pode-se gerar uma curva ROC muito boa e uma curva ROC muito ruim.

Buscando melhorar a avaliação dos sistemas CADs utilizando-se a curva ROC, este trabalho propõe um método que utiliza a técnica de cobertura estrutural na seleção dos dados de entrada para a geração da curva.

Definiu-se, então, um método que procura normalizar a escolha dos casos de teste através da análise de cobertura. O objetivo do método proposto é evitar possíveis distorções no formato da curva que possam advir da seleção inadequada de casos de teste.

Palavras-chave: CAD. Curva ROC. Critérios de Cobertura Estrutural.

ANDRADE, Marisa Silvana de Souza. **Utilização de Teste Estrutural para Aperfeiçoar a Avaliação de um Sistema de Auxílio ao Diagnóstico**. 2005. 73 f. Dissertação (Mestrado em Ciência da Computação) - Centro Universitário Eurípides de Marília, Fundação de Ensino Eurípides Soares da Rocha, Marília/SP, 2005.

ABSTRACT

Traditionally, the ROC curve is used in CAD (computer-aided diagnosis) schemes evaluation. This method allows to evaluates the scheme by the contraposition of the number of false-positives outcomes against the number of true-positives outcomes, for different configurations of the systems. The choice of data used in the generation of the ROC curve can distort the results.

There is not a standard method to select the input data to generate the curve. Depending on the set of input data it is possible to generate a good ROC curve or a bad ROC curve. With the purpose of improving the evaluation of CADs systems using a ROC curve, this text proposes a method that uses the structural technique in the selection of input data to generate the curve.

Thus, it was defined a method that aims at normalize the selection of tests cases using coverage anslysis. The objective of the proposed method is to avoid possible distortions on the shape of the curve that can occur due to the inadequate selection of test cases.

Keywords: CAD. Curve ROC. Coverage Structural Criteria.

Sumário

Lista de Figuras

Lista de Abreviaturas

CAD = *Computer-Aided Diagnosis*

FN = Falso-Negativo

FP = Falso-Positivo

OMS = Organização Mundial de Saúde

RCA = Realce Coeficiente Atenuação

RCC = Realce Curva Característica

RFP = Redução Sinais Falsos-Positivos

ROC = *Receiver Operating Characteristic*

RTH = Realce Transformação Histograma

TAP = Transformação Área-Ponto

VN = Verdadeiro-Negativo

VP = Verdadeiro-Positivo

cra = Com Relação à

Lista de Tabelas

1 Introdução

Esquemas de diagnóstico auxiliados por computador (CAD - *computer-aided diagnosis*) realizam uma análise computadorizada e utilizam técnicas de processamento de imagem, tendo, por exemplo, imagens radiológicas como fonte de dados, gerando resultados que fornecem uma segunda opinião para o profissional da saúde (GIGER, 2000).

É importante ressaltar que o computador é utilizado somente como uma ferramenta para obtenção de informação adicional, sendo o diagnóstico final sempre feito pelo profissional da saúde.

A finalidade do CAD é melhorar a qualidade do diagnóstico, assim como a consistência da interpretação da imagem utilizada como fonte de entrada, mediante o uso da resposta do computador como referência. A resposta do computador pode ser útil, uma vez que o diagnóstico do profissional da saúde é baseado em avaliação subjetiva, estando sujeito a variações intra e interpessoais, bem como à perda de informação devido à natureza sutil da imagem, baixa qualidade da imagem, sobreposição de estruturas, fadiga visual ou distração (MARQUES, 2001).

Os sistemas CAD elaborados por computador oferecem inúmeras vantagens com relação aos humanos: são superiores em análise multidimensional, sugerem decisões consistentes, apesar de alguma variação no dado fornecido, e não estão sujeitos ao cansaço. Porém, pelo menos com a tecnologia atual, a capacidade inigualável do raciocínio de um observador humano não pode ser contestada por um sistema de computador (DOI; GIGER; HOFFMANN, 1999).

Uma das formas mais utilizadas na avaliação do desempenho dos esquemas CAD é a curva ROC (*Receiver Operating Characteristic*) que se tornou padrão obrigatório na avaliação de observadores, devido ao seu caráter gráfico. As curvas ROC são obtidas com base nos resultados que o CAD produz, levando-se em conta a experiência do observador (RODRIGUES; FRERE, 2000).

Um dos desafios no desenvolvimento de tais esquemas é a sua avaliação, pois os resultados podem variar em função do conjunto de dados de entrada. Para minimizar esta questão são propostas soluções como a disponibilização de bases de imagens comuns e a padronização de critérios e avaliação, como as curvas ROC (NISHIKAWA et al., 1994).

O método que utiliza a curva ROC consiste em contrapor a quantidade de acertos e erros do sistema em avaliação, de forma que uma representação gráfica possa fornecer uma medida de eficiência. Com isso, obtém-se a avaliação do desempenho de um CAD, através da apresentação da relação entre sensibilidade e especificidade. Tradicionalmente, durante a avaliação de sistemas não é empregada uma forma de se selecionar casos adequados de teste para a geração da curva. Geralmente, são considerados todos os casos disponíveis ou, ainda, selecionam-se casos aleatoriamente, o que pode causar desvios na medição.

Foram desenvolvidas várias técnicas de teste, dentre elas temos a técnica estrutural, que se baseia no código fonte do programa de um software. Tal técnica apresenta vários critérios de teste que serão melhores abordados no capítulo 2. Foram utilizados critérios de teste estrutural para gerar casos de teste necessários para os testes realizados neste trabalho.

1.1 Motivação

A seleção aleatória de casos de teste pode produzir resultados bastante díspares na avaliação de um CAD. Não existe um método padrão para se selecionar os dados de entrada para gerar a curva e isso pode gerar problemas. Dependendo do conjunto de dados de entrada pode-se gerar uma curva ROC muito boa e uma curva ROC muito ruim. A idéia básica é que devem ser selecionados casos de teste que tenham características distintas entre si.

1.2 Objetivo da pesquisa

O objetivo final deste trabalho é propor um método que encontre uma forma canônica para a geração da curva ROC, fazendo com que esta reflita o comportamento do sistema, independente do conjunto de dados utilizado na sua geração.

Para atingir o objetivo, foram utilizados critérios de teste estrutural para a seleção de casos de teste usados para gerar a curva ROC. O objetivo do método é evitar possíveis distorções no formato da curva que possam advir da seleção inadequada de casos de teste.

1.3 Estrutura do Trabalho

O presente trabalho constitui-se de 5 capítulos. No capítulo 1 foi apresentada uma intodução ao tema e os objetivos da pesquisa. O capítulo 2 mostra a importância do teste de software, e se concentra na apresentação da técnica estrutural. O capítulo 3 apresenta conceitos sobre CAD e alguns exemplos de Sistema de Auxílio ao Diagnóstico e conceitos sobre a Curva ROC que é o método padrão de avaliação dos CADs. O capítulo 4 apresenta uma técnica para seleção de dados de teste e geração da curva ROC e um estudo de caso que avalia a eficácia do método, comparando-o com a seleção aleatória os resultados dos testes realizados. E, por fim o trabalho é concluído no capítulo 5 que apresenta a conclusão e propostas para sua continuidade.

Ao final, estão disponibilizadas as referências bibliográficas citadas ao longo do texto desta dissertação.

2 Teste Estrutural

Neste capítulo é feita uma breve revisão sobre a técnica de teste estrutural. O objetivo é apresentar as principais características dos critérios de teste estrutural e o tipo de análise estática necessária para a aplicação desses critérios.

A motivação para estudo da técnica é a sua utilização na definição do método de seleção de casos de teste, apresentado no capítulo 4.

2.1 Atividade de Teste

O principal objetivo da Engenharia de Software é produzir software com baixo custo e alta qualidade. E para que isso ocorra é necessário realizar teste de software. A atividade de teste tem como objetivo aumentar a confiabilidade e melhorar a qualidade dos produtos desenvolvidos (MALDONADO et al., 2000).

O processo de desenvolvimento de software envolve uma série de atividades humanas na qual a possibilidade de inclusão de falhas e erros, no projeto, pode ser enorme e ocasionar um grande dano. Mesmo com o uso de métodos, técnicas e ferramentas de desenvolvimento, ainda podem permanecer erros no produto, os quais podem ocasionar dificuldades e custos adicionais para o seu aperfeiçoamento. Assim, a atividade de teste é de vital importância na garantia da qualidade do software, representando a última revisão da especificação, projeto e codificação (PRESSMAN, 1995).

A atividade de teste consiste em executar um programa com o objetivo de encontrar erros, através dos seguintes passos:

1-construção de conjunto de casos de teste.

2-execução de um programa com esse conjunto de casos de teste.

3-análise do comportamento do programa para determinar se o mesmo é o esperado.

Esses passos se repetem até que se consiga achar um bom caso de teste. Um bom caso de teste pode ser definido como aquele que tem uma alta probabilidade de revelar defeitos no software e um caso de teste bem sucedido é aquele capaz de revelar erros ainda não descobertos (MALDONADO et al., 2000) (RAPPS; WEYUKER, 1985).

Os testes podem sem conduzidos em três níveis (PRESSMAN, 1995):

Teste de unidade: concentra esforços na menor unidade do projeto de software (módulo), ou seja, procura intensificar erros de lógica e de implementação em cada módulo do software separadamente;

Teste de integração: é uma técnica sistemática para a construção da estrutura de programa, realizando-se, ao mesmo tempo, teste para descobrir erros associados às interfaces. O objetivo é, a partir dos módulos testados no nível de unidade, construir a estrutura de programa que foi determinada pelo projeto;

Teste de sistema: é, na verdade, uma série de diferentes testes cujo objetivo é verificar se todos os elementos do sistema foram adequadamente integrados e se realizam as funções atribuídas.

Em geral, não se consegue, por meio de testes, provar que um programa está correto; portanto, testar, contribui no sentido de aumentar a confiança de que o software executa corretamente as funções especificadas (MALDONADO et al., 2000).

2.2 Técnicas de Teste de Software

Uma das principais atividades em teste de software é o projeto e avaliação de casos de teste. Utiliza-se um conjunto de técnicas, métodos e critérios. As técnicas podem ser classificadas, basicamente em: funcional, estrutural e baseada em erros.

A técnica funcional orienta a seleção de casos de teste apoiada na especificação do software, enquanto a técnica estrutural apoia-se essencialmente na implementação. A técnica baseada em erros utiliza informações de erros típicos cometidos no processo de desenvolvimento de software para derivar requisitos de teste (MALDONADO et al., 2000).

Para assegurar um teste de melhor qualidade, essas técnicas devem ser aplicadas em conjunto, pois, de acordo com (MALDONADO et al., 1998), nenhuma técnica de teste é completa, e nenhuma delas sozinha é suficiente para garantir a qualidade da atividade de teste.

Devido à diversidade de critérios de teste existentes e à necessidade desses critérios serem aplicados em conjunto, surge a questão de qual estratégia de teste utilizar, ou seja, como escolher os critérios de teste, de forma que as vantagens de cada um desses critérios sejam combinadas para se obter resultado com o menor custo.

O programa deveria ser executado com todos os valores possíveis do domínio de entrada. Sabe-se, entretanto, que o teste exaustivo é impraticável devido a restrições de tempo e custo para realizá-lo. Dessa forma, é necessário determinar quais casos de teste utilizar, de modo que a maioria dos erros existentes possam ser encontrados e que o número de casos de teste não seja tão grande a ponto de ser impraticável (MALDONADO et al., 2000) (RAPPS; WEYUKER, 1985).

Através de testes não é possível provar que um programa está correto. Os testes, se conduzidos com critério, contribuem para aumentar a confiança de que o software desempenha as funções especificadas e apresentar algumas características mínimas do ponto de vista da qualidade do produto.

Das técnicas de verificação e validação, o teste de software é a atividade mais utilizada e também a mais onerosa, podendo, em alguns casos, consumir 40% dos custos de desenvolvimento do software (PRESSMAN, 1995). Buscando sistematizar a atividade de teste e reduzir os custos associados, técnicas e critérios de teste têm sido desenvolvidos, contribuindo para o aprimoramento dessa atividade e, conseqüentemente, do produto em desenvolvimento.

Alguns fatores básicos são necessários para comparar os critérios de teste:

O **Custo** - esforço necessário para que o critério seja usado.

A **Eficácia** - capacidade que um critério possui de detectar um maior número de erros em relação a outro.

A **Dificuldade de Satisfação** - probabilidade de satisfazer-se um critério tendo satisfeito outro (MALDONADO et al., 2000).

Técnicas e critérios de teste têm sido elaborados com o objetivo de fornecer uma maneira rigorosa e sistemática para selecionar um subconjunto do domínio de entrada e, ainda assim, ser eficiente para apresentar os erros existentes, preocupando-se com o tempo e custo associados a um projeto de software.

Segundo Weyuker (1980), um critério de teste permite avaliar os testes realizados. Entretanto, mesmo com a obtenção de conjuntos de casos de testes adequados a um determinado critério, não é possível afirmar que o programa em teste está correto. O propósito dos testes é descobrir os erros que porventura existam e não demonstrar que o programa em teste está correto, porém a obtenção de conjuntos de casos de teste que satisfazem um bom critério de teste em relação a um Programa P fornecem fortes evidências de que P está quase correto (WEYUKER, 1980).

Portanto, critério de teste é um conjunto de condições que devem ser satisfeitas para que a atividade de teste seja realizada com sucesso. Este conjunto de condições no teste estrutural, pode ser, em termos práticos, o conjunto de nós ou de arcos do grafo do programa, determinadas associações entre definições e uso de variáveis, determinados tipos de caminhos no programa, entre outros. Independente do critério considerado, tem-se em qualquer caso, um conjunto de elementos requeridos pelo critério. O conjunto de elementos requeridos no teste estrutural de software consiste no conjunto de componentes estruturais do programa, selecionados pelo critério que o testador deve exercitar para satisfazer tal critério e, com isso, encerrar com sucesso sua tarefa. Critérios existentes para o teste de software podem ser usados como critérios de parada, critérios de seleção de dados de teste e critérios de adequação de dados de teste.

Um critério de parada estabelece parâmetros que devem ser satisfeitos para que se encerre o processo de teste, fornecendo, portanto, uma condição de suficiência para o fim do processo de teste. Um critério de parada pode ser baseado em tempo, cobertura, número de erros restantes (probabilisticamente estimados), entre outros.

Um critério de seleção de dados de teste estabelece parâmetros para a escolha dos dados de teste de um determinado domínio de entrada. Um critério de adequação de dados de teste estabelece parâmetros para que se possa considerar um conjunto de dados de teste apropriado. É um meio de avaliar o quanto um conjunto de dados de teste utilizado satisfaz as condições impostas pelo critério. Existe uma correspondência entre a seleção de dados de teste e a adequação dos mesmos sob algum critério. Para um dado critério de adequação, existe um critério de seleção que busca satisfazê-lo. Nada impede,

entretanto, que se utilize um determinado critério de teste para selecionar os dados e um outro para medir a adequação dos mesmos.

Satisfazer um critério de teste significa exercitar todos os elementos requeridos pelo critério para um dado programa, mediante a execução de casos de testes. O conjunto de casos de teste que exercitou todos esses elementos é considerado adequado ao critério. Ou seja, um conjunto de casos de teste T é C-adequado a um programa P, se e somente se T satisfaz o critério C para P. Em alguns casos, diz-se que um critério é exigente, ou ainda, que um critério é conservador. Isto significa que tal critério exige mais elementos requeridos, ou requer elementos mais difíceis de serem exercitados que outros critérios (que, comparados ao critério exigente, são considerados mais fracos). A utilização de critérios de seleção/adequação de dados de teste é importante, pois aumenta a probabilidade de encontrar dados mais eficazes, ou seja, que tenham maior capacidade de causar falhas e/ou detectar defeitos. A idéia de eficácia de um critério de teste está relacionada à habilidade do critério em levar o testador a selecionar dados que tenham uma boa chance de revelar os defeitos do programa. Intuitivamente, bons critérios são aqueles que, se satisfeitos para um dado programa, garantem um bom nível de confiabilidade (um atributo de qualidade) do mesmo. Portanto, quanto maior a eficácia de um critério, maior confiança pode-se ter de que um programa testado, segundo esse critério, sem apresentar nenhuma falha, esteja realmente correto.

Naturalmente, existe um compromisso entre o custo e o benefício no processo de teste. Idealmente, os critérios devem requerer um número aceitável de elementos, de forma que o teste de grandes módulos seja factível. Por outro lado, devem fazer com que o testador selecione dados que exercitem o programa de forma profunda e completa que boa parte dos defeitos existentes sejam revelados - idealmente todos os defeitos (RAPPS; WEYUKER, 1985).

Dentro do escopo do teste estrutural, vários critérios são utilizados para julgar a adequação do conjunto de casos de teste gerado. Critérios já adotados na prática como, Todos os Ramos/Nós, verificam se os casos de teste fazem com que todos os Ramos/Nós do programa foram visitados e já estão contemplados em ferramentas de teste comerciais (RAPPS; WEYUKER, 1985).

Na próxima seção será dada ênfase ao teste estrutural.

2.3 Teste Estrutural

O teste estrutural enfoca a implementação e a estrutura de controle e de dados do projeto procedimental. A técnica de teste estrutural tem como objetivo caracterizar um conjunto de elementos do software que devem ser exercitados. Ela é baseada no conhecimento da estrutura interna do programa. A estrutura interna do programa pode ser representada por um grafo de fluxo de controle, que é um grafo com um único nó de entrada e um único ou mais nós de saída. Conforme *Weyuker* (1985), o arco, denominado ramo ou aresta, representa um fluxo de controle, ou seja, é a transferência entre blocos. Um nó do grafo representa uma seqüência máxima de comandos tal que, se o primeiro comando for executado, então, necessariamente, todos os demais subseqüentes no bloco, também serão executados. (RAPPS; WEYUKER, 1985).

A técnica estrutural baseada em fluxo de controle utiliza informações de controle da execução do programa, tais como comandos e desvios como base para a seleção de dados de teste, de tal forma que determinados tipos de estrutura do grafo do programa sejam exercitados (MALDONADO et al., 2000) (MALDONADO et al., 1998).

A técnica estrutural baseada em fluxo de dados utiliza informações do fluxo dos dados existente no programa, visando identificar atribuições e utilizações das variáveis através do programa, gerando componentes elementares a serem exercitados pelo testador (MALDONADO et al., 2000) (MALDONADO et al., 1998).

Nó de entrada é o nó correspondente ao bloco cujo primeiro comando é também o primeiro comando do programa. Nó de saída é um nó que não possui sucessor. Um caminho é uma seqüência finita de nós (n_1, n_2,..,n_k), k$\geq$2 tal que exista um arco de n_i para n_i+1, para i = 1,2,3,...,k-1 (RAPPS; WEYUKER, 1985).

Um caminho simples é tal que todos os nós, exceto possivelmente o primeiro e o último, são distintos. Um caminho é livre de laço se todos os nós pertencentes a ele são distintos (MALDONADO et al., 2000)(MALDONADO et al., 1998).

Um caminho completo é um caminho cujo nó inicial é um nó de entrada e o nó final é um nó de saída.

A análise de fluxo de dados focaliza as ocorrências das variáveis nos programas, que podem ser: definição ou uso de variável. Uma definição de variável ocorre quando um valor é armazenado em uma posição de memória. Segundo (MALDONADO, 1991),

em geral, uma ocorrência de variável é uma definição, se ela estiver:

i) no lado esquerdo de um comando de atribuição;

ii) em um comando de entrada;

iii) em chamadas de procedimentos como parâmetro de saída.

A ocorrência de uma variável como uso se dá quando a referência a esta variável não a estiver definindo, ou seja, há uma recuperação de um valor em uma posição de memória associada a esta variável. Um uso pode afetar diretamente uma computação que está sendo realizada ou permitir que o resultado de uma variável definida anteriormente seja analisado. Nestes casos, o uso associado a um nó do grafo do programa analisado pode também afetar diretamente o fluxo de controle do programa; este uso está associado a um arco do grafo (RAPPS; WEYUKER, 1985) (MALDONADO et al., 2000).

Um caminho (i, n_1, ..., n_m, j), m$\geq$0 que não contenha nenhuma definição de uma dada variável x nos nós n_1,...,n_m é chamado de caminho livre de definição com respeito a x do nó i ao nó j e do nó i ao arco (n_n,j).

Um nó i possui uma definição global de uma variável x, se ocorrer uma definição de x no nó i e existir um caminho livre de definição x de i para algum nó ou para algum arco que contenha um uso da variável x.

Uma definição de x no nó i é uma definição local, se existirem usos-associados ao nó - da variável x somente neste nó i e que sucedam esta definição; e não existir nenhuma definição da mesma variável x entre aquela definição e aqueles usos.

A Figura 2.1 apresenta um programa mostrando os blocos de 1 a 9, os desvios A,B,C,D,E e instruções (comandos) do programa em alguma linguagem de programação (SPOTO; PERES; BUENO, 1995).

Cria-se um grafo def/uso a partir de um grafo de fluxo de controle, associando-se aos nós e arcos, definições e usos de variáveis ocorridos nesses nós ou arcos.

O nó 1 do grafo def/uso (Figura 2.2) é o nó de entrada.

O nó 9 do grafo def/uso (Figura 2.2) é um nó de saída. Podem ser identificados vários tipos de caminhos:

Bloco	Instrução
1	INICIO
1	LE, X,Y
1	IF Y < 0 THEN
2	VAR ← -Y
3	ELSE
3	VAR ← Y
4	K ← 1
5	WHILE VAR <> 0 DO
6	K←K*X
6	VAR← VAR -1
7	IF Y < 0 THEN
8	K←K+1
9	RESP ← K
9	IMPRIME RESP

Figura 2.1: Exemplo de Programa com uma linguagem de programação

Caminho Simples: exemplos de caminhos simples do grafo def/uso (Figura 2.2) (1,2,4,5,7,8,9), (1,3,4,5,7,9).

Caminho livre de laço: exemplo de caminho livre de laço do grafo def/uso (Figura 2.2) (1,2,4,5,7,8,9), (1,3,4,5,7,9).

Caminho Completo: exemplo de caminho completo do grafo def/uso (Figura 2.2) (1,3,4,5,7,9), (1,2,4,5,6,5,7,9), (1,3,4,5,6,5,7,9)

Segundo o grafo def/uso (Figura 2.2), as variáveis X e Y possuem definição apenas no nó 1;a variável VAR possui definições nos nós 2, 3 e 6; a variável K nos nós 4, 6 e 8. A variável X possui uso apenas no nó 6; Y possui uso nos nós 2 e 3 e nos arcos (1,2), (1,3), (7,8) e (7,9); a variável K possui uso nos nós 6, 8 e 9; e VAR possui uso no nó 6 e nos arcos (5,6) e (5,7).

Alguns caminhos livres de definição c.r.a variável K, segundo o grafo def/uso (Figura 2.2) são (1,2,4), (1,3,4), (4,5,6), (4,5,7,8), (6,5,7,8).

São elementos requeridos por critérios de teste estrutural os Nós e Arcos do grafo

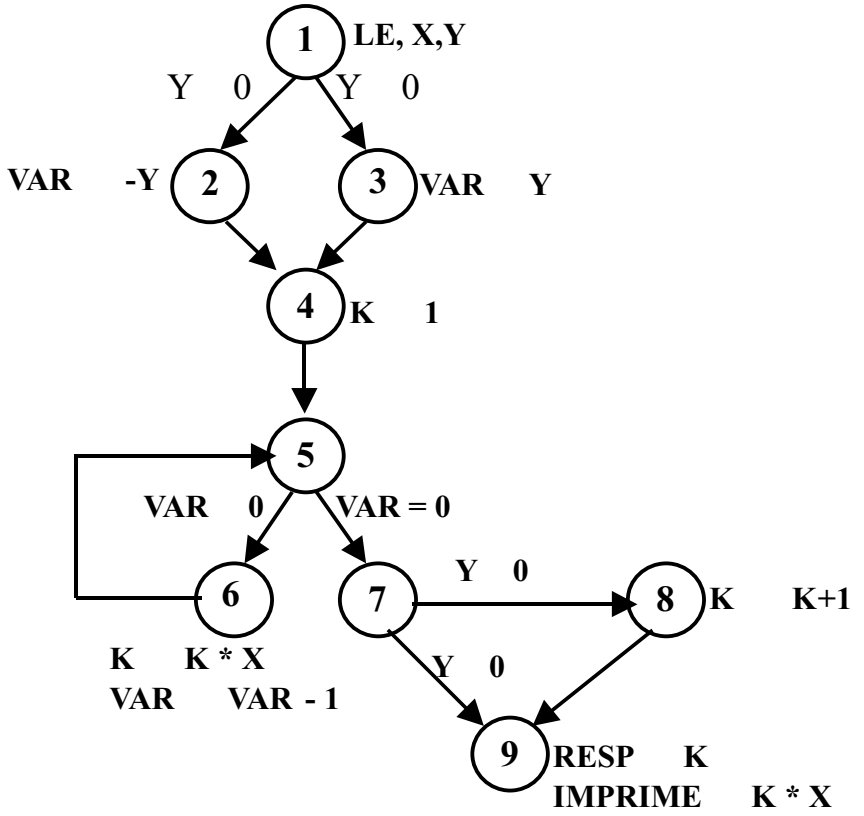


Figura 2.2: Grafo def/uso que ilustra o programa da Figura 2.1

do programa, ou certos caminhos e certas associações presentes nos programas, os quais exigem que estes componentes sejam cobertos por caminhos executados por dados de entrada. Quando se utiliza de técnicas estruturais, o objetivo é caracterizar um conjunto de componentes do software que devem ser exercitados, através de dados de teste. Um caminho completo é executável se existir alguma associação de valores às variáveis de entrada, variáveis globais e parâmetros do programa que cause a execução do caminho. Um caminho é executável, se ele é um subcaminho de algum caminho completo executável.

Somente existe uma **associação** de fluxo de dados se a mesma posição de memória recebe um valor, e este valor, inalterado, é subseqüentemente usado. A vantagem dessa abordagem é que ela captura precisamente os fluxos de dados estabelecidos nas associações definição-uso (RAPPS; WEYUKER, 1985) (MALDONADO et al., 2000).

Uma associação é dita executável, se existir ao menos um caminho executável que a cubra. Um arco ou nó é executável, se residir em algum caminho executável. Não existe algoritmo de propósito geral que decida para programas, se um dado caminho é executável ou não (RAPPS; WEYUKER, 1985), pois a determinação da executabilidade de um caminho também depende da semântica e não somente da sintaxe do programa.

Existem critérios que requerem a execução de elementos que são não executáveis, causando então, a impossibilidade de satisfazer esses critérios; ou seja, se um critério C para um dado programa P exige que se exercite um componente para o qual não existe um caminho executável que o cubra, então não existe nenhum conjunto de casos de teste T, capaz de satisfazer o critério C para um programa P. Devido a este problema, critérios deixam de satisfazer a propriedade da aplicabilidade.

Um critério C satisfaz a propriedade da aplicabilidade, se e somente se, para todo programa P, existir algum conjunto T de casos de teste que é C-adequado para P. Se houver elementos não-executáveis, nenhum conjunto de casos de teste será C-adequado para P; portanto, o critério não será aplicável.

2.4 Critérios de Teste Estrutural

Os critérios de teste estrutural, também conhecido como teste de caixa branca, baseiam-se no conhecimento da estrutura interna da implementação. Os critérios de teste estrutural são, em geral, classificados em: Critérios baseados em Fluxo de Controle, Critérios baseados em Fluxo de Dados.

2.4.1 Critérios Baseados em Fluxo de Controle

Utilizam apenas características de controle da execução do programa, como comandos e desvios, para determinar quais estruturas são necessárias. Os critérios mais conhecidos dessa classe são:

Todos-Nós: Exige que a execução do programa passe, ao menos uma vez, em cada vértice do grafo de fluxo, ou seja, que cada comando do programa seja executado pelo menos uma vez.

Todos-Ramos: Requer que cada aresta do grafo, ou seja cada desvio de fluxo de controle do programa, seja exercitada pelo menos uma vez.

Todos-Caminhos: Requer que todos os caminhos possíveis do programa sejam exercitados.

O teste Todos-Nós e o teste Todos-Ramos podem deixar de detectar um grande número de defeitos; a utilização mostra-se viável, mas pouco eficaz, porque grande parte

dos defeitos podem não ser revelados. O teste Todos-Caminhos é geralmente impraticável, visto que programas que contêm laços, possuem um número infinitamente grande de caminhos; além disso, exercitar todos os caminhos não garante que todos os defeitos serão detectados, pois os dados utilizados podem ser não reveladores.

Para algum dado de teste T_1, pertencente ao subdomínio do domínio de entrada que executa um dado caminho P, o defeito é revelado, enquanto que, para o dado de teste T_2, pertencente ao mesmo subdomínio, o defeito não é sensibilizado. Neste caso, ao selecionar-se T_2, tem-se um dado de teste não revelador. Tais problemas motivaram a introdução dos critérios de teste baseados em fluxo de dados.

2.4.2 Critérios Baseados em Fluxo de Dados

Uma motivação para a introdução dos critérios baseados na análise de fluxo de dados foi a indicação de que, mesmo para programas pequenos, o teste baseado unicamente no fluxo de controle pode não ser eficaz para revelar a presença até mesmo de erros simples e triviais. A introdução dessa classe de critérios procura fornecer uma hierarquia entre os critérios Todos-Ramos e Todos-Caminhos, procurando tornar o teste mais rigoroso, já que o teste Todos-Caminhos é, em geral, impraticável.

Em Weyuker (1985) pondera-se que não se pode acreditar na exatidão de uma computação, se o resultado desta computação nunca foi utilizado; afirmação que apóia a idéia básica dos critérios baseados em fluxo de dados.

Rapps e Weyuker (1985) propuseram o Grafo Def-Uso, que consiste em uma extensão do grafo de programa. No grafo são adicionadas informações a respeito do fluxo de dados do programa, caracterizando associações entre pontos do programa onde são atribuídos um valor a uma variável (chamado de definição da variável) e pontos onde esse valor é utilizado (chamado de referência ou uso de variável). Os requisitos de teste são determinados com base em tais associações. A Figura 2.2 ilustra o Grafo Def-Uso do programa. Conforme o modelo de fluxo de dados definido em (MALDONADO, 1991), uma definição de variável ocorre quando um valor é armazenado em uma posição da memória, como já citado na sessão de teste estrutural.

Dois tipos de usos são distinguidos: c-uso e p-uso. O primeiro tipo, o c-uso afeta diretamente uma computação sendo realizada ou permite que o resultado de uma definição anterior possa ser observado.Vale ressaltar que quando não existir nenhuma

definição de variável precedente a um c-uso no mesmo bloco tem se um c-uso global, caso contrário tem-se um c-uso local (RAPPS; WEYUKER, 1985). O segundo tipo, o p-uso afeta diretamente o fluxo de controle do programa, ou seja, a variável é usada como predicado em um comando de decisão ou de repetição (RAPPS; WEYUKER, 1985).

O critério mais básico dos critérios baseados em análise de fluxo de dados é o critério Todas-Definições e faz parte da família de critérios definidos por Rapps e Weyuker (RAPPS; WEYUKER, 1985). Entre os critérios dessa família o critério Todos-Usos tem sido um dos mais utilizados e investigados.

Todas-Definições: Requer que cada definição de variável seja exercitada pelo menos uma vez, não importa se por um c-uso ou por um p-uso.

Todos-Usos: Requer que todas as associações entre uma definição de variável e seus subseqüentes usos (c-usos e p-usos) sejam exercitadas pelos casos de teste, através de pelo menos um caminho livre de definição, ou seja, um caminho onde a variável não é redefinida.

A maior parte dos critérios baseados em fluxo de dados para requerer um determinado elemento (caminho, associação, etc.) exige a ocorrência explícita de um uso de variável e não garante, necessariamente , a inclusão dos critérios Todos-Ramos na presença de caminhos não executáveis, presentes na maioria dos programas.

A seguir é definida uma série de conceitos estabelecidos a partir do grafo-def/uso do programa, conforme (RAPPS; WEYUKER, 1985). Para tanto, considera-se i,j,k como sendo nós distintos do grafo do programa:

def(i) - é o conjunto de variáveis que possuem uma definição global no nó i.

c-uso(i) - é o conjunto de variáveis que possuem um c-uso global no nó i.

p-uso(i,j) - é o conjunto de variáveis para as quais a aresta (i,j) contém um p-uso.

dcu(x,i) - seja $x \in$ def(i), é o conjunto de todos os nós j tal que $x \in$ c-uso(j) e para o qual existe um caminho livre de definição com respeito a x de i até j.

dpu(x,i) - seja $x \in$ def(i), é o conjunto de todas as arestas (j,k) tal que $x \in$ p-uso(j,k) e para o qual existe um caminho livre de definição com respeito a x de i até j.

du-caminho - um caminho (n_1, ..., n_j, n_k) é um du-caminho com respeito a x se

n_1 tiver uma definição global x e: 1- n_k tiver um c-uso de x e (n_1, ..., n_j, n_k) é um caminho simples livre de definição com respeito a x; ou 2- (n_j,n_k) tem um c-uso de x e (n_1, ..., n_j) é um caminho livre de laço e livre de definição com respeito a x. Um du-caminho com respeito a x é executável, se existir algum conjunto de valores de entrada capaz de executar um caminho completo que o inclua.

Conforme Weyuker (1985), dado G como um grafo def/uso, e P como um conjunto de caminhos completos de G então:

P satisfaz o critério Todos-Nós, se todo nó de G está incluído em P.

P satisfaz o critério Todos-Arcos, se toda aresta de G está incluída em P.

P satisfaz o critério Todos-Caminhos, se todos os caminhos completos de G são incluídos em P.

P satifaz o critério Todas-Definições, se para todo nó i de G e todo $x \in$ def(i), P inclui um caminho livre de definição com respeito a x de i para todos elementos de dcu(x,i) ou dpu(x,i).

P satisfaz o critério Todos-p-usos, se para todo nó i e todo $x \in$ def(i),P inclui um caminho livre de definição com respeito a x de i para todos elementos de dpu(x,i).

P satisfaz o critério Todos-c-usos/Alguns-p-usos, se para todo nó i e todo $x \in$ def(i), P inclui algum caminho livre de definição com respeito a x de i para todo nó em dcu(x,i); se dcu(x,y) é vazio, então P deve incluir um caminho livre de definição com respeito a x de i para alguma aresta contida em dpu(x,i). Este critério requer que todo c-uso de uma variável x, definida em nó i, deve ser incluído em algum caminho de P.

P satisfaz o critério Todos-p-usos/Alguns-c-usos, se para todo nó i e todo $x \in$ def(i), P inclui o caminho livre de definição com respeito a x de i para todos elementos de dpu(x,i); se dpu(x,y) é vazio, então P deve incluir um caminho livre de definição com respeito a x de i para algum nó contido em dcu(x,i). Este critério requer toda definição que é constantemente usada em algum caminho de P.

P satisfaz o critério Todos-Usos se para todo nó i e para todo $x \in$ def(i), P inclui um caminho livre de definição com respeito a x do nó i até cada um dos elementos de dcu(x,i) e até cada elemento de dpu(x,i).

P satisfaz o critério Todos-du-caminhos se todo nó i e para toda variável x E

def(i), P inclui todos os du-caminhos com respeito a x.

Critério baseado em Fluxo de Dados: Todos Usos, Todos c-usos, Todos p-usos, Todos potenciais-usos (PU), Todos potenciais-usos/du (PUDU), Todos potenciais-du-caminhos(PDU).

A relação de Inclusão é uma importante propriedade dos critérios, sendo utilizada para avaliá-los do ponto de vista teórico. O critério Todos-Arcos por exemplo, inclui o critério Todos-Nós, ou seja qualquer conjunto de casos de teste que satisfaz o critério Todos-Arcos também satisfaz o critério Todos-Nós. Quando não é possível estabelecer essa ordem de inclusão para dois critérios, como é o caso de Todas-Definições e Todos-Arcos, diz-se que tais critérios são incomparáveis (RAPPS; WEYUKER, 1985).

2.5 Cobertura de Elementos Requeridos dos Critérios de Teste

O termo cobertura por si só não tem significado e só faz sentido se estiver relacionado com um critério de teste. Um determinado valor de cobertura expressa a idéia da proporção de elementos requeridos que já foram exercitados em relação ao número total de elementos requeridos do critério de teste. O valor de 100% de cobertura indica que todos os elementos requeridos do critério foram exercitados ao se aplicar um conjunto de dados de teste (CRESPO, 1997).

Torna-se de extrema importância, o controle do que está ou não sendo testado no código. Neste contexto, a cobertura de código pode ser uma medida de avaliação do teste para verificar se um conjunto de dados de teste consegue exercitar as várias regiões do código. Ou seja, a cobertura é uma métrica para garantir que um lote de dados de teste exercite adequadamente todas as regiões do código. O uso da informação da cobertura permite também avaliar a qualidade dos dados aplicados em relação a um critério de adequação adotado na atividade de teste. Se um dado de teste consegue exercitar uma região do código ainda não exercitada, ou seja, se há um crescimento na cobertura, então conclui-se que o dado de teste aplicado é de boa qualidade. Assim, quando há crescimento de cobertura é porque houve exercício de uma nova região no código, aumentando a chance de se revelar novos defeitos (CRESPO, 1997).

A Cobertura dos elementos requeridos exercitados de um critério de teste pode

ser um parâmetro de controle da capacidade do critério de teste para avaliar a qualidade dos dados de teste gerados. Se um critério de teste atingiu 100% de cobertura de seus elementos requeridos, então o progresso desse teste tende a gerar dados que exercitarão os mesmos elementos requeridos já exercitados.

Deve ser ressaltado, neste caso, que a cobertura é um parâmetro que deve ser utilizado com cautela, pois nem sempre um critério de teste atinge 100% de cobertura devido à presença de caminhos não executáveis (CRESPO, 1997).

Acredita-se que, utilizando a informação da cobertura, cresce a confiança de que a atividade de teste revele adequadamente a confiabilidade do software.

2.6 Considerações Finais

Para conseguir produzir software com baixo custo e alta qualidade é necessário realizar teste de software. Em geral, não se consegue, por meio de testes, provar que um programa está correto. A atividade de teste tem como objetivo aumentar a confiabilidade e melhorar a qualidade dos produtos desenvolvidos.

Técnicas e critérios de teste têm sido elaborados com o objetivo de fornecer uma maneira rigorosa e sistemática para selecionar um subconjunto do domínio de entrada e, ainda assim, ser eficiente para apresentar os erros existentes, preocupando-se com o tempo e custo associados a um projeto de software.

A técnica de teste estrutural tem como objetivo caracterizar um conjunto de elementos do software que devem ser exercitados. Ela é baseada no conhecimento da estrutura interna do programa.

O objetivo em estudar e escrever sobre critérios estruturais é que a idéia básica do método proposto nesse trabalho é que devem ser selecionados casos de teste que tenham características distintas entre si. O que se faz é selecionar apenas casos de teste que progressivamente contribuam para a cobertura do critério de teste utilizado. Se um caso de teste não contribui para a aumento da adequação ao critério, ele é descartado por considerar-se que no conjunto total, já existem casos de teste com as suas características.

3 Sistemas de Auxílio ao Diagnóstico

Esquemas de diagnóstico auxiliados por computador (CAD) realizam uma análise computadorizada e utilizam técnicas de processamento de imagem, tendo usualmente imagens médicas como fonte de dados, gerando resultados que fornecem uma segunda opinião para o profissional da saúde (GIGER, 2000).

A proposta do CAD é funcionar como um segundo especialista. É importante ressaltar que o computador é utilizado somente como uma ferramenta para obtenção de informação adicional, sendo o diagnóstico final sempre feito pelo profissional da saúde.

A finalidade do CAD é melhorar a qualidade do diagnóstico, assim como a consistência da interpretação da imagem utilizada como fonte de entrada, mediante o uso da resposta do computador como referência. A resposta do computador pode ser útil, uma vez que o diagnóstico do profissional da saúde é baseado em avaliação subjetiva, estando sujeito a variações intra e interpessoais, bem como perda de informação devido à natureza sutil da imagem, baixa qualidade da imagem, sobreposição de estruturas, fadiga visual ou distração (MARQUES, 2001).

Os sistemas baseados em computadores oferecem inúmeras vantagens com relação aos humanos: são superiores em análise multidimensional, sugerem decisões consistentes, apesar de alguma variação no dado fornecido, e não estão sujeitos ao cansaço. Porém, pelo menos com a tecnologia existente, a capacidade inegualável de raciocínio de um observador humano não pode ser contestada por um sistema baseado em computador (DOI; GIGER; HOFFMANN, 1999). Devido a isso, os esquemas de diagnóstico auxiliados por computador têm sido propostos com o objetivo de auxiliar os profissionais da saúde, fornecendo uma segunda opinião.(DOI; GIGER; HOFFMANN, 1999).

Os resultados obtidos pelos CADs podem ser indicativos de possíveis anomalias,

chamando assim a atenção dos profissionais da saúde para alguma área que possa passar despercebida, auxiliando na tarefa de emitir um laudo. Na prática, é difícil estimar quantitativamente se esta segunda opinião irá melhorar a precisão do diagnóstico final (DOI; GIGER; HOFFMANN, 1999) (YU; GUAN, 2000).

Em geral, os sistemas CAD utilizam-se de técnicas provenientes de duas áreas do conhecimento: visão computacional, que envolve o processamento de imagem para realce, segmentação e extração de atributos, e inteligência artificial, que inclui métodos para seleção de atributos e reconhecimento de padrões (MARQUES, 2001).

Por ter base conceitual genérica e ampla, a idéia do CAD pode ser aplicada a todas as modalidades de obtenção de imagem, incluindo radiografia convencional, tomografia computadorizada, ressonância magnética, ultra-sonografia e medicina nuclear. Pode-se, também, desenvolver esquemas de CAD para todos os tipos de exame de todas as partes do corpo, como crânio, tórax, abdome, ossos e sistema vascular, entre outros. Porém, os principais objetos de pesquisa para o desenvolvimento de sistemas CAD têm sido as áreas de mamografia, para a detecção precoce do câncer de mama; tórax, para a detecção de nódulos pulmonares, lesões intersticiais e pneumotórax; e angiografia, para a análise quantitativa de estenoses e de fluxo sanguíneo (MARQUES, 2001).

Pesquisas em CAD vêm sendo desenvolvidas em diversas áreas da radiologia. Na área de mamografia, vários grupos de pesquisas têm proposto sistemas para detectar, realçar e classificar estruturas de interesse. A análise computadorizada para detecção do câncer de mama está baseada em técnicas de processamento de imagens, como detecção de bordas, realce de contraste e métodos baseados em processamento hierárquico com árvore de decisão. Alguns pesquisadores vêm utilizando morfologia matemática em seus projetos, além de redes neurais artificiais, que são úteis na tarefa de classificar estruturas.

Neste trabalho utilizou-se o CAD para o auxílio diagnóstico de câncer de mama desenvolvido no trabalho de Nunes (2001). Por isso são apresentados neste capítulo alguns conceitos necessários para o conhecimento sobre CAD e uma breve descrição do sistema citado, que utiliza técnicas de processamento de imagens para identificar, realçar e classificar estruturas de interesse.

3.1 Câncer de Mama e Mamografia

O câncer de mama é um dos tipos de câncer que representa em todo o mundo uma das principais causas de morte em mulheres. As estatísticas indicam o aumento de sua freqüência tanto nos países desenvolvidos quanto nos países em desenvolvimento. Segundo a Organização Mundial de Saúde (OMS), nas décadas de 60 e 70 registrou-se um aumento de 10 vezes em suas taxas de incidências ajustadas por idade nos registros de câncer de base populacional de diversos continentes (INCA, 2005).

No Brasil, o câncer de mama é o que mais causa mortes entre as mulheres. E dos 467.440 novos casos de câncer com previsão de serem diagnosticados em 2005, o câncer de mama foi o segundo mais incidente entre a população feminina, sendo responsável por 49.470 novos casos. Estima-se que mais de 150.000 mulheres pelo mundo inteiro morrem de câncer (INCA, 2005).

Há uma evidência que mostra que o diagnóstico precoce e o tratamento de câncer de mama podem aumentar significantemente a chance de sobrevivência de pacientes. Quanto mais cedo é detectado, maior é a chance que um tratamento adequado possa ser prescrito. De todos os métodos disponíveis para detectar o câncer de mama, a mamografia é um método prático e confiável para identificar a doença no seu estágio inicial, pois é capaz de realçar estruturas que podem estar associadas ao início da formação de um tumor. Entre essas estruturas estão as microcalcificações que são depósitos de cálcio de tamanho variável (YU; GUAN, 2000).

Embora a mamografia seja o melhor método para a detecção de câncer de mama, entre 10% e 30% das mulheres que têm a doença e passam pelo exame têm o mamograma negativo. Em um estudo citado por (GIGER; MACMAHON, 1996), a avaliação da imagem não indicou anomalias em aproximadamente dois terços dos mamogramas verificados, mostrando que o radiologista falhou em detectar o câncer que foi restrospectivamente evidente(GIGER; MACMAHON, 1996).

Um aspecto relevante é em relação à composição do tecido mamário, que exerce influência na nitidez da imagem mamográfica. As mulheres mais jovens apresentam mamas com maior quantidade de tecido glandular. Com o passar do tempo, o tecido mamário vai se atrofiando e sendo substituído progressivamente por tecido gorduroso, até se constituir, quase que exclusivamente de gordura e resquícios de tecido glandular na fase pós-menopausa. A densidade da mama é um fator decisivo na qualidade da imagem radiológica

e, conseqüentemente, na exatidão do diagnóstico (NUNES, 2001).

A formação da imagem mamográfica depende dos diferentes graus de densidade dos diversos tecidos mamários. A leitura manual, que é o procedimento de exame corrente, é um trabalho intensivo, exigindo tempo e concentração. Há o risco de que os radiologistas possam perder algumas anormalidades sutis (por exemplo, mamas densas que podem esconder lesões devido à qualidade da imagem deficiente, ao cansaço, ou descuido do radiologista)(NISHIKAWA et al., 1994).

3.2 Processamento de Imagem

O uso de imagens médicas vem adquirindo uma importância relevante para o diagnóstico e auxílio na intervenção médica. O tempo despendido para trabalhar com essas imagens, a subjetividade dos atributos extraídos e a necessidade contínua de investigação para o progresso na área têm feito surgir novas técnicas auxiliares no tratamento das imagens (NUNES, 2001).

Existem três níveis de área de processamento de imagem: baixo, intermediário e alto nível.

O primeiro passo após uma imagem digitalizada ter sido adquirida é o pré-processamento, que corresponde ao processamento de baixo nível. Um grande número de técnicas de pré-processamento são disponíveis na prática. Estas incluem realce da imagem, filtragem de ruído, isolamento de regiões, correção geométrica, restauração, reconstrução e segmentação.

A próxima etapa de processamento corresponde ao nível intermediário. A tarefa neste nível consiste em realizar um mapeamento do processamento realizado no baixo nível, para entidades significativas do próximo nível, na etapa de classificação. Sistemas de reconhecimento de padrões geralmente consideram um espaço de características dentro do qual um vetor de observação é mapeado. Com este vetor de observação mapeado, pode-se verificar à qual classe ele pertence. O propósito da extração de características é a redução da quantidade de dados ou redução da dimensionalidade dos dados, obtidas por meio da observação de certas características ou propriedades que distinguem os padrões de entrada (GONZALEZ, 2000) (FACON, 2002).

A última etapa corresponde ao processamento de alto nível. Fazem parte deste

nível de processamento o reconhecimento de objetos e a interpretação da cena. A tarefa de classificação consiste em associar rótulos a um dado objeto, com base, na informação fornecida pelos seus descritores (vetor de características). Um classificador pode ser implementado a partir de técnicas de aprendizagem supervisionada ou não supervisionada. Na categoria supervisionada, os classificadores aprendem com a ajuda de conjuntos de treinamento, enquanto que na categoria não-supervisionada os classificadores aprendem sem a ajuda desses conjuntos (GONZALEZ, 2000) (FACON, 2002).

Uma das principais etapas em uma imagem médica é a segmentação de estruturas de interesse. A segmentação de imagens ocorre no nível intermediário e refere-se ao processo de dividir a imagem em diversas partes (segmentos), permitindo a análise dessas partes isoladamente. Para realizar esta tarefa é de extrema importância um conhecimento prévio da imagem em questão. O processo de segmentar a imagem utiliza modelos matemáticos/geométricos na descrição, análise e classificação das estruturas. Este processo de segmentação tem várias etapas, que vão desde o realce da estrutura até a identificação da imagem. A partir daí, a imagem resultante pode ser usada para outras tarefas importantes em um processo de classificação (GONZALEZ, 2000)(FACON, 2002).

Os algoritmos de segmentação para imagens são geralmente baseados em uma das seguintes propriedades básicas de valores de níveis de cinza: descontinuidade e similaridade. Na primeira categoria, a abordagem é fragmentar a imagem baseada em mudanças bruscas nos níveis de cinza. As principais abordagens da segunda categoria baseiam-se em limiarização, crescimento, divisão e fusão de regiões.

3.3 Avaliação de Esquemas CAD

Um dos desafios do desenvolvimento de esquemas CAD é a sua avaliação. Existe a necessidade de avaliar os resultados obtidos com um conjunto significativo de imagens, a fim de fornecer um embasamento maior para as conclusões.

A escolha de casos clínicos usados para criar e testar um esquema CAD pode afetar os resultados obtidos. Por isso, é difícil avaliar seguramente a exatidão desse tipo de sistema. O desempenho medido dos esquemas CADs é extremamente sensível para a dificuldade dos casos usados para testá-los. A comparação de diferentes esquemas CAD não pode ser válida, a menos que os mesmos casos sejam usados para testá-los (NUNES, 2001).

Existem vários índices de desempenho que podem ser utilizados na avaliação de sistemas de auxílio ao diagnóstico. Uma medida possível e muito utilizada é a análise da característica de resposta do observador, introduzida no contexto das imagens médicas. A análise por curva ROC é fundamentada na teoria de detecção de sinal, tendo como base a idéia de que para qualquer sinal sempre existirá um fundo ruidoso que tem variação aleatória sobre um valor médio. Quando um estímulo está presente, a atividade que ele cria no sistema de obtenção de imagem é adicionada ao ruído existente naquele momento. Este ruído pode estar dentro do próprio sistema ou fazer parte do padrão de entrada. A tarefa do sistema é determinar se o nível de atividade no sistema é devido apenas ao ruído ou resultado de um estímulo adicionado ao ruído. Em sua forma mais simples, a tarefa de diagnosticar consiste na apresentação de imagens contendo ou não uma anormalidade associada (MARQUES, 2001).

As curvas ROC tornaram-se parâmetro obrigatório na avaliação de observadores e sistemas, especialmente na avaliação de esquemas CAD, devido ao seu caráter gráfico que, muitas vezes, pode trazer mais informações qualitativas para a análise final do que a quantidade de informações que eventualmente poderiam ser extraídas através de tabelas e índices (NUNES, 2001).

A partir da seção 3.5 deste capítulo apresentam-se mais profundamente os conceitos sobre a curva ROC.

3.4 Exemplos de Esquemas de Diagnóstico Auxiliado por Computador

Como já mencionado, vários grupos de pesquisa têm o desenvolvimento de sistemas CAD como foco de seus projetos. A seguir, são citados alguns trabalhos que têm se destacado devido aos resultados obtidos.

Nishikawa et al. (1994) desenvolveram um esquema CAD para análise em massas contidas em mamogramas digitalizados e imagens de ultra-som, distinguindo-as em benignas e malignas. O esquema tem as seguintes etapas: identificação manual da massa, extração da lesão, extração automatizada de caracteríticas e aplicação de uma rede neural artificial para dar estimativa da probabilidade de malignidade. Foram executadas uma correção de fundo, equalização do histograma e um crescimento de região para extrair automaticamente a lesão. Para as imagens de ultra-som, um médico delineou as margens

das massas. Nas imagens mamográficas foram analisadas as seguintes características: textura da lesão, nitidez da borda e grau de especularidade. O desempenho dos métodos foi analisado por curvas ROC.

Os pesquisadores do *Kurt Rossman Laboratory*, da Universidade de Chicago, desenvolveram um método computadorizado para a detecção de massas em mamogramas que é baseado nas diferenças de simetria entre as mamas consideradas normais, sendo que essas correspondem a massas em potencial. O método envolve uma técnica não-linear de subtração bilateral que destaca assimetrias. Foram utilizadas técnicas de análise de características baseadas no tamanho, forma e contraste das possíveis massas, com o objetivo de reduzir o número de diagnósticos falsos-positivos. Em um estudo com 154 pares de mamogramas, o computador atingiu um nível de acerto de 85% com 3 ou 4 detecções falso-positivas por imagem(YARUSSO et al., 2000).

O Departamento de Radiologia da Universidade de Michigan desenvolveu um CAD que executa um filtro adaptativo, seguido de detecção de bordas e crescimento de região para detectar as estruturas da mama. Foram extraídas características de textura e morfológicas de cada estrutura detectada. Foram avaliados mamogramas obtidos de 93 pacientes, sendo que os casos foram divididos em 2 conjuntos: Um dos conjuntos continha 97 massas identificadas por um radiologista durante exames clínicos e posterior biópsia. Outro conjunto era constituído de 42 filmes destas mesmas pacientes obtidos entre 1 a 4 anos antes da biópsia. A sensibilidade do algoritmo foi medida caso a caso, usando-se duas visões de cada mama.

O algoritmo teve uma sensibilidade de detecção de 87% com 1,7 falsos-positivos por imagem para o primeiro dos conjuntos, incluindo-se a detecção de 94% para casos malignos. Para o segundo conjunto, houve a sensibilidade de detecção de 73% com 2 falsos-positivos por imagem, incluindo-se a detecção de 67% para casos malignos.

Os pesquisadores Yu e Guan (2000) desenvolveram um sistema CAD que é utilizado para a detecção automática de microcalcificações agrupadas em mamogramas digitalizadas. Em um primeiro passo, pixels relacionados a microcalcificações potenciais são segmentados pelo uso de características mistas, consistindo de características ondulatórias e características do nível de cinza. No segundo passo, microcalcificações individuais são detectadas pelo uso de um conjunto de 31 características extraídas dos objetos de identificação anteriormente. O poder discriminatório dessas características foi analisado, usando redes neurais de regressão. Os classificadores usados nesses dois passos são ambos de redes

neurais de várias camadas. O método foi aplicado para um banco de dados de 40 mamogramas, contendo 105 aglomerações de microcalcificações. As características da curva de operação de resposta foram usadas para avaliar o desempenho. Os resultados mostraram que o sistema apresentou um desempenho de detecção verdadeira, média de 90% com 0,5 falso-positivo por imagem, quando as características mistas foram usadas no primeiro passo e 15 características selecionadas por um método de seleção oposta foram usadas no segundo passo.

3.5 Esquema CAD utilizado no Trabalho

O CAD para diagnóstico de câncer de mama, desenvolvido no trabalho de Nunes (2001) foi utilizado na realização dos experimentos deste trabalho. No capítulo 4, na seção 4.2 foi apresentado de uma maneira detalhada as técnicas utilizadas no tratamento de imagens e a configuração que apresentou os melhores resultados sobre o CAD para diagnóstico de indícios de câncer de mama, desenvolvido no trabalho de Nunes (2001).

3.6 Método de Avaliação

As curvas ROC foram introduzidas na década de 50, tendo como foco principal de atuação a avaliação de desempenho (percentual de acertos versus percentual de erros) no lançamento de mísseis balísticos. Posteriormente, na década de 70, foi observado que o método estatístico das curvas ROC poderia ser utilizado na avaliação da qualidade de imagens, principalmente as imagens da área médica (GALPAROZO; FERNANDEZ, 1998).

O método das curvas ROC é um procedimento estatístico que leva em conta o aspecto subjetivo envolvido em determinado evento (EVANS, 1981). Para se construir uma curva, as entradas (imagens, no caso de esquemas CAD) são divididas em duas categorias: as positivas (possuem uma estrutura ou sinal de interesse) e as negativas (não possuem sinal). As curvas ROC permitem a avaliação do desempenho de um evento, através da apresentação da relação entre a sensibilidade e a especificidade. Define-se sensibilidade e especificidade como duas medidas de precisão de um teste diagnóstico, dadas pelas equações 3.1 e 3.2 (BRAGA, 2000).

$$Sensibilidade = \frac{\text{número de decisões verdadeiras positivas}}{\text{número de casos realmente positivos}} \tag{3.1}$$

$$Especificidade = \frac{\text{número de decisões verdadeiras negativas}}{\text{número de casos realmente negativos}} \quad (3.2)$$

Define-se também, valor de corte, como sendo um valor que pode ser selecionado arbitrariamente entre os valores possíveis para a variável de decisão.

3.7 Categorias para Construir uma Curva ROC

No caso de esquemas CAD, para se contruir uma curva as entradas são divididas em duas categorias: as positivas e as negativas. As imagens positivas possuem pelo menos uma estrutura ou sinal de interesse. As imagens negativas não possuem estruturas, nem sinal de interesse (BRAGA, 2000).

Dentro dessas categorias, são definidos quatro tipos de resultados possíveis em um sistema CAD conforme demonstra a Tabela 3.1.

Tabela 3.1: Tipos de Resultados em um Sistema CAD

Imagens	Tipo
Positivas	Verdadeiro Positivo-VP (foi indicada uma detecção em uma imagem positiva) Falso Negativo-FN (não foi indicada uma detecção em uma imagem positiva)
Negativas	Verdadeiro Negativo-VN (não foi indicada uma detecção em uma imagem negativa) Falso Positivo-FP (foi indicada uma detecção em uma imagem negativa)

O traçado da curva ROC é feito levando-se em conta as probabilidades de ocorrência de verdadeiros-positivos (VP) em função da probabilidade da ocorrência de falsos-positivos (FP) para cada ponto de operação notado na curva, escolhido através de um critério

pré-determinado. O diagnóstico verdadeiro-positivo ocorre quando o sistema identifica e detecta uma estrutura na imagem e ele realmente existe; o diagnóstico falso-positivo é quando o sistema detecta uma estrutura na imagem, mas a região processada não contém anomalias. A área compreendida entre a curva experimental e o eixo horizontal é o principal parâmetro utilizado na comparação de duas curvas. Essa área pode ser interpretada como a capacidade do sistema em prever a saída e, por isso, é utilizada como medida de eficiência do sistema que está sendo testado. A área sob a curva é calculada pela equação 3.3.

$$\mathbf{A_z = VP.P + VN.N} \tag{3.3}$$

onde:

A_z = área sob a curva.

VP = porcentagem de resultados verdadeiros-positivos.

P = porcentagem de casos positivos.

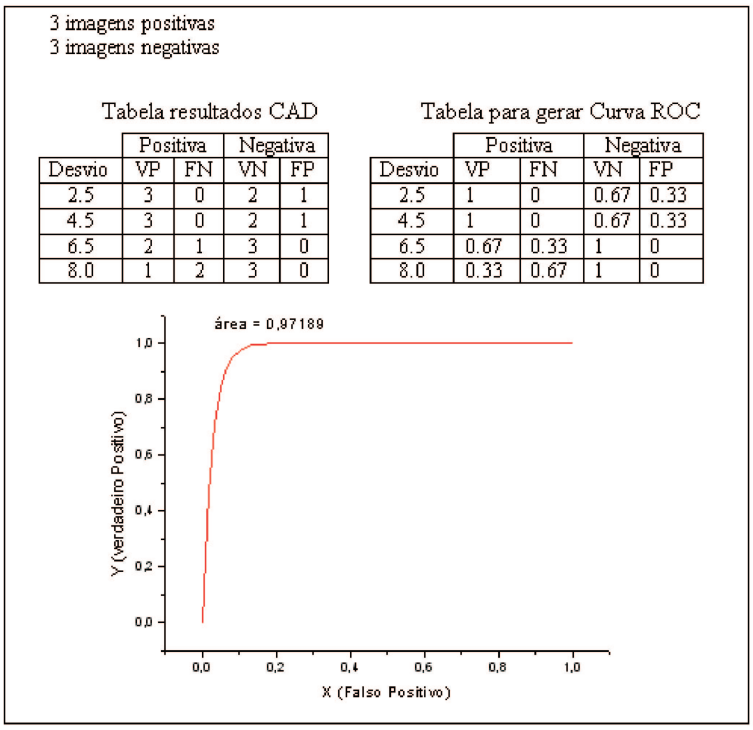
VN = porcentagem de resultados verdadeiros-negativos.

N = porcentagem de casos negativos.

Através da equação 3.3 é possível perceber que quanto mais o valor da área sob a curva A_z se aproximar da unidade , melhor será o comportamento do sistema. Isso significa que o sistema avaliado tem grande percentual de acerto (VP), sendo um sistema de alta sensibilidade (NUNES, 2001).

A curva ROC nos proporciona uma representação global da exatidão diagnóstica. A curva ROC é necessariamente crescente, propriedade que reflete o compromisso existente entre a sensibilidade e especificidade; modifica-se o valor de corte para se obter maior sensibilidade; isto só pode ser feito se diminuir ao mesmo tempo a especificidade. A exatidão do teste aumenta à medida que a curva se desloca da diagonal até o vértice superior esquerdo (GALPAROZO; FERNANDEZ, 1998).

Um primeiro grupo de métodos para construir a curva ROC é constituído pelos chamados métodos não paramétricos. Caracterizam-se por não fazer nenhuma suposição sobre a distribuição dos resultados do teste diagnóstico. O mais simples destes métodos é aquele que se costuma conhecer como empírico, que consiste simplesmente em representar



	Positiva		Negativa	
Desvio	VP	FN	VN	FP
2.5	3	0	2	1
4.5	3	0	2	1
6.5	2	1	3	0
8.0	1	2	3	0

	Positiva		Negativa	
Desvio	VP	FN	VN	FP
2.5	1	0	0.67	0.33
4.5	1	0	0.67	0.33
6.5	0.67	0.33	1	0
8.0	0.33	0.67	1	0

Figura 3.1: Exemplo de Curva ROC

todos os pares (FP, VP), ou seja, todos os pares (especificidade,sensibilidade) para todos os possíveis valores de corte (GALPAROZO; FERNANDEZ, 1998).

Os métodos paramétricos se baseiam em postular um determinado tipo de distribuição para a variável de decisão nas suas populações que se trata distribuir. O modelo mais freqüentemente utilizado é o binormal, que supõe a normalidade das variáveis tanto na população sadia como na enferma (GALPAROZO; FERNANDEZ, 1998).

3.8 Considerações Finais

A finalidade do CAD é melhorar a qualidade do diagnóstico, assim como a consistência da interpretação da imagem utilizada como fonte de entrada. A proposta do CAD é funcionar como uma segunda opinião, podendo chamar a atenção do profissional da saúde para áreas da imagem que talvez passem despercebidas.

Como mencionado, um dos desafios do desenvolvimento de esquemas CAD é a sua avaliação. Existe a necessidade de avaliar os resultados obtidos com um conjunto significativo de imagens, a fim de fornecer um embasamento maior para as conclusões.

O método das curvas ROC é um procedimento estatístico que leva em conta o

aspecto subjetivo envolvido em um determinado evento. As curvas ROC permitem a avaliação do desempenho de um evento, através da apresentação da relação entre a sensibilidade e a especificidade, modifica-se o valor de corte para se obter maior sensibilidade, isto só pode ser feito se diminuir ao mesmo tempo a especificidade.

As curvas ROC tornaram-se parâmetro obrigatório na avaliação de observadores e sistemas, especialmente na avaliação de esquemas CAD, devido ao seu caráter gráfico que, muitas vezes, pode trazer mais informações qualitativas para a análise final.

O desempenho medido dos CADs é extremamente sensível para a dificuldade dos casos usados para testá-los. A comparação de diferentes esquemas CAD não pode ser válida, a menos que os mesmos casos sejam usados para testá-los.

4 Procedimentos Experimentais

Com base em estudos realizados por pesquisadores e também pelos exemplos de esquemas CAD demonstrados no capítulo 3, na seção 3.4, a utilização da curva ROC geralmente não leva em conta um fator importante: a qualidade dos casos de teste utilizados na sua geração. Não existe um método padrão para selecionar os dados de entrada para gerar a curva ROC. Em geral, o que se faz é selecionar um conjunto de imagens, dentre as quais algumas cujo resultado esperado é positivo, outras cujo resultado esperado é negativo. Isso equivale, do ponto de vista da seleção de casos de teste, a utilizar-se um critério de teste funcional como o particionamento em classes de equivalência com apenas duas classes definidas.

Tal método de seleção, além de extremamente fraco, do ponto de vista da eficácia em revelar defeitos, pode também causar distorções na geração da curva ROC. Para se ter uma dimensão do problema, utilizou-se um sistema CAD para o diagnóstico de câncer de mama desenvolvido no trabalho de Nunes (2001) e um conjunto de 200 regiões de interesse extraídas de imagens mamográficas como universo de estudo. Suponha-se que, por decisão de projeto, decidiu-se pela utilização de um subconjunto de 50 imagens para a geração de uma curva ROC para avaliação do sistema. A seleção aleatória desses 50 casos de teste pode produzir resultados bastante díspares, como demonstram as Figura 4.1 e 4.2.

Para gerar a curva da Figura 4.1 foram utilizadas 50 imagens para as quais o sistema obteve próximo de 100% de acerto, resultando em uma área próxima a 1,0, como mostra a curva da Figura 4.1, indicando uma boa performance do sistema.

Para gerar a curva da Figura 4.2 foram utilizadas 50 imagens cujo processamento originou uma curva com uma área aproximada 0,21, que indica péssima performance do sistema, como mostra a curva da Figura 4.2.

Nas Figuras 4.1 e 4.2 pode-se observar duas curvas, geradas a partir de diferentes

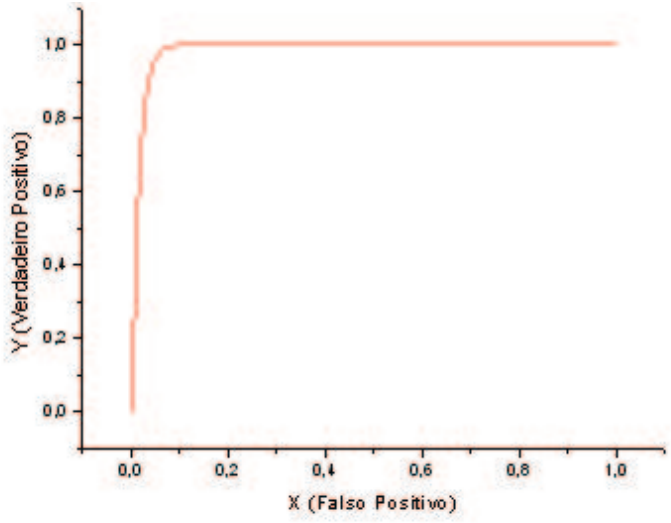


Figura 4.1: Curva ROC - Boa Performance do Sistema

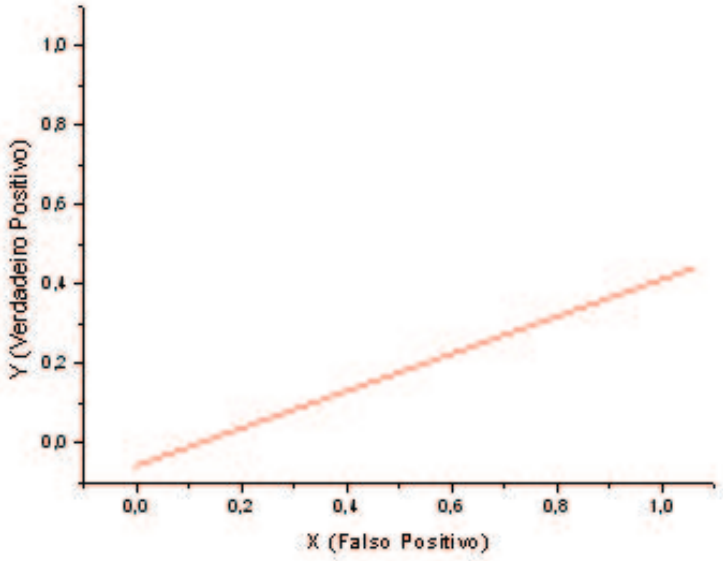


Figura 4.2: Curva ROC - Péssima Performance do Sistema

subconjuntos de imagens. Embora os conjuntos de imagens não tenham sido selecionados aleatoriamente e sim escolhidos, a probabilibade de escolher um ou outro é a mesma.

Pode-se constatar que, dependendo do tipo de conjunto de imagens usado, os resultados obtidos na geração da curva ROC são diferentes, podendo ser obtida uma curva boa ou uma curva ruim.

Buscando melhorar a avaliação dos sistemas CADs utilizando-se a curva ROC, foi proposto um método que utiliza a técnica de cobertura estrutural na seleção dos dados de entrada para a geração da curva. Esse método é apresentado na seção 4.1.

Na seção 4.2 é discutido o experimento que utiliza o método proposto. Na seção 4.3 são apresentados os resultados obtidos e na seção 4.4 são feitas as considerações finais sobre o capítulo.

4.1 Descrição do Método

Com base no exposto na seção anterior, foi desenvolvido um método para a geração da curva ROC que leva em consideração a adequação a critérios de teste, para a seleção de casos de teste. Pela facilidade de entendimento e de utilização e pela disponibilidade de ferramentas, estão sendo utilizados critérios de teste estrutural, baseados no fluxo de controle do programa. Porém, qualquer critério que forneça uma medida objetiva de adequação poderia ser utilizado.

A idéia básica do método é que devem ser selecionados casos de teste que tenham características distintas entre si. No caso, tais características estão relacionadas com a cobertura do código do programa. Por exemplo, tomando-se dois casos de teste t_1 e t_2, que executem exatamente o mesmo caminho dentro do programa avaliado, é provável que se obtenha o mesmo resultado, seja ele o esperado ou não. Generalizando esse conceito, se forem selecionados muitos casos de teste que executam o mesmo caminho, tem-se sempre o mesmo resultado (por exemplo, sempre um Falso Positivo), o que faz a curva ROC tender para um dos extremos, como exibido nas Figuras 4.1 e 4.2.

Definiu-se, então, um método que procura normalizar a escolha dos casos de teste através da análise de cobertura. Os passos propostos para o método são descritos a seguir. Considera-se nessa descrição o programa P como sendo o sistema avaliado, um conjunto finito T de dados de teste (imagens mamográficas, por exemplo) e um critério de teste C, como, por exemplo, o critério todos-os-nós. Os programas P_1, P_2, ..., P_k são diferentes versões de P, considerando-se as diferentes configurações usadas na produção da curva ROC. Por exemplo, para o sistema de Nunes (2001), pode-se considerar P_1 como sendo o sistema ajustado para utilizar o valor 2.5 como valor de ponto de corte (parâmetro utilizado para variar a sensibilidade do sistema durante o procedimento de segmentação da imagem), P_2 ajustado para utilizar 4.5, e assim por diante.

Os passos do método são:

1. É medida C(P_i,T), a adequação de T em relação a cada P_i e ao critério C. O conjunto T é o conjunto de todos os dados de teste disponíveis para execução da avaliação.

Por exemplo, no caso do experimento descrito a seguir, dispunha-se de um conjunto de 200 imagens manograficas (que seria o conjunto T) do qual um subconjunto de imagens deveria ser selecionado para a geração da curva ROC. Inicialmente, é essencial conhecer-se a adequação do conjunto de casos de teste total. Para que se tenha uma avaliação realmente confiável, é necessário que essa adequação seja de 100%. Sabe-se, porém, que, na prática, muitas vezes, não se dispõe de material (como por exemplo, imagens médicas) para que essa meta seja alcançada. Nesse caso, deve-se ter a consciência de que o método pode ser influenciado pela qualidade geral de T.

2. Executa-se o seguinte procedimento, considerando as versões P_1,...,P_k;

$T_i = \{\}$, para $i = 1, 2, ...k$
$T_{total} = \{\}$
repita
  seleciona-se aleatoriamente $t \in T$
  executa-se cada P_i com t
  se $C(P_i, T_i) < C(P_i, T_i \cup \{t\})$ para algum $i = 1, 2, \ldots, k$
    $T_{total} = T_{total} \cup \{t\}$
até que $C(P_i, T_i) = C(P_i, T)$ para todo $i = 1, 2, \ldots, k$

3. Constrói-se a curva ROC utilizando-se o conjunto T_{total}

No passo 2 acima, o que se faz é selecionar apenas casos de teste que progressivamente, contribuam para a cobertura do critério de teste utilizado. Se um caso de teste não contribui para o aumento da adequação ao critério, ele é descartado por considerar-se que no conjunto T_{total}, já existem casos de teste com as suas características. Com isso, pretende-se eliminar redundâncias e obter-se uma curva ROC mais padronizada.

4.2 Descrição do Experimento

Nessa seção será apresentado o experimento que utilizou o sistema CAD para a detecção de indícios do câncer de mama desenvolvido no trabalho de Nunes (2001). Para os testes do sistema foi utilizado um conjunto de 200 regiões de interesse extraídas de imagens mamográficas como universo de estudo, sendo 100 regiões positivas, que continham estruturas de interesse (microcalcificações) e 100 regiões negativas, provenientes de imagem de casos normais.

O sistema é composto por módulos com funções definidas, sendo possível executá-los separadamente. Na execução dos testes para este trabalho foram executados os módulos representados na Figura 4.3.

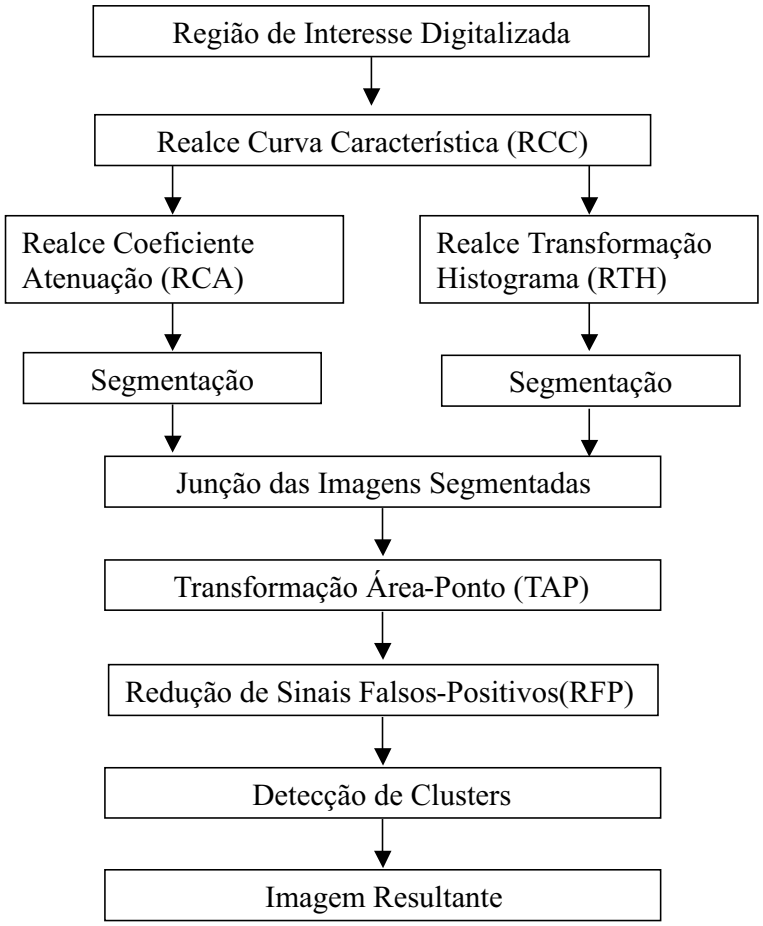


Figura 4.3: Diagrama esquemático da configuração final do esquema CAD - Nunes (2001)

O primeiro passo é aplicar a técnica de Realce da Curva Característica (RCC), sobre a imagem original. Esta técnica consiste em redistribuir os pixels na imagem fazendo uma linearização do histograma, a partir de considerações sobre a curva característica do filme mamográfico utilizado na aquisição da imagem.

A imagem resultante da técnica RCC é submetida a outras duas técnicas de realce de contraste: realce a partir dos coeficientes de atenuação dos materiais (RCA) e realce a partir da modificação do histograma (RTH).

A técnica RCA considera os coeficientes de atenuação dos materiais usados na radiologia para realçar estruturas de interesse na imagem e a técnica RTH divide o histograma em classes para realçar porções de interesse das imagens.

As duas imagens resultantes das técnicas citadas são submetidas ao processo

de segmentação, que produz uma imagem binarizada, na qual os pixels pertencentes às estruturas de interesse são brancos e aqueles pertencentes ao fundo da imagem são pretos.

No passo seguinte as imagens segmentadas são juntadas por um procedimento que soma os pixels das duas imagens binarizadas. A imagem resultante é submetida ao procedimento de transformação área-ponto (TAP) que transforma cada estrutura de interesse em um único pixel. Na seqüência, o procedimento Redução de falsos-positivos (RFP) elimina sinais falsos, a fim de que no último passo os pixels resultantes sejam contados e reunidos, indicando a presença de um aglomerado de microcalcificações.

A Figura 4.4 mostra as imagens resultantes do CAD Nunes (2001) após a aplicação de cada uma das técnicas mencionadas.

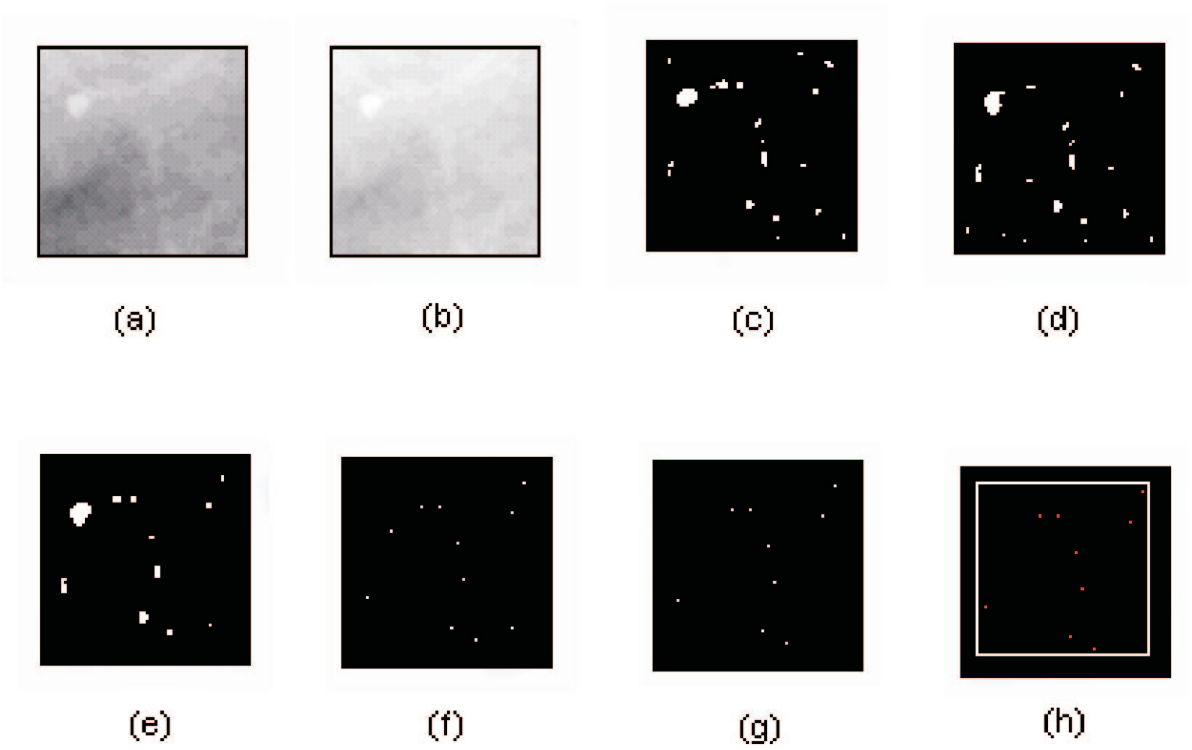


Figura 4.4: Exemplo das técnicas utilizadas (a) imagem original; (b) imagem após RCC; (c) imagem após RCA e segmentação; (d) imagem após RTH e segmentação; (e) junção das imagens; (f) imagem após TAP; (g) imagem após RFP; (h) imagem resultante após detecção de cluster

O sistema CAD utilizado é totalmente parametrizado, o que permite adequar a sua execução ao conjunto de imagem utilizado no processamento. Para a execução deste trabalho foram utilizados os parâmetros padrões, com exceção do parâmetro desvio-padrão, que permite variar o nível de sensibilidade do processamento durante o procedi-

mento de segmentação da imagem. Este é, portanto, o parâmetro usado como ponto de corte para gerar as curvas ROC.

A Figura 4.5 exibe a tela do sistema CAD de Nunes (2001) para fornecimento dos parâmetros no procedimento de segmentação da imagem, onde se pode variar o valor do desvio-padrão.

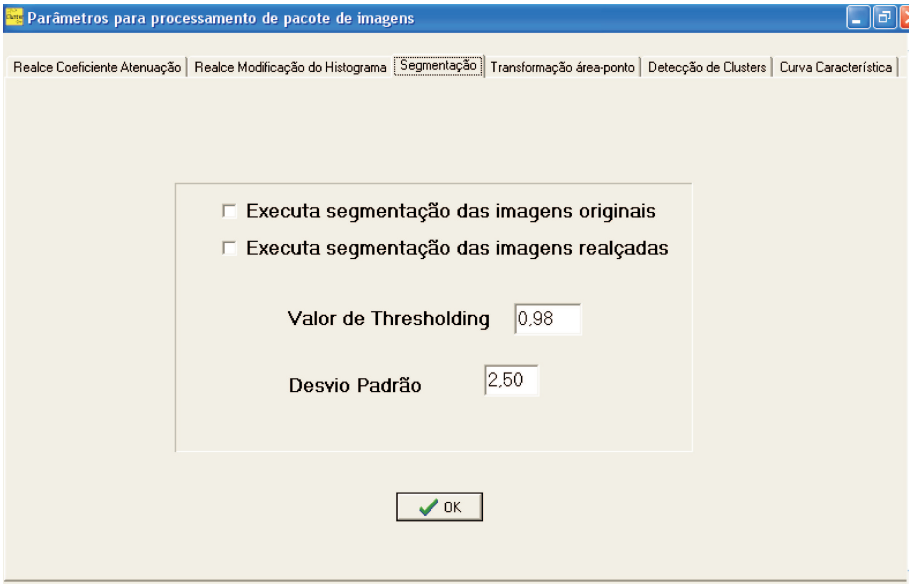


Figura 4.5: Tela do Sistema CAD Nunes (2001)

Para medir a cobertura foi utilizada uma ferramenta da Cyamon Software a *Discover* (CYAMON, 2005) que permite medir a cobertura do código fonte. A técnica utilizada pela ferramenta *Discover* é o teste estrutural utilizando o critério de todos-os-nós, baseado no fluxo de controle do programa.

A Tabela 4.1 apresenta os resultados de cobertura do esquema CAD Nunes (2001) após a aplicação de cada uma das técnicas mencionadas anteriormente para um conjunto de 200 imagens.

Para a geração da curva ROC foi utilizada uma ferramenta da Microcal Origin 6.0 (MICROCAL, 2005).

O objetivo do experimento é melhorar a avaliação do CAD Nunes (2001). Assim, foi realizada uma análise ao método proposto com o propósito de compará-lo a um método que não utiliza a técnica de cobertura. Assim é possível avaliar os resultados com respeito à confiabilidade, a fim de aumentar a precisão dos resultados do ponto-de-vista do desenvolvedor no contexto de um ambiente de estudo acadêmico para dissertação de mestrado.

Tabela 4.1: Cobertura Alcançada em cada Técnica do CAD

Técnicas de Imagens	Cobertura
Realce Curva Característica (RCC)	13%
Realce Coeficiente Atenuação (RCA)+Segmentação	40%
Realce Transformação Histograma (RTH)+Segmentação	48%
Junção das Imagens Segmentadas	51%
Transformação Área Ponto (TAP)	57%
Redução de Sinais Falso Positivo (RFP)	67%
Detecção de Clusters	80%

O procedimento de coleta para a realização do experimento está descrito a seguir.

Primeiramente foi encontrada a porcentagem de cobertura máxima para um universo de 200 imagens o que corresponde ao passo 1 do método proposto, conforme descrito na seção 4.1; a porcentagem máxima encontrada foi de 80%. Isto foi realizado para obter-se como parâmetro a porcentagem de cobertura máxima encontrada para as imagens em questão.

Foram gerados 30 conjuntos de testes utilizando a técnica de análise de cobertura aos quais, daqui para frente, daremos o nome de (T_1,..,T_{30}). Para encontrar cada conjunto foi utilizado o caminho descrito no passo 2 do método.

Quando uma imagem utilizada não apresentava alteração na cobertura com relação ao conjunto, ela era descartada porque a idéia é que devem ser selecionados casos de teste que tenham características distintas entre si.

Para gerar o conjunto de teste T_{total} foram realizados os seguites passos:

1- Gerar uma seqüência aleatória para a seleção das imagens.

2- Executar o sistema CAD para diferentes pontos de corte. Foram usados os pontos de corte (2.5 - 4.5 - 6.5 - 8.0).

3- Utilizar a ferramenta Discover para medir a cobertura utilizando a técnica de teste estrutural.

4- Se a imagem selecionada não apresenta nenhuma alteração de cobertura para algum dos pontos de corte, ela é descartada.

5- repetir até atingir a cobertura máxima, obtida para T (conjunto das 200 ima-

gens).

Foram realizados os 5 passos descritos anteriormente para cada conjunto de teste T_1,..,T_{30}). Em seguida, foram gerados 30 conjuntos de testes, com as mesmas cardinalidades de T_1,..,T_{30}, sem utilizar a técnica de análise de cobertura, ou seja, apenas selecionando-se aleatoriamente as imagens. Daremos o nome de T'_1,..,T'_{30} a esses conjuntos.

A Tabela 4.2 apresenta o tamanho dos conjuntos de teste e a quantidade de imagens descartadas utilizando a ferramenta *Discover* para medir a cobertura no caso dos conjuntos T_1,..,T_{30}.

Foi selecionado o conjunto de teste T_{30} como exemplo para demonstrar o procedimento realizado para cada conjunto de teste. A Figura 4.6 mostra as imagens que constituem o conjunto de teste T_{30}, a situação original das imagens e para cada ponto de corte, qual foi a resposta do CAD Nunes (2001). Em seguida avaliou-se para cada ponto de corte, a quantidade de casos verdadeiros-positivos (VP) e falsos-negativos (FN) para as imagens positivas e a quantidade de casos verdadeiros-negativos (VN) e falsos-positivos (FP) para as imagens negativas. Com base nesses dados, geraram-se as informações para chegar na curva ROC do conjunto de teste T_{30}. Como se pode observar na Figura 4.6, foram selecionadas 8 imagens, 5 positivas e 3 negativas, a tabela 1 da Figura 4.6 mostra o resultado da execução do CAD, com base nos resultados da tabela de execução do CAD é gerado a tabela 2 da Figura 4.6 para obter a curva ROC e em seguida é realizado o cálculo da área sob a curva.

Esse procedimento que foi mostrado com o conjunto de teste T_{30} foi realizado para todos os conjuntos de teste T_1,..,T_{30} e T'_1,..,T'_{30}.

A Tabela 4.3 mostra, na primeira coluna, o nome dos conjuntos de teste; na segunda coluna, a área obtida para cada conjunto de teste utilizando a técnica de análise de cobertura; na terceira coluna, o nome dos conjuntos de teste, e na quarta coluna, a área obtida para cada conjunto de teste, sem utilizar a técnica de análise de cobertura.

Foi feito o experimento conforme demonstra os resultados na Tabela 4.3 e foram calculados a média e o desvio padrão das áreas obtidas em ambos os casos: com ou sem o uso de cobertura. O objetivo é avaliar a variabilidade dos valores obtidos. Esperava-se que uma técnica mais precisa para seleção dos dados para gerar a curva ROC apresentase resultados sempre parecidos, enquanto que a simples seleção aleatória deveria criar

Tabela 4.2: Tamanho dos Conjuntos de Teste

Conjuntos)	Imagens	Descartadas
T_1	6	9
T_2	7	20
T_3	4	3
T_4	5	9
T_5	8	18
T_6	7	24
T_7	6	32
T_8	7	2
T_9	6	9
T_{10}	9	20
T_{11}	8	4
T_{12}	5	17
T_{13}	6	14
T_{14}	4	1
T_{15}	8	31
T_{16}	8	14
T_{17}	8	9
T_{18}	9	31
T_{19}	5	1
T_{20}	5	24
T_{21}	7	59
T_{22}	8	14
T_{23}	7	20
T_{24}	5	9
T_{25}	6	7
T_{26}	5	6
T_{27}	8	65
T_{28}	6	11
T_{29}	6	78
T_{30}	8	52
Média	6,6	20,43

conjuntos cujos resultados diferissem significativamente entre si.

Observando-se, porém, o desvio padrão das áreas obtidas dos conjuntos utilizando-se o método proposto dp(T_1,..,T_{30}) e o desvio padrão dos conjuntos selecionados aleatoriamente, dp(T'_1,..,T'_{30}) notou-se que esses valores diferem significativamente. Na verdade, dp(T_1,..,T_{30}) é superior a (dp(T'_1,..,T'_{30})).

Para explicar esse fato, observando os resultados, verificou-se que os conjuntos utilizados T_1,..,T_{30} e T'_1,..,T'_{30} possuem uma cardinalidade pequena. Como a curva

Conjunto T30 → 8 imagens

original = 5 imagens positivas e 3 imagens negativas

imagens descartadas 52

Conjunto T30

Imagem	Original	2.5	4.5	6.5	8.0
0937280507991ECC2	Negativa	Positiva	Positiva	Negativa	Negativa
3272712906931DCC3	Negativa	Negativa	Negativa	Negativa	Negativa
3272712906931DML4	Negativa	Negativa	Negativa	Negativa	Negativa
0447570207911DCC1	Positiva	Positiva	Positiva	Positiva	Negativa
0592700112991EML1	Positiva	Negativa	Negativa	Negativa	Negativa
1692022709991DCC1	Positiva	Positiva	Positiva	Positiva	Negativa
1692022709991EML1	Positiva	Positiva	Positiva	Positiva	Negativa
3060171301941EML1	Positiva	Positiva	Positiva	Positiva	Positiva

Tabela (1)

	Positiva		Negativa	
Desvio	VP	FN	VN	FP
2.5	4	1	2	1
4.5	4	1	2	1
6.5	4	1	3	0
8.0	1	4	3	0

Tabela (2)

	Positiva		Negativa	
Desvio	VP	FN	VN	FP
2.5	0,80	0,20	0,67	0,33
4.5	0,80	0,20	0,67	0,33
6.5	0,80	0,20	1	0
8.0	0,20	0,80	1	0

Tabela (1) Resultados obtidos após a execução do CAD para o conjunto T30.

Tabela (2) Resultados obtidos com base na tabela 1 para gerar a curva ROC do conjunto T30 utilizando a ferramenta Origin.

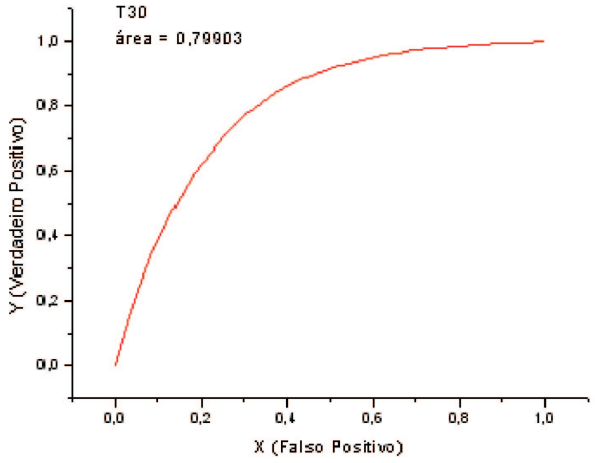


Figura 4.6: Dados para gerar a Curva ROC e Curva ROC do conjunto T_{30}

ROC leva em consideração o número de acertos versus o tamanho total do conjunto, uma pequena variação no número de acertos (uma unidade, por exemplo) causa uma variação muito grande no comportamento da curva. Como conseqüência, a variação das áreas, expressa pelo desvio padrão, tende a ser alta. Para tentar eliminar os desvios causados pela cardinalidade reduzida, foram avaliadas duas possíveis estratégias, ambas com o intuito de criarem-se conjuntos maiores.

A primeira proposta consiste em utilizar um critério de cobertura mais exigente,

Tabela 4.3: Resultados Obtidos

Conjunto	Área(com cobertura)	Conjunto	Área(sem cobertura)
T_1	0,97189	T'_1	0,99999
T_2	0,77054	T'_2	0,78484
T_3	0,46616	T'_3	0,66983
T_4	0,37500	T'_4	0,99999
T_5	0,79903	T'_5	0,82648
T_6	0,67678	T'_6	0,90000
T_7	0,63826	T'_7	0,99999
T_8	0,97839	T'_8	0,61884
T_9	0,99999	T'_9	0,99999
T_{10}	0,84412	T'_{10}	0,93546
T_{11}	0,73774	T'_{11}	0,99134
T_{12}	0,50266	T'_{12}	0,61195
T_{13}	0,90000	T'_{13}	0,68543
T_{14}	0,96088	T'_{14}	0,99999
T_{15}	0,82009	T'_{15}	0,90000
T_{16}	0,79190	T'_{16}	0,58306
T_{17}	0,87300	T'_{17}	0,75000
T_{18}	0,77054	T'_{18}	0,91500
T_{19}	0,99999	T'_{19}	0,96930
T_{20}	0,87398	T'_{20}	0,99999
T_{21}	0,47687	T'_{21}	0,99999
T_{22}	0,97492	T'_{22}	0,69200
T_{23}	0,97680	T'_{23}	0,62558
T_{24}	0,66500	T'_{24}	0,87500
T_{25}	0,98322	T'_{25}	0,70371
T_{26}	0,77661	T'_{26}	0,99999
T_{27}	0,74298	T'_{27}	0,83500
T_{28}	0,47542	T'_{28}	0,50063
T_{29}	0,95691	T'_{29}	0,83500
T_{30}	0,79903	T'_{30}	0,91500
Média	0,78929	Média	0,83745
Desvio	0,18487	Desvio	0,15600

como, por exemplo, critérios de fluxo de dados ou baseados em mutação. Com a utilização de tais critérios certamente seriam obtidos conjuntos maiores e também, mais significativos. Embora essa abordagem seja a mais apropriada, não foi possível utilizá-la pela falta de ferramenta que suportasse a aplicação de tais critérios para a linguagem Delphi, usada no CAD de estudo.

A segunda abordagem possível, e que foi utilizada no experimento, seria unir alguns dos conjuntos gerados, de forma a manter as suas características em termos de

cobertura e obter conjuntos mais numerosos. Por exemplo, unindo os conjuntos T_1 e T_2 obtém-se um conjunto com cardinalidade maior e que contém casos de teste que cobrem todos os requisitos de teste (pelo menos duas vezes).

Fez-se, então, a união de pares de conjuntos ($T_1 \cup T_2$=$T_{1,2}$, $T_3 \cup T_4$=$T_{3,4}$, etc), criando-se 15 novos conjuntos que utilizam a análise de cobertura e 15 que não utilizam, que seriam as uniões dos pares de conjuntos de T'_1,..,T'_{30}.

A Tabela 4.4 mostra o tamanho que os conjuntos ficaram depois da união efetuada. Na primeira coluna é disposto, o nome dos conjuntos de teste e na segunda coluna, a quantidade de imagens de cada conjunto.

Tabela 4.4: Tamanho dos Conjuntos com a União de 2 Conjuntos

Conjuntos	Imagens
$T_{1,2}$	13
$T_{3,4}$	9
$T_{5,6}$	15
$T_{7,8}$	13
$T_{9,10}$	15
$T_{11,12}$	13
$T_{13,14}$	10
$T_{15,16}$	16
$T_{17,18}$	17
$T_{19,20}$	10
$T_{21,22}$	15
$T_{23,24}$	12
$T_{25,26}$	11
$T_{27,28}$	14
$T_{29,30}$	14
Média	13,13

Para obter a área sob a curva foi necessário gerar a curva ROC para cada conjunto de teste $T_{1,2}$,..,$T_{29,30}$ e $T'_{1,2}$,..,$T'_{29,30}$. Após o cálculo da área sob a curva, foram calculados a média e o desvio padrão dos novos conjuntos obtidos.

A Tabela 4.5 mostra os resultados obtidos após este procedimento. Na primeira coluna aparece, o nome dos conjuntos de teste; na segunda coluna, a área obtida para cada conjunto de teste utilizando a técnica de cobertura; na terceira coluna, o nome dos conjuntos de teste e, na quarta coluna, a área obtida para cada conjunto de teste, sem utilizar a técnica de cobertura.

Tabela 4.5: Resultados Obtidos - União de 2 Conjuntos

Conjunto	Área(com cobertura)	Conjunto	Área(sem cobertura)
$T_{1,2}$	0,83331	$T'_{1,2}$	0,92858
$T_{3,4}$	0,59700	$T'_{3,4}$	0,83000
$T_{5,6}$	0,70257	$T'_{5,6}$	0,78580
$T_{7,8}$	0,82629	$T'_{7,8}$	0,83631
$T_{9,10}$	0,83243	$T'_{9,10}$	0,98670
$T_{11,12}$	0,75078	$T'_{11,12}$	0,83272
$T_{13,14}$	0,83331	$T'_{13,14}$	0,83002
$T_{15,16}$	0,83022	$T'_{15,16}$	0,64596
$T_{17,18}$	0,80501	$T'_{17,18}$	0,85000
$T_{19,20}$	0,98192	$T'_{19,20}$	0,97683
$T_{21,22}$	0,68153	$T'_{21,22}$	0,89026
$T_{23,24}$	0,83972	$T'_{23,24}$	0,60705
$T_{25,26}$	0,92795	$T'_{25,26}$	0,84646
$T_{27,28}$	0,66175	$T'_{27,28}$	0,65311
$T_{29,30}$	0,81267	$T'_{29,30}$	0,89000
Média	0,79443	Média	0,82599
Desvio	0,10068	Desvio	0,11350

Pode-se observar que, com esse novo conjunto de teste houve um aumento na cardinalidade dos conjuntos e o desvio padrão é menor quando utiliza a técnica de análise de cobertura.

Para complementar o estudo, fez-se, necessário a união de trios de conjuntos ($T_{1,2,3}$,..,$T_{28,29,30}$), criando-se mais 10 conjuntos de teste e fez-se também a união de trios de conjuntos ($T'_{1,2,3}$,..,$T'_{28,29,30}$) criando-se outros 10 conjuntos de teste. Os conjuntos continuam com as mesmas características, somente com a cardinalidade ainda mais elevada.

A Tabela 4.6 mostra o tamanho dos conjuntos depois da união de 3 conjuntos. Na primeira coluna, o nome dos conjuntos de teste; na segunda coluna, a quantidade de imagens de cada conjunto.

Da mesma forma, descrita anteriormente, para obter a área sob a curva foi necessário gerar a curva ROC de cada conjunto de teste $T_{1,2,3}$,..,$T_{28,29,30}$ e $T'_{1,2,3}$,..,$T'_{28,29,30}$. Após o cálculo da área sob a curva, foram calculados a média e o desvio padrão da união de 3 conjuntos utilizando a técnica de análise de cobertura e foi calculado a média e o desvio padrão da união de 3 conjuntos sem utilizar a técnica de análise de cobertura.

Tabela 4.6: Tamanho dos Conjuntos com a União de 3 Conjuntos

Conjuntos	Imagens
$T_{1,2,3}$	17
$T_{4,5,6}$	20
$T_{7,8,9}$	19
$T_{10,11,12}$	22
$T_{13,14,15}$	18
$T_{16,17,18}$	25
$T_{19,20,21}$	17
$T_{22,23,24}$	20
$T_{25,26,27}$	19
$T_{28,29,30}$	20
Média	19,70

A Tabela 4.7 mostra os resultados obtidos com os novos conjuntos. Na primeira coluna, tem-se o nome dos conjuntos de teste; na segunda coluna, a área obtida para cada conjunto de teste utilizando a técnica de análise de cobertura; na terceira coluna, o nome dos conjuntos de teste e, na quarta coluna, a área obtida para cada conjunto de teste, sem utilizar a técnica de cobertura.

Tabela 4.7: Resultados Obtidos - União de 3 Conjuntos

Conjunto	Área(com cobertura)	Conjunto	Área(sem cobertura)
$T_{1,2,3}$	0,74385	$T'_{1,2,3}$	0,84117
$T_{4,5,6}$	0,67442	$T'_{4,5,6}$	0,86299
$T_{7,8,9}$	0,85694	$T'_{7,8,9}$	0,88171
$T_{10,11,12}$	0,76791	$T'_{10,11,12}$	0,87759
$T_{13,14,15}$	0,84556	$T'_{13,14,15}$	0,87255
$T_{16,17,18}$	0,80186	$T'_{16,17,18}$	0,70264
$T_{19,20,21}$	0,76254	$T'_{19,20,21}$	0,98947
$T_{22,23,24}$	0,87582	$T'_{22,23,24}$	0,68139
$T_{25,26,27}$	0,82843	$T'_{25,26,27}$	0,82368
$T_{28,29,30}$	0,73745	$T'_{28,29,30}$	0,82137
Média	0,78948	Média	0,83545
Desvio	0,06328	Desvio	0,08914

Foi calculada a média e medido o desvio padrão para esses novos conjuntos, tendo sido observado que o desvio padrão é maior quando se utiliza a técnica de análise de cobertura.

4.3 Resultados Obtidos

As Figuras 4.7 e 4.8 mostram os resultados da Tabela 4.3 na forma gráfica. O cálculo da média da área das curvas ROC dos conjuntos T_1,..,T_{30} é 0,78929, para um desvio padrão de 0,18487 e a média da área das curvas ROC dos conjuntos T'_1,..,T'_{30}) é 0,83745 para um desvio padrão de 0,15600.

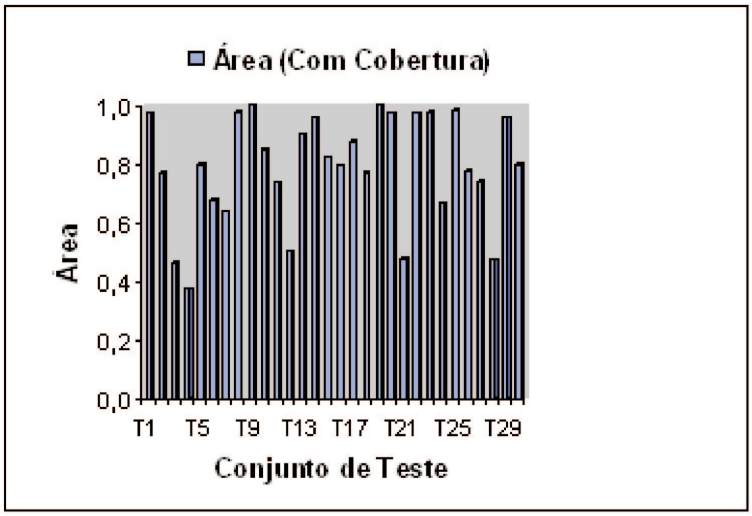


Figura 4.7: Gráfico área com cobertura

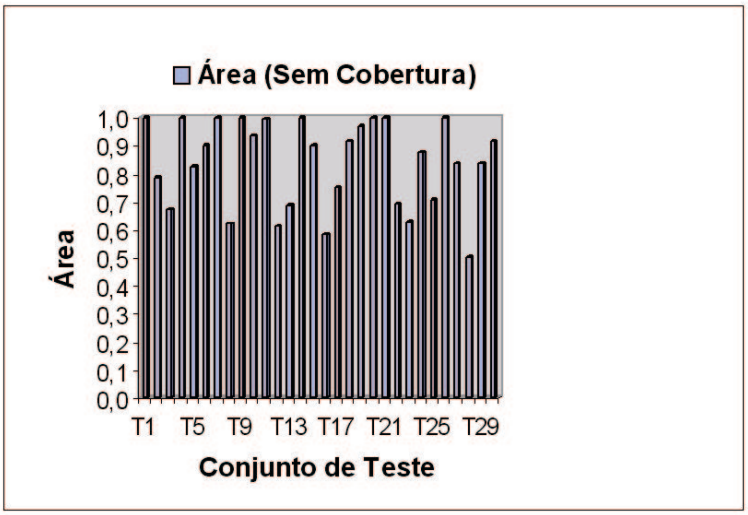


Figura 4.8: Gráfico área sem cobertura

Isto demonstra que o desvio padrão dos conjuntos T_1,..,T_{30} é maior que o desvio padrão dos conjuntos T'_1,..,T'_{30}, o que contraria a expectativa inicial: que os conjuntos de teste que utilizavam a técnica de análise de cobertura possuíssem uma área mais próxima da média e que o desvio padrão fosse menor.

Repetiu-se, então o experimento com a união de 2 conjuntos como explicado na seção anterior. Os conjuntos obtidos continuam com as mesmas características, mas com uma cardinalidade maior, conforme mostra a Tabela 4.4.

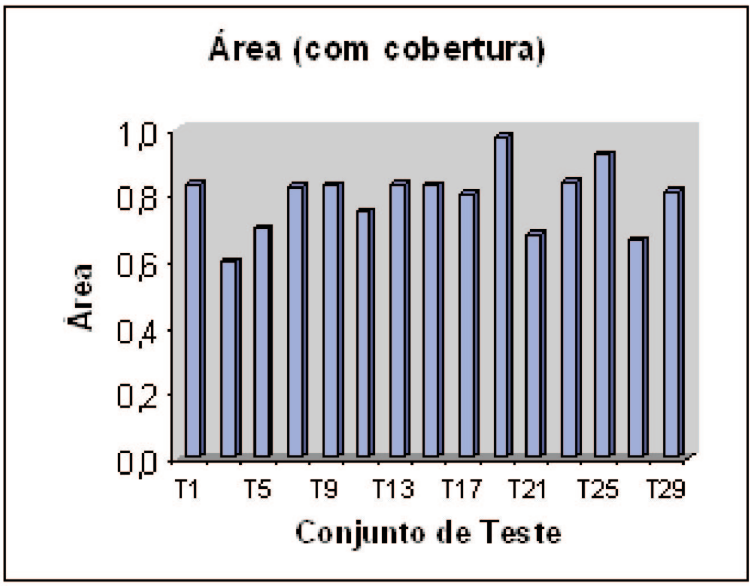


Figura 4.9: União de 2 conjuntos com cobertura

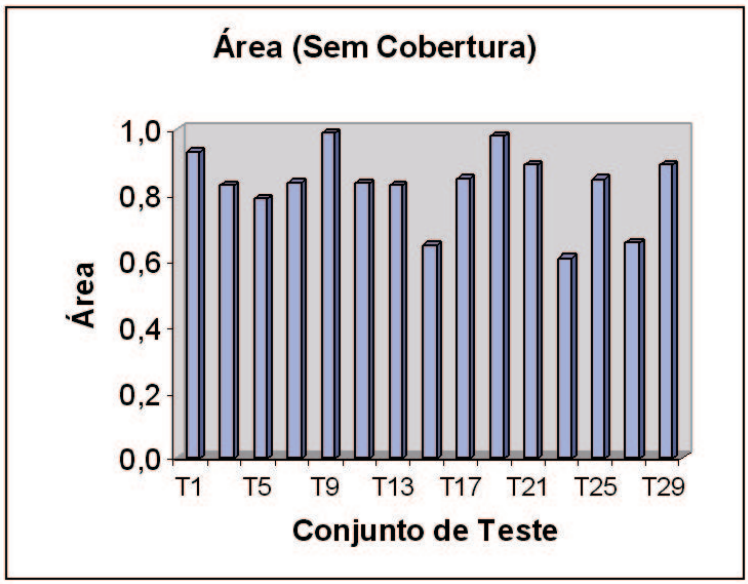


Figura 4.10: União de 2 conjuntos sem cobertura

As Figuras 4.9 e 4.10 mostram os resultados da Tabela 4.5 na forma gráfica. O cálculo da média da área das curvas ROC dos conjuntos $T_{1,2},..,T_{29,30}$ é 0,79443, para um desvio padrão de 0,10068 e a média da área das curvas ROC dos conjuntos $(T'_{1,2},..,T'_{29,30})$ é 0,82599, para um desvio padrão de 0,11350. Isto desmonstra que o desvio padrão dos conjuntos $T_{1,2},..,T_{29,30})$ é menor que o desvio padrão dos conjuntos $(T'_{1,2},..,T'_{29,30})$, que

era o resultado esperado. Então o aumento da cardinalidade foi positivo em relação à hípotese inicial.

Repetiu-se o experimento com a união de 3 conjuntos. Os conjuntos continuam com as mesmas características, mas com uma cardinalidade maior, conforme mostra a Tabela 4.6.

As Figuras 4.11 e 4.12 mostram os resultados da Tabela 4.7 na forma gráfica. A média da área das curvas ROC dos conjuntos $T_{1,2,3},..,T_{28,29,30}$ é 0,78948, para um desvio padrão de 0,06328 e a média da área das curvas ROC dos conjuntos $T'_{1,2,3},..,T'_{28,29,30}$ é 0,83546, para um desvio padrão de 0,08914. Isto desmonstra que o desvio padrão dos conjuntos $T_{1,2,3},..,T_{28,29,30}$ é maior que o desvio padrão dos conjuntos $T'_{1,2,3},..,T'_{28,29,30}$.

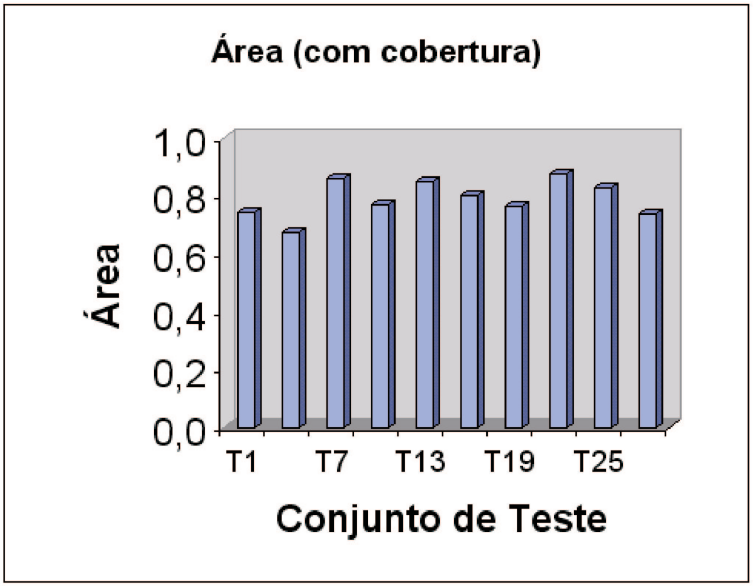


Figura 4.11: União de 3 conjuntos com cobertura

Pode-se observar que os conjuntos $T_1,..,T_{30}$ utilizaram a técnica de análise de cobertura para seleção dos conjuntos de teste, tendo uma média de área menor que os conjuntos $T'_1,..,T'_{30}$ que não utilizaram a técnica de cobertura para seleção dos conjuntos, demonstrando, assim, que os conjuntos $T_1,..,T_{30}$ geram curvas ROC com áreas menores em comparação aos conjuntos $T'_1,..,T'_{30}$. Os conjuntos de teste $T_1,..,T_{30}$ passam por caminhos que conseguem fazer com que o sistema em questão apresente resultados mais reais.

A Tabela 4.8 resume os dados obtidos nesse experimento. Comparando os valores da média e do desvio padrão obtidos através dos conjuntos de teste utilizando a técnica de cobertura e sem utilizar a técnica de cobertura (utilizando conjuntos selecionados aleatori-

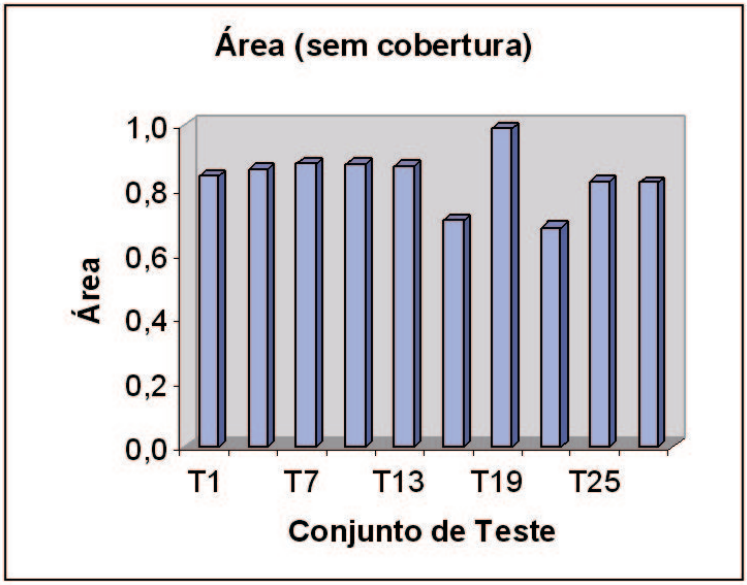


Figura 4.12: União de 3 conjuntos sem cobertura

Tabela 4.8: Resultados do calculo da média e desvio padrão

Com Cobertura	Média	Desvio Padrão
T1 a T30	0,78929	0,18487
União 2 Conjuntos	0,79443	0,10068
União 3 Conjuntos	0,78948	0,06328
Sem Cobertura	Média	Desvio Padrão
T1 a T30	0,83745	0,15600
União 2 Conjuntos	0,82599	0,11350
União 3 Conjuntos	0,83546	0,08914

amente), verificou-se que o desvio padrão, utilizando a técnica de cobertura, é ligeiramente menor. Portanto é um grande passo para obter-se uma curva mais padronizada, evitando, assim, possíveis distorções no formato da curva que possam advir da seleção inadequada de casos de teste.

Por outro lado, a pequena diferença nos valores de desvio padrão obtidos, juntamente com o fato de que no primeiro conjunto de dados (1ª e 4ª linhas da tabela 4.8), o desvio padrão foi inferior para a seleção aleatória, não permitem uma conclusão mais precisa sobre a eficácia da técnica proposta. O motivo principal para isso é, certamente, a fragilidade do critério de cobertura utilizado.

A Tabela 4.9 mostra a cobertura obtida para cada um dos conjuntos T'_1,..,T'_{30}, em relação ao critério todos-os-nós. Nota-se que as medidas de cobertura se aproximam bastante dos valores obtidos pelos conjuntos gerados pela técnica proposta. Isso, de

Tabela 4.9: Resultados obtidos de cobertura para conjuntos aleatórios

Conjunto	Desvio 2.5	Desvio 4.5	Desvio 6.5	Desvio 8.0
T'_1	79%	80%	79%	70%
T'_2	79%	79%	80%	80%
T'_3	79%	79%	79%	80%
T'_4	79%	79%	79%	70%
T'_5	79%	79%	79%	70%
T'_6	79%	79%	79%	70%
T'_7	79%	79%	79%	79%
T'_8	79%	79%	79%	79%
T'_9	79%	79%	79%	70%
T'_{10}	79%	79%	80%	70%
T'_{11}	79%	79%	80%	70%
T'_{12}	79%	79%	79%	70%
T'_{13}	79%	80%	80%	69%
T'_{14}	79%	80%	70%	69%
T'_{15}	80%	80%	71%	70%
T'_{16}	79%	79%	79%	79%
T'_{17}	79%	80%	70%	69%
T'_{18}	79%	80%	80%	70%
T'_{19}	79%	79%	79%	79%
T'_{20}	79%	79%	79%	69%
T'_{21}	79%	79%	79%	70%
T'_{22}	79%	79%	80%	70%
T'_{23}	80%	80%	80%	80%
T'_{24}	78%	79%	79%	70%
T'_{25}	79%	79%	79%	69%
T'_{26}	79%	79%	79%	79%
T'_{27}	80%	80%	80%	71%
T'_{28}	79%	80%	80%	79%
T'_{29}	79%	79%	79%	79%
T'_{30}	79%	80%	80%	70%

certa forma, justifica os resultados do desvio padrão semelhantes, afinal, estão sendo comparados conjuntos de coberturas bastante semelhantes.

Outro problema enfrentado pelo experimento que está relacionado com o critério é a cardinalidade dos conjuntos de teste que é muito pequena. Como a curva ROC leva em consideração o número de acertos em relação ao tamanho total do conjunto, uma pequena variação no número de acertos (uma unidade por exemplo) causa uma variação muito grande no comportamento da curva.

Tabela 4.10: Resultados obtidos de cobertura para conjuntos aleatórios - união de 2 conjuntos

Conjunto	Desvio 2.5	Desvio 4.5	Desvio 6.5	Desvio 8.0
$T'_{1,2}$	80%	80%	80%	80%
$T'_{3,4}$	79%	79%	80%	80%
$T'_{5,6}$	79%	79%	79%	70%
$T'_{7,8}$	79%	79%	79%	79%
$T'_{9,10}$	79%	79%	80%	70%
$T'_{11,12}$	79%	80%	80%	70%
$T'_{13,14}$	79%	80%	80%	70%
$T'_{15,16}$	80%	80%	80%	80%
$T'_{17,18}$	79%	80%	80%	70%
$T'_{19,20}$	79%	79%	79%	79%
$T'_{21,22}$	79%	79%	80%	71%
$T'_{23,24}$	80%	80%	80%	80%
$T'_{25,26}$	79%	79%	79%	79%
$T'_{27,28}$	80%	80%	80%	80%
$T'_{29,30}$	79%	80%	80%	80%

A Tabela 4.10 mostra a cobertura obtida para cada um dos conjuntos de teste $T'_{1,2}$,..,$T'_{29,30}$, em relação ao critério todos-os-nós.

Tabela 4.11: Resultados obtidos de cobertura para conjuntos aleatórios - união de 3 conjuntos

Conjunto	Desvio 2.5	Desvio 4.5	Desvio 6.5	Desvio 8.0
$T'_{1,2,3}$	80%	80%	80%	80%
$T'_{4,5,6}$	79%	79%	79%	70%
$T'_{7,8,9}$	79%	79%	79%	79%
$T'_{10,11,12}$	79%	80%	80%	70%
$T'_{13,14,15}$	80%	80%	80%	70%
$T'_{16,17,18}$	80%	80%	80%	80%
$T'_{19,20,21}$	79%	79%	79%	79%
$T'_{22,23,24}$	80%	80%	80%	80%
$T'_{25,26,27}$	80%	80%	80%	80%
$T'_{28,29,30}$	79%	80%	80%	80%

A Tabela 4.11 mostra a cobertura obtida para cada um dos conjuntos de teste $T'_{1,2,3}$,..,$T'_{28,29,30}$, em relação ao critério todos nós.

Uma pergunta fundamental a respeito dos resultados de um experimento é até quando são válidos os resultados. Validade adequada refere-se àquela em que os resultados

devem ser válidos para a população de interesse. Antes de tudo, os resultados devem ser válidos para a população à qual o exemplo foi desenvolvido.

Há diferentes planos de classificação para diferentes tipos de ameaças à validade de um experimento. Campbell e Stanley (2002) definem dois tipos: ameaças para a validade interna e externa. Traçaram uma lista de quatro tipos de ameaças à validade de resultados experimentais. Os quatro tipos de ameaças são: validade do resultado, validade interna, validade construtiva, validade externa (WOHLIN et al., 2002).

A validade do resultado é algumas vezes citada como validade do resultado estatístico. As ameaças à validade do resultado são relacionadas a problemas que afetam a capacidade de desenvolver o resultado correto sobre as relações entre o tratamento e a conseqüência de um experimento. Esses problemas incluem, por exemplo, escolha de testes estatíticos, escolha dos tamanhos da amostra, cuidados tomados ao implementar e medir um experimento (WOHLIN et al., 2002).

Ameaças à validade interna dizem respeito a problemas que podem indicar uma relação causal (WOHLIN et al., 2002).

Ameaças à validade construtiva referem-se à extensão que o ambiente do experimento verdadeiramente reflete a construção em estudo (WOHLIN et al., 2002).

Ameaças à validade externa dizem respeito à capacidade de generalizar os resultados do experimento fora do ambiente do experimento (WOHLIN et al., 2002).

No experimento realizado temos várias ameaças; por exemplo, utilizamos um conjunto de imagens e não conseguimos chegar a 100% de cobertura dos requisitos executáveis. Esta é uma ameaça que demonstra que o método pode ser influenciado pela qualidade geral de T, e pode ser classificada como validade construtiva.

Outro tipo de ameaça é que o experimento realizado utilizou o CAD de Nunes (2001), não podendo ser utilizados os resultados desse experimento para outro CAD. Esta é uma ameaça que pode ser classificada como validade externa.

Outra ameaça para o experimento é que o critério de teste usado (todos-os-nós) é muito fraco porque satisfaz facilmente todos os requisitos e pode conter defeitos que não conseguem revelar. Esta é uma ameaça que pode ser classificada como validade do resultado.

4.4 Considerações Finais

Através de estudos empíricos realizados chegou-se a algumas conclusões sobre os resultados obtidos no experimento.

Conforme esperado, através de experimentos, que dependendo do conjunto de imagens que é utilizado para construir a curva ROC, pode-se gerar uma curva boa ou uma curva ruim.

Foi desmonstrado através de números e gráficos que utilizando a técnica de cobertura para um conjunto de teste que possui uma cardinalidade maior, o desvio padrão entre as curvas é menor.

Também analisando os gráficos observa-se, que com o aumento da cardinalidade dos conjuntos e a utilização da técnica de análise de cobertura, geraram curvas com áreas menores do que as geradas sem a utilização da técnica de análise de cobertura. A técnica de seleção de casos de teste, utilizando o teste estrutural, obriga o sistema a passar por caminhos que conseguem fazer com que o sistema em questão apresente resultados mais reais. Salientando que aleatoriamente também se consegue obter um conjunto de teste que demonstre resultados mais reais; só que a chance de acontecer é a mesma de não acontecer.

Analisando os resultados obtidos verificou-se que o critério de teste usado (todos-os-nós) é muito fraco porque os requisitos de teste gerados pelo critério de teste são satisfeitos facilmente e pode conter defeitos não revelados. A solução correta seria utilizar algum critério mais forte como todos-usos, que faria que os conjuntos fossem maiores. Infelizmente, não havia uma ferramenta para isso.

Utilizando a técnica de cobertura no sistema em questão foi percebido que o sistema não executa alguns caminhos. Isto acontece, porque o conjunto de imagens analisado não passa por determinados caminhos realmente e, segundo, existem caminhos que, para nenhum conjunto de imagens, será coberto (caminhos executados quando há intervenção do usuário por exemplo).

O objetivo do método proposto é evitar possíveis distorções na construção da curva que possam advir da seleção inadequada de casos de teste.

Analisando os resultados obtidos através de experimentos chegou-se à conclusão

que o método proposto atingiu o objetivo que era obter curvas que possuem uma forma semelhante, curvas mais padronizadas. Para se obter uma conclusão definitiva sobre o método o experimento tem que ser refeito utilizando um critério de teste mais forte.

5 Conclusão

Tradicionalmente, utiliza-se a curva ROC na avaliação de esquemas de diagnóstico auxiliado por computador (CAD). Este método permite que se avalie o esquema através da contraposição do número de resultados falso-positivos contra o número de resultados verdadeiro-positivos, para diferentes configurações do sistema. A escolha dos dados utilizados na geração da curva ROC pode distorcer os resultados obtidos.

A curva ROC pode induzir a avaliações enganosas porque não leva em consideração a seleção dos casos de teste.

Não existe um método padrão para se selecionar os dados de entrada para gerar a curva e isso pode gerar problemas. Dependendo do conjunto de dados de entrada pode-se gerar uma curva ROC muito boa e uma curva ROC muito ruim.

Buscando melhorar a avaliação dos sistemas CADs utilizando-se a curva ROC, foi proposto um método que utiliza a técnica de cobertura estrutural na seleção dos dados de entrada para a geração da curva.

Definiu-se, então, um método que procura normalizar a escolha dos casos de teste através da análise de cobertura. A idéia básica do método é que devem ser selecionados casos de teste que tenham características distintas entre si. O que se faz é selecionar apenas casos de teste que progressivamente contribuam para a cobertura do critério de teste utilizado. Se um caso de teste não contribui para a aumento da adequação ao critério, ele é descartado por considerar-se que no conjunto total, já existem casos de teste com as suas características. Com isso, pretende-se eliminar redundâncias e obter-se uma curva ROC mais padronizada.

O objetivo do método proposto é evitar possíveis distorções no formato da curva que possam advir da seleção inadequada de casos de teste.

Foram realizados estudos empíricos que visam demonstrar a validade e utilidade

do método proposto. Utilizando-se o sistema CAD desenvolvido por Nunes (2001) e o conjunto de 200 imagens disponíveis, aplicou-se o método descrito no capítulo 4 na seção 4.1 que utiliza a técnica de análise de cobertura utilizando-se diversas seqüências aleatórias de casos de teste.

Comparando os resultados obtidos do método proposto com um método que não utiliza a técnica de análise de cobertura (utilizando conjuntos aleatoriamente) observou-se que o desvio padrão obtido com o método proposto é ligeiramente menor.

Utilizando o método proposto na escolha dos casos de teste através da análise de cobertura foram obtidas curvas que possuem uma forma semelhante que representaria a forma canônica da curva para o sistema em estudo. Porém tem que ser refeito o experimento utilizando um critério de teste mais forte, para se chegar a uma conclusão definitiva sobre o método.

5.1 Contribuições e Trabalhos Futuros

Este trabalho contribuiu para demonstrar a importância de ter um método de entrada de dados para evitar possíveis distorções nos resultados e a idéia de selecionar casos de teste com características distintas entre si são válidas, o critério de teste todos-os-nós apresentou-se frágil para este estudo, para obter-se um resultado mais expressivo é necessário utilizar um critério de teste mais exigente.

Como continuidade deste trabalho e com o objetivo de obter um maior grau de precisão dos experimentos realizados, novos experimentos podem ser realizados, como:

- Refazer o experimento utilizando critérios de teste mais fortes (mais exigentes).
- Fazer novos experimentos utilizando outros CAD para medir o comportamento, mas utilizando o mesmo contexto (mesmo conjunto de imagens).
- Refazer o experimento utilizando o CAD Nunes (2001), com outro conjunto de imagens para medir o comportamento.
- Desenvolvimento de uma ferramenta que utilize a técnica de análise de cobertura com um critério mais apropriado.

Referências Bibliográficas

BRAGA, A. C. S. *Curvas ROC: Aspectos Funcionais e Aplicações*. Tese (Doutorado em Engenharia de Produção e Sistemas) — Universidade do Minho, Braga, Dez 2000.

CRESPO, A. *Modelo de Confiabilidade de Software Baseados em Cobertura de Critérios Estruturais de Teste*. Tese (Doutorado em Engenharia Elétrica) — Universidade Estadual de Campinas, Campinas, 1997.

CYAMON, S. *Discover para Delphi*. 2005. Disponível em: http://www.cyamon.com.discover1.html [Acesso em 12 fevereiro de 2005].

DOI, K.; GIGER, M.; HOFFMANN, K. Computer-aided diagnosis in medical imaging. In: COMPUTER-AIDED., I. W. on (Ed.). Chicago: Elsevier, 1999.

EVANS, A. L. The evolution of medical images. *Adam Hilger Ltda and Bristol and Great Britain*, Adam Hilger Ltda and Bristol and Great Britain, 1981.

FACON, J. Processamento e análise de imagens. Curitiba: PUC, Feb 2002.

GALPAROZO, L. U.; FERNANDEZ, P. Unidad de epidemiologia clínica y bioestadística. *Complexo Hospitalario Juan Canalejo, CAD ATEN PRIMARIA*, A Coruña (España), v. 5, n. 4, p. 229–235, 1998.

GIGER, M. L. Computer-aided diagnosis of breast lesions in medical images. *Computing in Science Engineering*, v. 2, p. 39–45, 2000.

GIGER, M. L.; MACMAHON, H. Image processing and computer-aided diagnosis. In: GREENES, R.; BAUMAN, R. (Ed.). *Radiologic Clinics of North America*. Philadelphia: Saunders Publishing Co, 1996. p. 565–596.

GONZALEZ, R. C. *Processamento de Imagens Digitais*. São Paulo: Edgard Blucher, 2000.

INCA, I. N. C. *Instituto Nacional de Câncer,Câncer de Mama*. 2005. Disponível em: http://www.inca.gov.br/conteudoview.asp?id=336 [Acesso em 22 setembro de 2005].

MALDONADO, J. C. *Critérios Potenciais-usos: Uma Contribuição ao Teste Estrutural de Software*. Dissertação (Engenharia Elétrica) — Faculdade de Engenharia Elétrica, Campinas, 1991.

MALDONADO, J. C. et al. Introdução ao teste de software. *São Carlos:ICMC-USP*, (Mini Curso), 2000.

MALDONADO, J. C. et al. Aspectos teóricos e empíricos de teste de cobertura de software. *São Carlos: ICMC/USP*, (Relatório Técnico,31), 1998.

MARQUES, P. M. de A. Diagnóstico auxiliado por computador na radiologia. Radiologia Brasileira, v. 34, n. 5, p. 285–293, 2001.

MICROCAL, S. *Data Analysis and Technical Graphics*. 2005. Disponível em: http://www.originlab.com [Acesso em 15 fevereiro de 2005].

NISHIKAWA, R. M. et al. Performance of automated cad schemes for detection and classification of clustered microcalcifications. In: GALE, A. G. e. a. (Ed.). *Digital Mammography*. Amsterdam: Elsevier, 1994. p. 13–20.

NUNES, F. L. S. *Investigações em Processamento de Imagens Mamográficas para Auxílio ao Diagnóstico de Mamas Densas*. Tese (Doutorado na área de Física Aplicada - Instituto de Física de São Carlos) — Universidade de São Paulo, São Carlos, 2001.

PRESSMAN, R. S. *Engenharia de Software*. 3ª. ed. São Paulo: Makron Books, 1995.

RAPPS, S.; WEYUKER, E. Selecting software test data using data flow information. *IEEE Transactions on Software Engineering*, v. 11, n. 4, p. 367–375, Apr 1985.

RODRIGUES, S. C. M.; FRERE, A. F. Método abrangente para avaliação dos algoritmos de processamento de imagens médicas. In: *Congresso Brasileiro de Engenharia Biomédica*. Florianópolis: [s.n.], 2000. p. 1227–1232.

SPOTO, E. S.; PERES, L. M.; BUENO, P. M. S. Um estudo de critérios de teste de software baseados em fluxo de dados. *Campinas SP:Unicamp*, 1995.

WEYUKER, J. E. Theories of program testing and the application of revealing subdomains. *IEEE Transactions on Software Engineering*, v. 6, n. 3, p. 236–246, May 1980.

WOHLIN, C. et al. *Experimentation in Software Engineering: An Introduction*. [S.l.]: Kluwer Academic, 2002.

YARUSSO, L. M. et al. Application of computer-aided diagnosis to full-field digital mammography. *5 th International Warkshop on Digital Mammography*, p. 421–426, 2000.

YU, S.; GUAN, L. A cad system for the automatic detection of clustered microcalcifications in digitized mammogram films. *IEEE Transactions on Medical Imaging*, v. 19, n. 2, p. 115–126, Feb 2000.