

**UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO**

GABRIELA FELIX PERSINOTI

**Auxílio ao Diagnóstico de Artrite Reumatóide
Através de Técnicas de Inteligência Artificial**

**Ribeirão Preto – SP
2009**

Livros Grátis

<http://www.livrosgratis.com.br>

Milhares de livros grátis para download.

GABRIELA FELIX PERSINOTI

**Auxílio ao Diagnóstico de Artrite Reumatóide
Através de Técnicas de Inteligência Artificial**

Dissertação apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para a obtenção do título Mestre em Ciências.

Área de concentração: Genética

Orientadora: Prof^a. Dr^a. Silvana Giuliatti

**Ribeirão Preto – SP
2009**

FICHA CATALOGRÁFICA

Persinoti, Gabriela Felix

Auxílio ao Diagnóstico de Artrite Reumatóide Através de Técnicas de Inteligência Artificial, São Paulo. Ribeirão Preto, 2009.

p. 100 : il. ; 30cm

Dissertação de mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP – Área de concentração: Genética.

Orientadora: Giuliatti, Silvana.

1. Artrite Reumatóide. 2. Inteligência Artificial. 3. Sistema *Web*.

FOLHA DE APROVAÇÃO

Gabriela Felix Persinoti

Auxílio ao Diagnóstico de Artrite Reumatóide Através de Técnicas de Inteligência Artificial

Dissertação apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Mestre em Ciências.

Área de concentração: Genética

Aprovada em:

Banca examinadora

Prof. Dr.

Instituição: Assinatura: _____

Prof. Dr.

Instituição: Assinatura: _____

Prof. Dr.

Instituição: Assinatura: _____

Ao meu filho, Heitor

Ao meu esposo, Leonardo e

Aos meus pais, Carmen e Antônio

Amo vocês.

Agradecimentos

Agradeço, especialmente, à Prof^a. Dr^a. Silvana Giuliatti pela oportunidade de integração ao Grupo de Bioinformática, por todo o apoio, respeito e incentivo durante a realização deste projeto. Obrigada pela orientação e amizade.

Agradeço a disposição dos membros da banca examinadora para avaliar e discutir o presente projeto.

Agradeço ao Prof. Dr. Geraldo Aleixo Passos e a Dra. Cristina Moraes Junta pela colaboração e apoio para a realização deste projeto.

Agradeço ao Prof. Dr. Paulo Louzada Jr e ao doutorando Renê de Oliveira pela colaboração prestada ao projeto, que foi de fundamental importância para sua conclusão.

Agradeço ao Departamento de Genética e à Faculdade de Medicina de Ribeirão Preto pela estrutura oferecida para minha formação na pós-graduação.

Agradeço à FAPESP e ao CNPq pela bolsa de Mestrado oferecida para a realização deste projeto e todo o suporte financeiro.

Agradeço à FAEPA pelo apoio financeiro dado ao projeto, principalmente para a participação em eventos.

Gostaria de agradecer aos meus grandes amigos do Grupo de Bioinformática, Saulo, André, Luciano, Daniel, Renato, Pablo, Vitor e Mariana pela amizade, pelo apoio e pelo convívio durante esse dois anos.

Agradeço ao meu esposo, Leonardo e meu filhinho Heitor, por todo o amor, carinho, apoio, compreensão e ajuda, em fim, por ser uma família incrível.

Agradeço aos meus pais, Carmen e Antônio pela educação, pelos ensinamentos, por todo o esforço que fizeram para eu estar aqui e por toda a ajuda que me deram e dão sempre.

Agradeço ao meu irmão Luiz Fernando, pela amizade, pelo apoio, por ser um ótimo irmão.

Agradeço aos meus avôs vó Lina e vô Álvaro por todo o apoio, incentivo, e por tudo que já fizeram por mim.

A toda a minha família que sempre apoiou minhas decisões e sempre se manteve ao meu lado em todos os momentos.

Agradeço aos meus amigos do Cofil, Rebeca, Mônica, Priscila, Bruno e Willian, grandes amigos que apesar da distância estão sempre presentes.

Aos colegas do laboratório da Profª Catarina e Profª Tiemi, Raquel, Ana Paula, Cristiano, Ana Cláudia, Paulo, Danilo (Chitão), Júnior, Aline e Danilo (Bolor).

Agradeço ao Prof. Dr. Renato Tinós e sua esposa Lúcia, por todo o apoio, incentivo e amizade.

Agradeço ao Prof. Dr. Ademilson E. Espencer Soares pela amizade e sua disposição, sempre brigando pelos alunos da pós-graduação.

Em fim, agradeço a todos que de alguma forma, direta ou indiretamente, colaboraram na realização deste projeto. Meu muito obrigado!

Apoio e Suporte Financeiro

Este projeto foi realizado com o apoio financeiro das seguintes entidades e instituições:

Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP
(Proc.: nº 06/59402-7).

Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq.

Fundação de Apoio ao Ensino, Pesquisa e Assistência – FAEPA.

Faculdade de Medicina de Ribeirão Preto – FMRP/USP.

Normalização Adotada

Este documento foi elaborado de acordo com:

Referências: Associação Brasileira de Normas Técnicas (ABNT) – NBR 6023

Universidade de São Paulo. Sistema Integrado de Bibliotecas. Diretrizes para apresentações de dissertações e teses da USP: documento eletrônico e impresso. São Paulo, 2004.

RESUMO

PERSINOTI, G. F. **Auxílio ao Diagnóstico de Artrite Reumatóide Através de Técnicas de Inteligência Artificial.** 2009. Dissertação (Mestrado). Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo. Ribeirão Preto, 2009.

A Artrite Reumatóide (AR) é uma doença auto-imune, crônica e inflamatória cujas manifestações são notadas, principalmente, nas articulações. Um ponto de extrema importância nesta doença é a necessidade de diagnóstico e início do tratamento dos pacientes o mais rápido possível, para que os danos às articulações e erosões ósseas sejam evitados, pois estes danos ocorrem no início da doença e são, muitas vezes, irreversíveis. O presente projeto teve como objetivo a construção de um sistema computacional via *Web* para o auxílio ao diagnóstico de pacientes portadores de artrite reumatóide, através da utilização de técnicas de inteligência artificial, como Redes Neurais Artificiais, Árvores de Decisão e Algoritmos Genéticos, e o sistema gerenciador de conteúdos Drupal. O objetivo do sistema é proporcionar um ambiente estruturado, no qual as informações clínicas e de expressão gênica de pacientes portadores de AR são cadastradas e categorizadas, para que, a partir destas informações, os algoritmos de aprendizado de máquinas possam ser treinados para selecionar características clínicas e genes relevantes para a predição da resposta dos pacientes ao tratamento com Drogas Anti-Reumáticas Modificadoras da Doença (DMARDs). Após o treinamento, os modelos gerados podem ser utilizados para a predição da resposta à DMARDs de novos pacientes. O sistema visa selecionar possíveis características que permitam identificar precocemente os pacientes que irão evoluir de forma mais agressiva e, assim, fornecer uma nova ferramenta para que os médicos possam indicar o melhor tratamento possível para cada paciente individualmente. O Sistema desenvolvido foi denominado ARIA e está disponível através do endereço <http://gbi.fmrp.usp.br/artrite>. Apenas médicos e colaboradores do projeto têm permissão para acessá-lo. Estão disponíveis os seguintes conteúdos: Cadastro de Pacientes, Ferramenta de Seleção de Atributos e Ferramenta de Predição da Resposta dos Pacientes à DMARDs. Além disso, o sistema disponibiliza diversas possibilidades de visualização e interação com tais conteúdos. Estão cadastrados, até o momento, 126 pacientes, sendo que 106 deles são utilizados para o treinamento da ferramenta de seleção de atributos e vinte deles para os testes da ferramenta de predição da resposta à DMARDs.

Palavras-chave: Artrite Reumatóide, Inteligência Artificial, Sistema *Web*.

ABSTRACT

PERSINOTI, G. F. **Diagnosis Aided to Rheumatoid Arthritis through Artificial Intelligence Techniques**. 2009. Dissertation (Master). Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo. Ribeirão Preto, 2009.

Rheumatoid Arthritis (RA) is a chronic, inflammatory, autoimmune disease, whose manifestations are observed, mainly in the joints. One point of extreme importance in this disease is the early diagnosis and the beginning of the treatment as quickly as possible. These procedures try to avoid joints damage and bone erosions, since these damages are often irreversible. This project aimed to develop a computational system to aid diagnosis of rheumatoid arthritis patients. Artificial intelligence techniques, such as Artificial Neural Networks, Decision Trees and Genetic Algorithms, and the content management system Drupal were employed to the system development. The system goal is to provide a structured environment in which clinical and gene expression information about RA patients are registered and categorized. From the information stored in the system database, the machine learning algorithms can be trained to select clinical information or genes which are relevant to predicting the patients' response to treatment with Disease Modifying Anti-Rheumatic Drugs (DMARDs). After the training, the generated models generated can be used for predicting the response to DMARDs of new RA patients. The system aims to select characteristics to early identify patients who might have more aggressively disease outcome and thus, it provides a new tool to physicians in the analysis of the best treatment for each patient individually. The system developed was called ARIA System and it is available at the web site <http://gbi.fmrp.usp.br/artrite>. Only doctors and collaborator researchers have permission to access it. The contents available in the system are the following: Patients Register, Feature Selection Tool and Prediction the Response to DMARDs Tool. Besides these contents, the system offers different kinds of displaying and interaction with the content. There are, until now, 126 patients registered in the ARIA System, 106 of them are used for training the features selection tool and twenty of them are used to test the tool for predicting patients' response to DMARDs.

Keywords: Rheumatoid Arthritis, Artificial Intelligence, Web System

Lista de Figuras

- Figura 1: Visão geral do funcionamento do Sistema ARIA. 1. Interação do usuário com o Sistema através da Internet. 2. Interação do CMS Drupal e a linguagem PHP com o banco de dados MySQL. 3. Linguagem Perl e o banco de dados MySQL para cadastro dos pacientes. 4. Integração do *software* Weka e a linguagem JAVA para a criação das ferramentas de seleção de atributos e de predição. 5. Sistema ARIA gerando resultados para os usuários. 19
- Figura 2: Exemplo de uma árvore de decisão indutiva. Os atributos utilizados para a construção da árvore foram: DAS 28, cujos valores são: alta, moderada, baixa e remissão clínica; Sexo, cujos valores são: masculino e feminino; E o atributo anti-CCP, cujos valores são positivo e negativo. O atributo classe é resposta ao tratamento com DMARDs cujos valores podem ser respondeu e não respondeu. 25
- Figura 3: Estrutura de uma Rede Neural Artificial. A camada de entrada representa os atributos do conjunto de treinamento. A camada intermediária realiza uma transformação não-linear dos dados de entrada e a camada de saída representa os valores do atributo classe, neste exemplo, Respondeu e Não Respondeu ao tratamento com DMARDs..... 27
- Figura 4: Pseudocódigo do algoritmo de treinamento *backpropagation*..... 28
- Figura 5: Pseudocódigo do Algoritmo Genético Padrão..... 29
- Figura 6: Ilustração da interação entre o algoritmo genético e a árvore de decisão, utilizados na seleção de atributos. A população inicial representa soluções aleatórias ao problema. Na fase de avaliação, uma árvore de decisão é treinada para avaliar cada uma das soluções candidatas. O resultado da avaliação é utilizado na reprodução dos indivíduos que irão compor a população atual. Este ciclo se repete até um critério de parada ser satisfeito. Adaptado de (Favretto et al., 2006)..... 30
- Figura 7: Exemplo de um cromossomo com codificação binária. Cada posição do cromossomo indica um gene que representa um atributo do conjunto de treinamento. Quando o valor do gene é 1, o atributo representado pelo respectivo gene será utilizado na solução candidata. Quando o gene assume o valor 0, o respectivo atributo não será utilizado na solução candidata..... 31
- Figura 8: Exemplos de cromossomos com codificação real, utilizando *setsize* = 5. Cada posição do cromossomo representa um atributo do conjunto de treinamento. O valor de cada posição indica o atributo que será utilizado na solução candidata. 32
- Figura 9: Diagrama de casos de uso: Visão Global dos Atores e *Use-Cases*. Os atores Administrador, Médico e Colaborador são Usuários do sistema e interagem com as *use-cases* que cada papel tem permissão de realizar. O ator visitante não é um Usuário cadastrado no sistema, podendo visualizar somente os conteúdos públicos. 39

Figura 10: Diagrama de Colaboração do caso de uso Cadastrar Usuário – Fluxo Básico. Representa todos os eventos que o Ator Administrador realiza para criar um novo usuário, descrevendo as entradas de dados necessárias e as mensagens para cada evento.	40
Figura 11: Diagrama de colaboração do caso de uso Cadastrar Categorias – Fluxo Básico. Representa os eventos que o ator Administrador realiza para criar uma nova categoria, descrevendo as entradas de dados necessárias e as mensagens para cada evento.	41
Figura 12: Diagrama de colaboração do caso de uso Criar Tipo de Conteúdo – Fluxo Básico. Representa todos os eventos que o ator Administrador deve realizar para criar um novo tipo de conteúdo no sistema, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento. .	42
Figura 13: Diagrama de colaboração do caso de uso Cadastrar Paciente – Fluxo Básico. Representa todos os eventos que o ator Usuário, ou seja, Administrador, Médico ou Colaborador, deve realizar para cadastrar um novo paciente no sistema, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento.	43
Figura 14: Diagrama de colaboração do caso de uso Enviar Submissão de Seleção de Atributos – Fluxo Básico. Representa todos os eventos que o ator Usuário (Administrador, Médico ou Colaborador) deve realizar para submeter uma seleção de atributos ao sistema, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento.	44
Figura 15: Diagrama de colaboração do caso de uso Enviar Predição da Resposta à DMARDs – Fluxo Básico. Representa todos os eventos que o ator Usuário (Administrador, Médico ou Colaborador) deve realizar para submeter uma predição da resposta à DMARDs ao sistema, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento.	45
Figura 16: Diagrama de colaboração do caso de uso Alterar Dados – Fluxo Básico. Representa todos os eventos que o ator Usuário (Administrador, Médico ou Colaborador, deve realizar para alterar dados previamente cadastrados no sistema, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento.	46
Figura 17: Diagrama de colaboração do caso de uso Visualizar Conteúdo – Fluxo Básico. Representa todos os eventos que o ator Usuário (Administrador, Médico ou Colaborador) deve realizar para visualizar um conteúdo do sistema, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento.	47
Figura 18: Diagrama de colaboração do caso de uso Visualizar Conteúdo Público – Fluxo Básico. Representa os eventos que o ator Visitante deve realizar para visualizar um conteúdo público, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento. Visualizar Conteúdo Público é o único caso de uso que não requer um usuário cadastrado no sistema.	48

Figura 19: Visão geral das funcionalidades do Sistema ARIA. As principais funcionalidades são: Cadastro de Pacientes, Seleção de Atributos e Predição da Resposta à DMARDs.	52
Figura 20: Página principal do sistema ARIA. Nesta página, pode ser observada a apresentação do projeto, informações sobre o grupo de pesquisa, o formulário de contato e o formulário de <i>Login</i> no sistema.....	53
Figura 21: <i>Front Page</i> do Sistema ARIA. Apresenta os diferentes tipos de conteúdos e diversas formas de acesso e visualização. a. Pesquisar. b. Pacientes. c. Ferramenta. d. Resultados. e. Usuário.	54
Figura 22: Formulário de Cadastro de Paciente do Sistema ARIA. a. Dados Pessoais. b. Parâmetros Clínicos devem ser preenchidos.	57
Figura 23: Continuação do Formulário de Cadastro de Pacientes. c. Parâmetros Laboratoriais. d. Informações de <i>Microarray</i> . e. Classe Resposta ao Tratamento.....	58
Figura 24: Lista de Pacientes do Sistema ARIA. a. <i>Link</i> para a página contendo todas as informações do paciente (ver Figura 24). b. Quadro resumo sobre o paciente.....	60
Figura 25: Visualização de um registro de um paciente.....	61
Figura 26: Filtro de Pacientes do Sistema ARIA. A pesquisa realizada neste exemplo apresenta os pacientes que possuem as seguintes características: Nível de fator reumatóide Negativo, Possui expressão gênica determinada através de <i>microarray</i> Sim e a Atividade da doença Alta.	63
Figura 27: Lista de Pacientes com Expressão Gênica do Sistema ARIA. As informações apresentadas nesta lista são: ID do Paciente, Tabagismo, Haplótipo do Epítipo Compartilhado, Nível de Fator Reumatóide, Nível da Proteína C Reativa, Nível de Anticorpos anti-CCP e Atividade da Doença.....	64
Figura 28: Visão geral do funcionamento da ferramenta de seleção de atributos disponível no sistema ARIA.....	66
Figura 29: Exemplo de um arquivo no formato arff. @relation define o nome do conjunto de dados. @attribute é utilizado para definir a lista de atributos. @data define os dados do conjunto.....	67
Figura 30: Ferramenta de Seleção de Atributos do Sistema ARIA. a. Selecionar o tipo de dado de treinamento a ser utilizado para a seleção de atributos. b. Parâmetros da busca genética. c. Selecionar o algoritmo de classificação a ser utilizado na seleção de atributos. d. Parâmetros dos algoritmos de classificação.....	68
Figura 31: Resultado da Seleção de Atributos utilizando como classificador uma árvore de decisão. a. Apresenta os atributos selecionados. b. Apresenta a árvore de decisão gerada. c. Apresenta a avaliação do classificador nos dados	

de treinamento após a seleção de atributos. d. Apresenta a matriz de confusão do classificador. e. Sensibilidade e Especificidade do classificador..... 70

Figura 32: Resultado da Seleção de Atributos, utilizando como classificador uma rede neural artificial. a. Apresenta os atributos selecionados. b. Apresenta a avaliação do classificador nos dados de treinamento após a seleção de atributos. c. Apresenta a matriz de confusão do classificador. d. Sensibilidade e Especificidade do classificador..... 71

Figura 33: Ferramenta de Predição da Resposta de um Paciente à DMARDs. O usuário deve selecionar um paciente e um modelo para submeter uma predição..... 72

Figura 34: Resultado da Predição da Resposta de um Paciente à DMARDs. . 73

Figura 35: Lista de submissões de seleção de Atributos. A lista de seleção de atributos apresenta as informações da seleção, como a data de submissão, tipo de dados, algoritmo de classificação, e o resultado da seleção, contendo o número de atributos, os atributos selecionados e o arquivo de resultado, que contém a avaliação e a matriz de confusão do classificador gerado..... 74

Figura 36: Lista de Seleção de Atributos realizadas através do sistema ARIA, nas quais Árvores de Decisão foram utilizadas como classificadores e os dados de treinamento foram dados clínicos..... 75

Figura 37: Lista de Seleção de Atributos realizadas através do sistema ARIA, nas quais Redes Neurais Artificiais, MLP foram utilizadas como classificadores e como dados de treinamento, os dados clínicos..... 76

Figura 38: Lista de Seleção de Atributos realizadas através do Sistema ARIA, nas quais os dados de treinamento foram dados de expressão gênica..... 77

Figura 39: Lista de Predições realizadas através do sistema ARIA. a. Identificador da Predição. b. Identificador do Paciente, link para a página contendo as informações do paciente. c. Modelo utilizado na predição, link para a página contendo as informações do modelo. d. Predição gerada..... 77

Figura 40: Árvore de decisão resultante das Submissões 257 e 259. Foram selecionados, nestas submissões, os atributos: VHS, DAS 28, gravidade da doença e PCR. Apenas o atributo gravidade não foi utilizado na construção da árvore. 79

Figura 41: Árvore de decisão gerada pela Submissão 258. Esta submissão selecionou os atributos: número de critérios ACR, VHS, DAS 28, manifestações extra-articulares e PCR, mas apenas DAS 28 e VHS foram utilizados na construção da árvore..... 80

Lista de Tabelas

Tabela 1: Permissões dos diferentes níveis de acesso ao sistema.	20
Tabela 2: Diferentes Tipos de Conteúdos disponíveis no Sistema ARIA.	22
Tabela 3: Categorias e termos disponíveis no Sistema ARIA.	23
Tabela 4: Características dos Pacientes*.....	50
Tabela 5: Métricas de avaliação dos modelos gerados utilizando como classificadores, Árvores de Decisão e dados clínicos como dados de treinamento.	78
Tabela 6: Métricas de avaliação no conjunto de teste dos modelos gerados utilizando como classificadores, Árvores de Decisão.....	81
Tabela 7: Métricas de avaliação dos modelos gerados utilizando como classificadores, Redes Neurais Artificiais MLP e dados clínicos como dados de treinamento.	82
Tabela 8: Métricas de avaliação no conjunto de teste dos modelos gerados utilizando como classificador, Redes Neurais Artificiais.....	83
Tabela 9: Métricas de avaliação dos modelos gerados utilizando dados de expressão gênica como dados de treinamento.	84

Lista de Abreviaturas e Siglas

ACR	<i>American College of Rheumatology</i> (Colégio Americano de Reumatologia)
AD	Árvore de Decisão Indutiva
AE	Algoritmos Evolucionários
AG	Algoritmo Genético
AM	Aprendizado de Máquinas
API	<i>Application Programming Interface</i> (Interface de Programação de Aplicativos)
AR	Artrite Reumatóide
ARIA	Artrite Reumatóide & Inteligência Artificial
CAD	<i>Computer Aided Diagnosis</i> (Diagnóstico Auxiliado por Computador)
CCK	<i>Content Construction Kit</i> (Quite de Construção de Conteúdo)
CCP	<i>Cyclic Citrulinated Peptide</i> (Peptídeo Citrulinado Cíclico)
CMS	<i>Content Management System</i> (Sistema Gerenciador de Conteúdo)
CSS	<i>Cascading Style Sheets</i> (Folhas de Estilo Encadeadas)
DAS 28	<i>Disease Activity Score</i> (Índice de Atividade da Doença)
DMARDs	<i>Disease-Modifying Anti-Rheumatic Drugs</i> (Drogas Anti-Reumáticas Modificadoras da Doença)
FAN	Fator anti-nuclear
FR	Fator Reumatóide
GBi	Grupo de Bioinformática do Departamento de Genética - FMRP
GPL	<i>General Public Licence</i> (Licença Pública Geral)
HC	Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto
HLA	Antígeno Leucocitário Humano
HTML	<i>Hypertext Markup Language</i> (Linguagem de Marcação de Hipertexto)
IA	Inteligência Artificial
MLP	<i>Multilayer Perceptron</i> (Perceptron de Múltiplas Camadas)
MySQL	<i>My Structured Query Language</i> (Linguagem de Consulta Estruturada)
PCR	Proteína C Reativa
PERL	<i>Practical Extraction and Report Language</i> (Linguagem Prática de Extração de Relatório)
PHP	<i>Hypertext Preprocessor</i> (Preprocessador de Hipertexto)
RN	Rede Neural Artificial
SAGE	<i>Serial Analysis Gene Expression</i> (Análise Serial de Expressão Gênica)
SE	<i>Shared Epitope</i> (Epítopo Compartilhado)
SGBD	Sistema Gerenciador de Banco de Dados

SQL	<i>Structured Query Language</i> (Linguagem de Consulta Estruturada)
TNF	<i>Tumor Necrosis Factor</i> (Fator de Necrose Tumoral)
UML	<i>Unified Modeling Language</i> (Linguagem de Modelagem Unificada)
URL	<i>Uniform Resource Locator</i> (Localizador Uniforme de Recursos)
VHS	Velocidade de Hemossedimentação
XHTML	<i>eXtensible Hypertext Markup Language</i> (Linguagem de Marcação de Hipertexto Extensível)

Sumário

1. INTRODUÇÃO	2
1.1 Doenças auto-imunes	2
1.2 Artrite Reumatóide	3
1.3 Bioinformática	7
1.4 Inteligência Artificial	8
1.5 Sistemas <i>Web</i> e Sistema Gerenciador de Conteúdos	10
2. OBJETIVOS	14
3. MATERIAIS E MÉTODOS.....	16
3.1 Aquisição dos Dados de Pacientes	16
3.2 Desenvolvimento do Sistema	17
3.2.1 Banco de Dados.....	19
3.2.2 Política de Segurança	20
3.2.3 Criação de Tipos de Conteúdos	21
3.2.4 Categorias e termos	22
3.3 Desenvolvimento da Ferramenta.....	24
3.3.1 Árvore de Decisão Indutiva	24
3.3.2 Redes Neurais Artificiais	26
3.3.3 Algoritmos Genéticos	28
3.3.4 Seleção de Atributos	29
3.3.5 Weka	33
4. RESULTADOS E DISCUSSÃO	50
4.1 Caracterização da Amostra de Pacientes	50
4.2 Desenvolvimento do Sistema	51
4.2.1 Pacientes.....	55
4.2.2 Ferramentas	65
4.2.3 Resultados	73
4.3 Análise dos Resultados.....	78
5. CONCLUSÕES	88
6. REFERÊNCIAS BIBLIOGRÁFICAS	91
7. ANEXOS.....	96

INTRODUÇÃO

1. INTRODUÇÃO

1.1 Doenças auto-imunes

As doenças auto-imunes são consideradas complexos desordens imunológicas, caracterizadas pela diminuição da tolerância aos componentes do próprio organismo, o que resulta em respostas imunes contra células e tecidos próprios. Essa resposta é causada por uma falha no mecanismo de distinção entre antígenos constituintes do organismo e antígenos externos, como vírus e bactérias (Leech, 1998).

Estas doenças compõem um grupo de doenças que ainda não são completamente esclarecidas. Dentre as doenças auto-imunes, as que apresentam maiores taxas de incidência são: diabetes melitus tipo 1, lúpus eritematoso sistêmico e artrite reumatóide. A Artrite Reumatóide é foco deste trabalho.

As doenças auto-imunes têm origem multifatorial. Fatores ambientais e fatores intrínsecos ao organismo, como predisposição genética, alterações hormonais ou alterações no sistema imuno-regulatório estão relacionadas com o desencadeamento da auto-imunidade. Suspeita-se que defeitos genéticos desempenhem papel importante na etiologia destas doenças. Porém, diferentemente de outras doenças genéticas, as doenças auto-imunes não são causadas pelo defeito de um único gene, mas sim pela disfunção de um grupo complexo de genes (Jawaheer e Gregersen, 2002; Karopka et al., 2006). Apesar da diversidade clínica das diversas doenças auto-imunes, elas apresentam similaridade quanto à disfunção do sistema imune (Davidson e Diamond, 2001).

1.2 Artrite Reumatóide

A Artrite Reumatóide (AR) é uma doença auto-imune, inflamatória, sistêmica e crônica, marcada por inflamação nas membranas sinoviais e estruturas articulares que podem levar à destruição ósteo-articular (Feldmann et al., 1996). É uma doença com prevalência estimada na população mundial de 0,5% a 1%, com incidência anual de 0,02% a 0,05%. Apresenta predomínio de 2 a 3 vezes maior no sexo feminino e pico de incidência por volta de 50 anos de idade (Alamanos e Drosos, 2005). Um estudo realizado em 2004 estimou a prevalência da AR na população brasileira em 0,46% (Senna et al., 2004).

A AR é a forma mais comum de poliartrite crônica e é caracterizada por envolvimento simétrico e bilateral das articulações periféricas, principalmente das mãos e dos pés. O comprometimento das articulações é devido à intensa inflamação da membrana sinovial, resultando na invasão da cartilagem e, conseqüentemente, a erosão óssea (Choy e Panayi, 2001). Estes danos ocorrem no início da instalação da doença, por volta do segundo ano, causando sérias deformidades e incapacidades funcionais importantes, o que faz com que a AR se torne economicamente relevante, uma vez que estas incapacidades interferem significativamente nas atividades profissionais dos pacientes (Cooper, 2000). Além do envolvimento articular, manifestações extra-articulares sistêmicas como febre, perda de peso, anemia e náuseas, também são notadas frequentemente nos pacientes.

Trata-se de uma doença multifatorial, resultado da interação entre fatores genéticos e ambientais. Os principais fatores de risco incluem a susceptibilidade genética, idade, sexo, tabagismo, agentes infecciosos, fatores hormonais, dieta, fatores socioeconômicos e étnicos (Alamanos e Drosos, 2005). Um dos fatores de susceptibilidade genética associado a AR são os genes HLA-DR do complexo de histocompatibilidade maior, responsáveis pela apresentação de antígenos às células T. Estes genes, localizados no cromossomo 6, são altamente polimórficos, porém

apresentam uma região conservada, o epítopo compartilhado. Esta região compreende os aminoácidos 70 a 74 da cadeia DR beta. Os genes HLA-DR são associados à susceptibilidade e a doença mais agressiva (Bértolo et al., 2001; Delgado-Vega e Anaya, 2007).

O diagnóstico da AR é feito baseando-se nos critérios estabelecidos pelo Colégio Americano de Reumatologia (ACR) (Arnett et al., 1988). Foram definidos sete critérios, sendo que o paciente deve apresentar pelo menos quatro deles, para que possa ter o diagnóstico de AR confirmado. Esses critérios incluem rigidez matinal, artrite em pelo menos três articulações, artrite nas mãos, artrite simétrica, nódulos reumatóides, fator reumatóide¹ e alterações radiográficas. Estes critérios foram estabelecidos em pacientes com a doença estabelecida, apresentando tempo médio de doença de 8 anos. Muitas discussões têm surgido sobre a necessidade de se utilizar novos critérios para que o diagnóstico possa ser confirmado nos estágios iniciais da doença e, assim, o tratamento possa ser iniciado o mais rápido possível. Nos estágios iniciais da AR, os critérios ACR podem estar ausentes (Aletaha et al., 2005). As alterações radiológicas, os nódulos reumatóides e o fator reumatóide (FR) positivo são menos freqüentes no início do estabelecimento da doença que em estágios mais avançados. O diagnóstico precoce e o início da administração de DMARDs em até três meses é de extrema importância para evitar danos às articulações, que ocorrem nos dois primeiros anos da doença e são muitas vezes irreversíveis (O'dell, 2004).

Quanto à gravidade da doença, a AR costuma ser dividida em leve, moderada e grave. Na AR leve o paciente apresenta dor, em pelo menos três articulações com

¹ Fatores reumatóides: são anticorpos dirigidos contra determinados antígenos e são encontrados no soro de cerca de 80% dos pacientes com artrite reumatóide moderada ou grave, mas em apenas 20% dos pacientes com artrite reumatóide juvenil. Os fatores reumatóides podem ser anticorpos IgM, IgG ou IgA, embora os testes sorológicos meçam somente o IgM (Dorland et al., 1999).

sinais de inflamação, nenhuma manifestação extra-articular, fator reumatóide costumeiramente negativo, elevação dos reatores de fase aguda e nenhuma evidência de erosão ou perda de cartilagem no estudo radiográfico (Dorland et al., 1999). Na AR moderada, o paciente apresenta de 6 a 20 articulações acometidas, doença comumente restrita às articulações, elevação de reatores de fase aguda, fator reumatóide positivo e evidência de inflamação no estudo radiográfico. Quando a AR é grave o paciente apresenta mais de 20 articulações persistentemente acometidas, elevação dos reatores de fase aguda, anemia crônica, hipoalbuminemia², fator reumatóide positivo, manifestações extra-articulares e estudos radiográficos demonstrando erosões e perda de cartilagem (Brasil, 2006).

Inicialmente a avaliação dos pacientes é feita através de sintomas de atividade da doença, os quais incluem graduação da dor, rigidez matinal, estado funcional do paciente, avaliação das articulações acometidas e a presença de manifestações extra-articulares. Avaliam-se também marcadores de inflamação sinovial, como a proteína C reativa e a velocidade de hemossedimentação, e a presença de dano ao estudo radiográfico (Guidelines, 2002).

O tratamento da AR tem como objetivo prevenir e controlar os danos articulares, melhorar a capacidade funcional do paciente, reduzir a dor e, principalmente, busca alcançar a remissão clínica. A terapia inicia-se com a educação e o aconselhamento do paciente. A terapia medicamentosa é o principal tratamento para a AR e, normalmente, é constituída por anti-inflamatórios não-esteróides, corticosteróides e drogas anti-reumáticas modificadoras da doença (DMARDs). As DMARDs são um grupo de medicamentos com potencial para reduzir e prevenir dano articular e preservar a integridade e a funcionalidade das articulações. As DMARDs sintéticas incluem os agentes cloroquina, hidroxicloroquina, sulfassalazina,

² Hipoalbuminemia: situação na qual os níveis de albumina no sangue estão abaixo do normal.

metotrexato, ciclosporina e leflunomida. Alguns estudos demonstraram que o metotrexato é o medicamento que apresenta a maior taxa de indução de resposta em pacientes com AR e este medicamento passou a ser o mais utilizado no início da terapêutica (O'dell, 2004). A dose administrada de metotrexato é no máximo 20 mg/m²/semana, durante 3 a 6 meses. Após esse período, o paciente é avaliado novamente e se não houver resposta, a terapia deve ser modificada, adicionando-se, por exemplo, outro DMARD como o Leflunomida ou uma droga anti-citocina e, desta forma, o paciente é considerado não respondedor à DMARDs.

Drogas anti-citocinas estão, atualmente, disponíveis para o tratamento da AR e são conhecidas por terapia biológica. Estão disponíveis três drogas que atuam como inibidores de TNF- α (Fator de Necrose Tumoral α), o Infliximabe, anticorpo monoclonal quimérico anti-TNF, o Etanercepte, que é um receptor solúvel de TNF e o Adalimumabe, anticorpo monoclonal anti-TNF totalmente humanizado. Uma quarta droga disponível é o Anakinra que inibe a ação da interleucina-2 (Choy e Panayi, 2001). Estas drogas constituem o tratamento mais eficiente disponível para a AR até o momento, mas ainda assim existem pacientes que não respondem a estes medicamentos e apresentam a doença descontrolada (O'dell, 2004).

O estabelecimento do tratamento adequado a cada paciente individualmente, se torna um grande desafio na AR. Uma estratégia possível é a utilização de biomarcadores para prever a evolução da doença. O fator reumatóide que é o biomarcador utilizado para diagnóstico da doença, é encontrado em 80% dos pacientes com AR estabelecida, mas é encontrado também em outras doenças. Porém, outros biomarcadores estão disponíveis atualmente e quando utilizados em conjunto aumentam o poder do prognóstico. Os anticorpos anti-peptídeos cíclicos citrulinados (anti-CCP) estão presentes em 79% dos pacientes com AR e se apresentam no início do estabelecimento da AR (Saleem et al., 2006). Outro biomarcador utilizado é a proteína C reativa, que é um marcador de inflamação de

fase aguda. Os biomarcadores estão se tornando uma ferramenta essencial na avaliação da AR.

1.3 Bioinformática

A Bioinformática tem crescido exponencialmente devido, principalmente, a grande quantidade de dados gerados pela Biologia Molecular (Baldi e Brunnak, 1998). O Projeto Genoma, iniciado em 1992 que teve como objetivo sequenciar o genoma humano, popularizou a Bioinformática, o que fez com que se intensificasse a pesquisa por novas técnicas e ferramentas computacionais para resolução de problemas em Biologia, especialmente os problemas da Biologia Molecular e Genética (Meidanis e Setúbal, 1997).

Recentemente, diversos grupos de pesquisas têm proposto projetos para o sequenciamento de genomas de diferentes espécies. Feito o sequenciamento dos genes de uma espécie, os cientistas se deparam com um problema muito mais complexo: a determinação da expressão gênica. Determinar o que cada gene faz dentro da célula é de fundamental importância para que a imensa quantidade de informação armazenada através do sequenciamento tenha valor prático para, por exemplo, o desenvolvimento de fármacos. Porém, a realização desta tarefa de forma manual, em laboratórios, mostra-se impraticável e muito custosa devido, principalmente, à grande quantidade de dados gerados.

Existem várias tecnologias para a determinação da expressão gênica em larga escala como *arrays* de oligonucleotídeos (Gene Chips), *microarrays* de cDNA e análise serial de expressão gênica (SAGE). Estas técnicas geram dados que podem ser usados para desenvolver uma compreensão mais completa da função, da regulação e das interações dos genes (Quackenbush, 2001).

A primeira tarefa relacionada à expressão gênica está em analisar os dados derivados das experiências de disposição do DNA. A estrutura numérica dos dados da expressão deve ser analisada por um método de *cluster* sem conhecimento biológico. Métodos tais como clusters hierárquicos, mapas auto-organizáveis (SOMs), recozimento simulado (*simulated annealing*) e k-médias têm sido empregados para a realização desta tarefa (Huang et al., 2003). Técnicas de Inteligência Artificial capazes de extrair informações sobre os genes de forma automática se tornam extremamente interessantes neste caso (Baldi e Brunak, 1998).

1.4 Inteligência Artificial

A Inteligência Artificial é uma área da ciência da computação dedicada a investigar formas ou métodos de habilitar o computador a resolver problemas e tarefas, nas quais até o momento, um ser humano teria um desempenho superior. (Rich e Knight, 1991). Constitui-se de um conjunto de tecnologias computacionais que fazem com que o computador possa executar habilidades humanas, como raciocínio, decisão e resolução de problemas. É uma área que teve um grande desenvolvimento nos últimos anos, devido a sua utilização em produtos comerciais, como por exemplo, em jogos de vídeo game.

Técnicas inteligentes podem ser aplicadas nos mais variados problemas, incluindo problemas biológicos, como por exemplo, no alinhamento de sequências biológicas. Neste caso, o objetivo é alinhar sequências relacionadas, dado uma função de custo. Com isto, explora-se o grau de similaridade entre cadeias de DNA, RNA e proteínas. Um grau alto de similaridade indica uma alta probabilidade das funções executadas pelas moléculas comparadas serem semelhantes. Os resultados obtidos permitem a descoberta, por exemplo, de funções de novas proteínas ou a identificação de possíveis mutações genéticas. O alinhamento pode ser utilizado também para

medir a distância evolutiva entre duas ou mais espécies (Gondro e Kinghorn, 2007; Nguyen et al., 2002; Notredame e Higgins, 1996; Notredame et al., 1997).

Outro problema de biologia molecular em que técnicas de IA podem ser utilizadas é na identificação de genes em sequências de DNA, ou seja, reconhecer e localizar cada gene em uma seqüência. Entre as abordagens utilizadas na realização desta tarefa, cita-se a determinação de padrões que podem ser relacionados à localização de genes, tais como posições de início de tradução das sequências de DNA ou mRNA (Craven e Shavlik, 1994). Em outro método, intitulado busca por conteúdo, procura-se por padrões nas sequências que determinem regiões do DNA que são efetivamente codificadas em proteínas. A medição dos níveis de ativação das sequências durante os diversos processos biológicos também pode ser utilizada para a determinação dos genes (D'haeseleer et al., 1999).

As técnicas de IA podem ser aplicadas também, na análise de dados de expressão gênica, por exemplo, dados de *microarray*. Estes dados normalmente apresentam um grande número de genes ou atributos e, relativamente, poucas amostras (Souto et al., 2003). Assim, o objetivo, neste caso, é agrupar ou selecionar genes baseando-se na similaridade dos valores da expressão gênica. Uma vez feito o agrupamento ou a seleção dos genes, pode-se classificá-los ou efetuar a busca por novas classes, por exemplo, sub-classes de uma determinada doença (Golub et al., 1999).

Outra aplicação que tem utilizado técnicas inteligentes são os esquemas de diagnóstico auxiliados por computador (*Computer-aided Diagnosis*). Neste tipo de sistema, uma análise computadorizada gera informações que auxiliam profissionais da área de saúde a tomar decisões com relação à definição de diagnósticos e terapias. Sistemas CAD são muito difundidos na radiologia, como auxílio na detecção e diagnóstico de lesões em imagens médicas (Vyborny e Giger, 1994). Recentemente, outras áreas da saúde têm investido na pesquisa destes sistemas aplicados às mais

diversas patologias como em pediatria, doenças neurológicas, doenças pulmonares e medicina em geral.

1.5 Sistemas Web e Sistema Gerenciador de Conteúdos

O crescimento e popularização da Internet no Brasil e no mundo permite que os mais diversos sistemas utilizem a infra-estrutura da Internet com plataforma de trabalho. Os sistemas *Web* são sistemas que funcionam da mesma forma que as soluções de *software* tradicionais, apresentando uma grande vantagem, a portabilidade oferecida por estes sistemas, ou seja, estes sistemas podem ser acessados por qualquer computador conectado a Internet, em qualquer lugar do mundo. Outras vantagens são a manutenção e atualização em tempo real, independência de sistema operacional e a abrangência de sistema *Web* que pode vir a ser mundial.

Os sistemas *Web* são utilizados nas mais diversas áreas inclusive na área médica e biológica. Atualmente, existem milhares de bancos de dados biológicos disponíveis para que pesquisadores de todas as partes do mundo possam utilizar como instrumento de busca, possibilitando um melhor direcionamento das pesquisas, economizando tempo e dinheiro.

Um sistema *Web* é uma aplicação dinâmica, que é composta por um banco de dados, um servidor web e por *scripts* que fazem uso das informações armazenadas no banco e as apresentam ao usuário por meio de um navegador. Um banco de dados é uma forma de armazenamento de dados, de maneira organizada e estruturada, a fim de prover facilidade nas consultas, atualizações e manutenção dos dados em questão. Em um banco de dados relacional, as informações são estruturadas em um conjunto de tabelas e os dados são organizados em linhas, onde cada linha representa um registro no banco de dados. Uma linha pode conter várias informações separadas

(campos). Cada campo no banco de dados contém uma informação distinta. Para controlar o acesso ao banco de dados é necessária a utilização de um Sistema de Gerenciamento de Banco de Dados (SGBD). O SGBD é o *software* responsável pelo gerenciamento da base de dados, ou seja, gerencia o acesso, a manipulação e a organização dos dados (Ferrari, 2007).

O MySQL (*My Structured Query Language*) é um sistema gerenciador de banco de dados relacional *open source*, multitarefa e multiusuário, que oferece mecanismos para armazenar os dados em tabelas e uma interface para a execução de consultas SQL. O SQL é a linguagem padrão utilizada para acessar banco de dados. As consultas SQL podem ser executadas por meio da linguagem de programação PHP e visualizadas no navegador *Web*. O PHP (*Hypertext Preprocessor*) é uma linguagem *open source* muito utilizada na *Web*. Através de um *script* PHP é possível se conectar ao SGBD MySQL, realizar as consultas SQL necessárias e criar interfaces para que o usuário possa visualizar os dados armazenados no banco (Meloni, 2000).

Um sistema *Web* deve ser implementado de forma dinâmica e gerenciada. Para isso, um bom planejamento é imprescindível. Existem ferramentas chamadas Gerenciadores de Conteúdos (CMS) que auxiliam na implementação e no gerenciamento destes sistemas. Um gerenciador de conteúdos é uma ferramenta utilizada para lidar com os diferentes métodos de publicação na *Web*, permite integrar e automatizar todos os processos relacionados à criação, catalogação, indexação, personalização, controle de acesso e disponibilização de conteúdos na *Web*. Um CMS, geralmente, pode ser configurado, adicionando ou removendo características, para que o sistema final possua apenas as características que são necessárias (Douglass et al., 2006).

O Drupal é um gerenciador de conteúdos *open source* que provê um ambiente estruturado e funcionalidades básicas para o desenvolvimento de *Web sites*. É uma

plataforma de desenvolvimento robusta, eficiente e flexível. É mantida por uma comunidade de milhares de desenvolvedores em todo o mundo (Mercer, 2006). Sua plataforma pode ser facilmente estendida através de módulos - blocos de código que provêm funcionalidade extra ou aprimoramentos ao desenvolvimento do *Web site*, permitindo novas implementações sem comprometer o sistema. Alguns módulos, chamados módulos-padrão, são instalados juntamente com o Drupal, enquanto outros podem ser obtidos, através do *site* do Drupal e instalados, individualmente. Os módulos extras são a base para estender e implementar soluções de gerenciamento de conteúdos personalizadas.

Algumas características oferecidas pelo Drupal são muito úteis em sistemas de Bioinformática, como os recursos de taxonomia (categorização do banco de dados), diferentes tipos de conteúdos, relacionamentos (conteúdo *versus* blocos), diferentes formatos de entrada de dados (Texto, HTML, PHP, etc.), URLs alternativas, indexação total para sistema de busca, manipulação de expressões para idiomas, códigos extremamente limpos, *Template* (temas) em PHP, XHTML, CSS, estatísticas, rastreador, controle de acesso definido por papéis e, principalmente, os *Snippets*, fragmentos de códigos que proporcionam customizações, associados com PHP, SQL e arquivos do *template*.

OBJETIVOS

2. OBJETIVOS

O presente projeto teve como objetivo o desenvolvimento de um sistema computacional para a *Web*, para o auxílio ao diagnóstico de Artrite Reumatóide, utilizando técnicas de Inteligência Artificial. O auxílio se dá na construção de modelos preditivos da resposta de pacientes portadores de AR ao tratamento com DMARDs.

O sistema é composto por um banco de dados de informações de pacientes portadores de AR e por ferramentas baseadas em técnicas de IA, como Redes Neurais Artificiais, Árvores de Decisão e Algoritmos Genéticos. O banco de dados armazena informações clínicas, laboratoriais e de expressão gênica de pacientes portadores de AR, previamente diagnosticados e com tratamento iniciado. As ferramentas de IA utilizam as informações armazenadas no banco para a seleção de características relevantes para a classificação destes pacientes de acordo com a resposta ao tratamento utilizando Drogas Anti-Reumáticas Modificadoras da Doença. Uma vez selecionadas, estas características, clínicas, laboratoriais ou de expressão gênica, podem ser utilizadas na predição da resposta ao tratamento com DMARDs de novos pacientes portadores de AR.

O sistema gera modelos de predição da resposta à DMARDs, executa uma avaliação e os disponibiliza para a predição da resposta de novos pacientes. Isso proporciona uma nova abordagem na avaliação da melhor terapia para os pacientes. Poderá se tornar uma importante ferramenta para identificar precocemente quais pacientes irão evoluir de forma mais agressiva, não respondendo aos DMARDs, para que uma terapia mais adequada possa ser indicada a cada paciente individualmente, a fim de prevenir danos severos às articulações.

***MATERIAIS E
MÉTODOS***

3. MATERIAIS E MÉTODOS

3.1 Aquisição dos Dados de Pacientes

Para a construção do sistema, denominado ARIA, foi necessário coletar dados de pacientes portadores de AR.

Os dados clínicos dos pacientes foram cedidos ao projeto pelo Prof. Dr. Paulo Louzada Jr. (Departamento de Clínica Médica – Divisão de Imunologia Clínica, HC-FMRP, USP). Tais dados foram coletados durante a vigência do projeto intitulado “*Determinação de Anticorpos Anti-peptídeo Citrulinado Cíclico em Pacientes com Artrite Reumatóide e Correlação com os Alelos HLA de Classe II e a Gravidade da Doença*” sob orientação do referido Professor. Este projeto foi aprovado pelo Comitê de Ética em pesquisa, de acordo com o processo HCRP nº 2958/2005.

Os dados de expressão gênica provenientes de experimentos de *microarrays* foram cedidos pela Dra Cristina Moraes Junta (Laboratório de Imunogenética Molecular do Departamento de Genética, FMRP, USP). Estes dados foram gerados durante a vigência do projeto intitulado “*Análise da Expressão Gênica Diferencial em Pacientes com Artrite Reumatóide do Adulto Estratificado Segundo o Padrão de Suscetibilidade Genética*” sob orientação do Prof. Dr. Geraldo Passos. O referido projeto obteve aprovação do Comitê de Ética em Pesquisa, de acordo com processo HCRP 8090/2001.

Os pacientes incluídos apresentam diagnóstico de Artrite Reumatóide, definidos pelos critérios de classificação diagnóstica propostos pelo Colégio Americano de Reumatologia (Arnett et al., 1988) e estão em seguimento no ambulatório de Doenças Reumáticas da Disciplina de Reumatologia do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto – Universidade de São Paulo. Os pacientes têm idade entre 18 e 75 anos e assinaram o termo de consentimento livre e esclarecido antes de participarem dos estudos. Além disso, os pacientes foram selecionados de acordo com

a presença de alelos HLA-DRB1 incluídos no epítipo compartilhado por meio da tipificação HLA-DRB1 (*0401/ 0404/ *0101/ *0405/ *1402).

Os aspectos éticos referentes à utilização de dados de pacientes portadores de AR foram aprovados junto ao Comitê de Ética em Pesquisa do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto – USP, de acordo com o Processo HCRP nº 4880/2007.

3.2 Desenvolvimento do Sistema

O Sistema ARIA está hospedado em um servidor HP ProLiant ML150 Pentium Xeon 3.0 com 2 Gb de memória RAM, 4 discos rígidos de 160 Gb e sistema operacional Fedora 6, localizado no laboratório do Grupo de Bioinformática do Departamento de Genética da Faculdade de Medicina de Ribeirão Preto (GBi).

Está implementado sobre o CMS Drupal que é um gerenciador de conteúdo *open source* criado por Dries Buytaert, distribuído sob a licença GPL (*General Public Licence*) e mantido por uma comunidade de usuários e desenvolvedores do mundo todo. O Drupal, em sua versão 5.9, foi o CMS utilizado para a implementação do Sistema ARIA e pode ser obtido através do *site* <http://drupal.org/>.

O CMS Drupal simplifica o processo de criação, gerenciamento e publicação de conteúdos como texto, imagens, arquivos e áudio na *Web*. É uma plataforma extensível que disponibiliza um conjunto muito variado de APIs e permite a construção de módulos personalizados, para possibilitar o desenvolvimento de *sites* customizados (Douglass et al., 2006). Com isso o Drupal se torna uma plataforma bastante flexível, crescendo de forma dinâmica de acordo com a especificação da aplicação em questão (Mercer, 2006).

O Drupal gerencia todos os tipos de conteúdos e serviços atribuídos à interface *Web*: o acesso dos usuários (*Login*), os métodos de busca, a inserção de novos

dados, atualizações, alterações, entre outros. A estrutura do CMS Drupal foi implementada na linguagem PHP (Meloni, 2000) (<http://www.php.net>) e o banco de dados MySQL (<http://www.mysql.com>). O PHP corresponde a um módulo de pré-processamento de hipertexto para o servidor *Web*, permitindo ler e interpretar códigos PHP incorporados em páginas *Web*. Além disso, o PHP oferece suporte para a interação com diversos bancos de dados, incluindo o MySQL. A plataforma Drupal permite também acoplar outras linguagens de programação como Perl (*Practical Extraction and Report Language*) (Wall et al., 2001) (<http://www.perl.com>), que é uma linguagem mais popular para escrita de *scripts* CGI (Guelich e Gundavan, 2001). A linguagem Perl é uma linguagem de programação especialmente desenvolvida para processamento de texto, que no sistema, foi utilizada para a inserção automática dos dados dos pacientes que estavam previamente armazenados em arquivos Excel.

A linguagem HTML (*Hypertext Markup Language*) juntamente com a linguagem de marcação CSS (*Cascading Style Sheets*) foram utilizadas na criação das páginas *Web* que irão interagir com as ferramentas de seleção de atributos e predição da resposta dos pacientes. Estas ferramentas, por sua vez, foram implementadas utilizando o pacote Weka³ (<http://www.cs.waikato.ac.nz/ml/weka>) e a linguagem de programação orientada a objetos Java (<http://java.sun.com>).

A Figura 1 ilustra a metodologia, transparente ao usuário, empregada no desenvolvimento do sistema, a qual será apresentada de forma mais detalhada nas próximas sessões. Em (1), ilustra-se a interação do usuário com o sistema através da Internet. Em (2), é ilustrada a posição central do CMS Drupal, a linguagem PHP e o banco de dados MySQL como núcleo do sistema ARIA. Em (3), ilustra-se a utilização da linguagem Perl integrada ao banco de dados MySQL para cadastro dos pacientes de forma automatizada. Em (4), é ilustrada a integração do *software* Weka e a

³ Software *open source* desenvolvido pela Universidade de Waikato e pode ser obtido através do endereço: www.cs.waikato.ac.nz/ml/weka/.

linguagem JAVA com o núcleo do sistema para a criação da ferramenta de seleção de atributos e da ferramenta de predição da resposta à DMARDs.

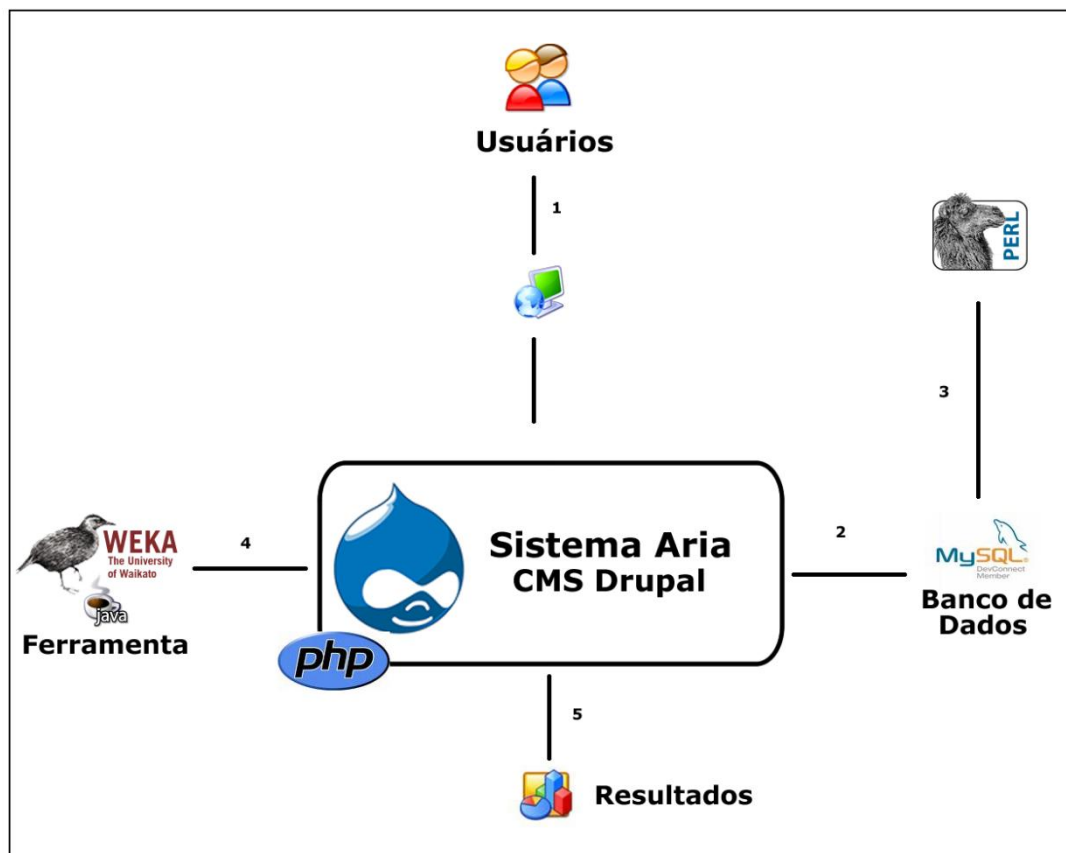


Figura 1: Visão geral do funcionamento do Sistema ARIA. 1. Interação do usuário com o Sistema através da Internet. 2. Interação do CMS Drupal e a linguagem PHP com o banco de dados MySQL. 3. Linguagem Perl e o banco de dados MySQL para cadastro dos pacientes. 4. Integração do *software* Weka e a linguagem JAVA para a criação das ferramentas de seleção de atributos e de predição. 5. Sistema ARIA gerando resultados para os usuários.

3.2.1 Banco de Dados

O Banco de Dados MySQL 5, um sistema de gerenciamento de bancos de dados relacional e de código aberto, foi o banco de dados selecionado para o desenvolvimento do sistema. O MySQL possui um sistema de privilégios e senhas que é bastante robusto, garantindo a segurança do banco de dados.

Estão armazenadas no banco todas as informações referentes aos dados e as configurações do sistema ARIA. A administração do banco de dados é feita através do

próprio sistema, utilizando o módulo *Database Administration*, que permite o gerenciamento da base de dados, a execução de funções importantes, como *backup* do banco e a execução de comandos SQL. O SQL é a linguagem padrão utilizada para manipular bancos de dados, possibilitando a criação, remoção e alteração de tabelas, inserção, alteração, remoção e buscas nos conteúdos das tabelas entre outras funções disponíveis.

3.2.2 Política de Segurança

O acesso aos conteúdos do sistema é aberto somente aos pesquisadores colaboradores do projeto e médicos envolvidos no acompanhamento dos pacientes. As medidas de segurança das informações são controladas por papéis e por módulos adicionados ao CMS Drupal. Um desses módulos é o *cck_field_permission* que permite ao administrador restringir a criação, atualização e visualização de campos específicos. O segundo módulo adicionado é o *Node Privacy By Role Quick Grant*. Este módulo permite controlar o acesso aos conteúdos do sistema baseado no tipo de papel que o usuário possui.

Os usuários do sistema podem ser cadastrados como: administrador, colaboradores, médicos e visitantes. Para cada tipo de usuário, diferentes níveis de acesso são definidos, os quais são apresentados na Tabela 1.

Tabela 1: Permissões dos diferentes níveis de acesso ao sistema.

Usuário	Permissão
Administrador	Acesso completo (Criar novo tipo de conteúdo, cadastrar usuário, cadastrar paciente, acesso às ferramentas e à área administrativa)
Médicos e Colaboradores	Acesso restrito (Cadastrar paciente, visualização dos conteúdos e acesso às ferramentas)
Visitantes	Acesso negado (Acesso somente à paginas informativas – públicas)

3.2.3 Criação de Tipos de Conteúdos

Os conteúdos específicos foram criados utilizando o módulo CCK (*Content Construction Kit*), que cria diferentes tipos de conteúdos de forma flexível, com administração no painel de controle, juntamente com o gerenciamento dos conteúdos inseridos no portal. Permite a criação de qualquer classe de conteúdo, como por exemplo, o cadastro de pacientes, além dos tipos já existentes, como matérias, páginas, fórum entre outros. Um tipo de conteúdo representa qualquer informação que, para ser inserida no banco de dados, necessite de campos específicos. O módulo CCK permite criar novos tipos de conteúdos e interagir diretamente com o banco de dados MySQL.

O administrador define os campos necessários para a inserção dos dados correspondente ao conteúdo criado, editando-os de acordo com a necessidade específica do conteúdo. Os campos podem ser texto simples, data e hora, imagem, área para composição de mensagem, lista de seleção, caixa de seleção, URL, galeria de imagem, endereço eletrônico, recursos de mídia e arquivos anexos.

Após criar um novo tipo de conteúdo, foram adicionados os tipos de campos necessários e algumas configurações sobre instruções de preenchimento, descrição, opções de visualização em listagem de conteúdos, requisição do campo e o valor do peso para a ordenação.

Para o desenvolvimento do sistema ARIA foi necessário criar três tipos de conteúdos específicos: Cadastro de Pacientes, Submissão de Seleção de Atributos e Predição da Resposta à DMARDs (Tabela 2). O tipo de conteúdo **Cadastro de Pacientes** foi criado para que os dados dos pacientes pudessem ser cadastrados no banco. Os tipos de conteúdos **Submissão de Seleção de Atributos** e **Predição da Resposta à DMARDs** foram criados para armazenar no banco de dados informações referentes às ferramentas de seleção de atributos e de predição da resposta à

DMARDs. Somente o usuário com privilégios administrativos tem permissão para criar, editar e apagar os tipos de conteúdos.

Tabela 2: Diferentes Tipos de Conteúdos disponíveis no Sistema ARIA.

Conteúdo	Descrição
Cadastro de Pacientes	Formulário de cadastro de um paciente portador de Artrite Reumatóide
Páginas	Páginas estáticas, como página de contato
Predição da Resposta à DMARDs	Formulário utilizado para submeter uma predição da resposta à DMARDs de um determinado paciente, através de um classificador previamente treinado
Submissão de Seleção de Atributos	Formulário utilizado para realizar a seleção de atributos, através do treinamento de um classificador (Árvore de Decisão/Rede Neural Artificial)

3.2.4 Categorias e termos

As categorias são utilizadas para classificar os conteúdos inseridos no sistema. A organização das categorias hierárquicas e a adição de sub-categorias para os conteúdos, relacionado em sistemas de nós, foram realizadas através do módulo *Taxonomy*. Neste módulo, as classes de categorias são chamadas de “Vocabulários”, que contêm seus respectivos termos, compondo a estrutura de categorias do sistema. Esta tarefa de categorização é importante para que as buscas, tanto por palavras chaves como através do módulo *Views*, possam ser adicionadas ao sistema. O módulo *Views* permite a criação de listas personalizadas de conteúdos específicos e interage diretamente com as categorias e termos do sistema, permitindo a criação de buscas customizadas. A Tabela 3 apresenta as categorias e seus respectivos termos disponíveis no sistema ARIA.

Tabela 3: Categorias e termos disponíveis no Sistema ARIA.

<i>Categoria</i>	<i>Termos</i>
Anticorpos anti-CCP	Positivo, Negativo, Informação Ausente
Classe Funcional	Classe I, Classe II, Classe III, Classe IV
Classe Resposta à DMARDs*	Respondeu à DMARDs, Não Respondeu à DMARDs
Atividade da Doença	Alta, Moderada, Baixa, Remissão Clínica
Determinação da Gravidade	Grave, Leve/Moderada
Erosões Radiológicas	Sim, Não
Etnia	Caucasiano, Afro descendente, Pardo, Indígena, Oriental
Expressão gênica determinada	Sim, Não
Fator anti-nuclear	Positivo, Negativo, Informação Ausente
Haploidia	SEISE, SEIX, XIX, Informação Ausente
Manifestações extra-articulares	Sim, Não, Informação Ausente
Nível de Fator Reumatóide	Título Elevado, Positivo, Negativo, Informação Ausente
Proteína C Reativa	Positivo, Negativo, Informação Ausente
Presença de DR de proteção	Sim, Não, Informação Ausente
Sexo	Masculino, Feminino
Tabagismo	Fumante, Não Fumante

* Não resposta, ou falência da terapêutica é definida como ausência de resposta a uma dose mínima de metotrexato (MTX) de 15 mg/m²/semana durante 3 a 6 meses.

3.3 Desenvolvimento da Ferramenta

Técnicas de Inteligência Artificial foram empregadas para construção das ferramentas de seleção de atributos e de predição da resposta à DMARDs. Dentre as técnicas selecionadas estão as Técnicas de Aprendizado de Máquinas (AM) que incluem: Árvores de Decisão (ADs), Redes Neurais Artificiais (RNs) e Algoritmos Genéticos (AGs). Estas técnicas de AM estudam algoritmos e métodos que melhoram seu desempenho em alguma tarefa através da experiência (treinamento).

3.3.1 Árvore de Decisão Indutiva

Árvore de decisão indutiva é um método de aprendizado de máquinas simbólico, amplamente utilizado na inferência indutiva. Um método de aprendizado simbólico é aquele no qual o classificador gerado deve ser legível e interpretável por humanos (Baranauskas, 2001).

Uma árvore de decisão é uma estrutura recursiva que pode ser definida como nós folha, que representam uma classe, nós de decisão, que realizam um teste sobre o valor de algum atributo, sendo que cada possível resultado do teste executado em um nó de decisão leva a uma sub-árvore. E cada sub-árvore possui a mesma estrutura da árvore (Mitchell, T., 1997).

A construção de uma árvore de decisão indutiva é feita da seguinte forma: a partir do conjunto de exemplos de treinamento, um atributo é escolhido, baseando-se em algum critério de seleção, de forma que este atributo particione o conjunto de exemplos inicialmente fornecido em subconjuntos. Para cada novo subconjunto um novo atributo é escolhido para particioná-lo novamente. Este processo de particionamento é mantido até que os subconjuntos possuam exemplos de uma única classe. Quando um subconjunto contiver apenas exemplos de uma classe, um nó folha é criado e recebe o nome da respectiva classe (Quinlan, 1986).

Na Figura 2, pode ser observada uma árvore de decisão indutiva. Os atributos selecionados para o particionamento do conjunto de treinamento, neste caso, foram: **DAS 28**, **Sexo** e **anti-CCP**. Na Figura, os atributos são representados por elipses e os valores desses atributos são apresentados nos ramos da árvore. Já os nós folhas, são representados por retângulos, contendo o valor do atributo classe e o número de exemplos cobertos pelo ramo. A classe neste exemplo é **Resposta ao Tratamento com DMARDs** E os valores do atributo classe são: **Respondeu** e **Não Respondeu**.

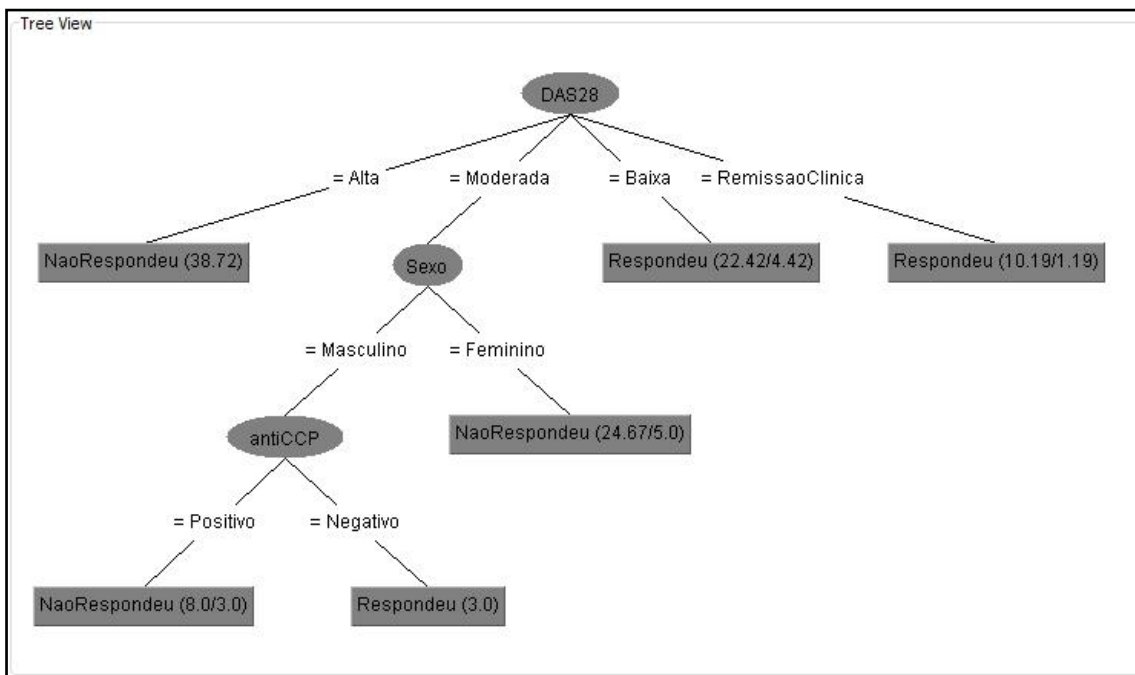


Figura 2: Exemplo de uma árvore de decisão indutiva. Os atributos utilizados para a construção da árvore foram: DAS 28, cujos valores são: alta, moderada, baixa e remissão clínica; Sexo, cujos valores são: masculino e feminino; E o atributo anti-CCP, cujos valores são positivo e negativo. O atributo classe é resposta ao tratamento com DMARDs cujos valores podem ser respondeu e não respondeu.

3.3.2 Redes Neurais Artificiais

As Redes Neurais Artificiais (RN) são modelos computacionais inspirados no cérebro humano que possuem a capacidade de aquisição e manutenção de informações. Uma Rede Neural Artificial pode ser definida como um processador paralelamente distribuído composto por unidades básicas de processamento, com a propriedade natural para armazenar conhecimento experimental e torná-lo disponível para o uso (Haykin, 2001). Fazendo uma analogia ao cérebro, as RNs são um conjunto de unidades de processamento (neurônios), interligadas por um grande número de interconexões (sinapses).

As RNs são classificadores não simbólicos, ou seja, não fornecem como resultado uma representação do classificador legível por humanos, porém apresentam outras características que as tornam interessantes. Uma RN possui capacidade de se adaptar. Quando uma rede é construída para uma determinada aplicação, ela poderá ser utilizada em aplicações de tempo real, sem necessidade de que sua arquitetura seja alterada a cada atualização, é necessário apenas que ela seja re-treinada com base nos novos dados que forem surgindo. Possui capacidade de generalização, ou seja, mesmo com dados incompletos ou imprecisos, as redes podem preencher as lacunas sem sofrer degradação. As redes possuem capacidade de tolerar falhas, pois as unidades da rede operam em paralelo, sendo assim, se um nó falhar isso não a torna inoperante, podendo até mesmo não causar grandes problemas em seu funcionamento.

Um tipo de rede neural muito utilizada é a rede de múltiplas camadas, ou Perceptron Multicamadas (*Multilayer Perceptron* - MLP). Uma rede MLP apresenta os neurônios dispostos em camadas, sendo uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída (Haykin, 2001). As camadas de neurônios intermediários utilizam funções não-lineares ponderadas pelos valores dos pesos das conexões associadas a elas. Seu treinamento é, normalmente, realizado de forma

supervisionada pelo algoritmo de retro-propagação de erro ou *backpropagation* (Hecht-Nielsen, 1989). A estrutura de uma Rede Neural Artificial MLP pode ser observada na Figura 3. Nesta rede, o conjunto de treinamento apresenta cinco atributos, Idade, DAS 28, Sexo, anti-CCP e PCR. A classe é a Resposta ao Tratamento com DMARDs, que pode assumir os valores respondeu e não respondeu.

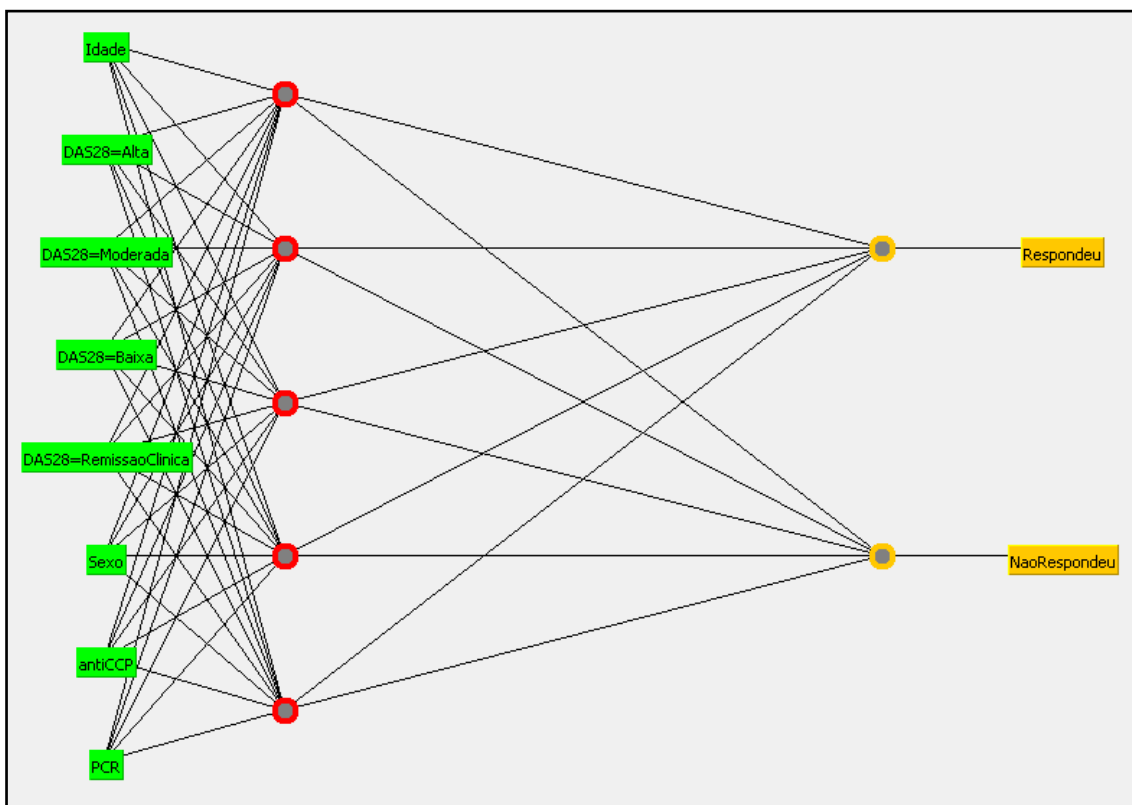


Figura 3: Estrutura de uma Rede Neural Artificial. A camada de entrada representa os atributos do conjunto de treinamento. A camada intermediária realiza uma transformação não-linear dos dados de entrada e a camada de saída representa os valores do atributo classe, neste exemplo, Respondeu e Não Respondeu ao tratamento com DMARDs.

O treinamento usando o algoritmo *backpropagation* consiste em duas etapas: *forward* e *backward*. Na fase *forward*, um exemplo de treinamento é apresentado à rede e esta produz uma saída. Na fase *backward*, que percorre a rede da camada de saída em direção à camada de entrada, os pesos das conexões são corrigidos a partir do erro calculado entre a saída desejada e a saída produzida pela rede (Lorena e Carvalho, 2003). O algoritmo *backpropagation* é apresentado na Figura 4.

```

1: Iniciar os pesos da rede com pequenos valores aleatórios
2: repita
3: erro_total = 0
4: para cada padrão de treinamento faça
5:   Apresentar padrão à rede
6:   Calcular saída_produzida pela rede
7:   erro = saída_produzida - saída_desejada
8:   erro_total = erro_total + erro
9:   para todo camada  $i$  da rede,  $i = n, n - 1, \dots, 1$  faça
10:    para cada nodo  $n_{ij}$  da  $i$ -ésima camada faça
11:     Ajustar pesos associados ao nodo
12:    fim para
13:   fim para
14: fim para
15: até erro_total < limiar

```

Figura 4: Pseudocódigo do algoritmo de treinamento *backpropagation*.

3.3.3 Algoritmos Genéticos

Os Algoritmos Evolucionários (AE) são uma classe de algoritmos de busca estocástica baseados nas idéias e princípios da seleção natural. Os AEs incluem Estratégias Evolucionárias, Programação Evolucionária e os Algoritmos Genéticos (AGs). Os AEs podem ser utilizados para várias tarefas, uma delas é a seleção de atributos (Yao, 1999).

Os AGs são métodos adaptativos baseados nos processos genéticos de organismos biológicos e na teoria da evolução natural que podem ser utilizados para resolver problemas de busca e otimização (Mitchell, M., 1996).

No AG padrão, um conjunto de indivíduos (cromossomos), representando soluções do problema, está sujeito às leis de evolução. A solução \mathbf{x}_i representada pelo cromossomo $i = 1, \dots, n$ é avaliada através de uma função de avaliação, ou *fitness*, $f(\mathbf{x}_i)$. Os cromossomos estão sujeitos, através das diferentes gerações, a operadores genéticos de transformação. Os mais utilizados são o *crossover* e a mutação. No

primeiro, dois indivíduos da população corrente, escolhidos através de um critério de seleção, têm alguns de seus genes trocados. O número de indivíduos escolhidos para *crossover* em uma geração é definido através de uma taxa p_c , chamada de taxa de *crossover*. No segundo operador, indivíduos têm alguns de seus genes alterados através de uma regra pré-definida. O número de genes alterados por mutação é definido através de uma taxa p_m , chamada de taxa de mutação (Linden, 2006). O AG padrão é apresentado na Figura 5.

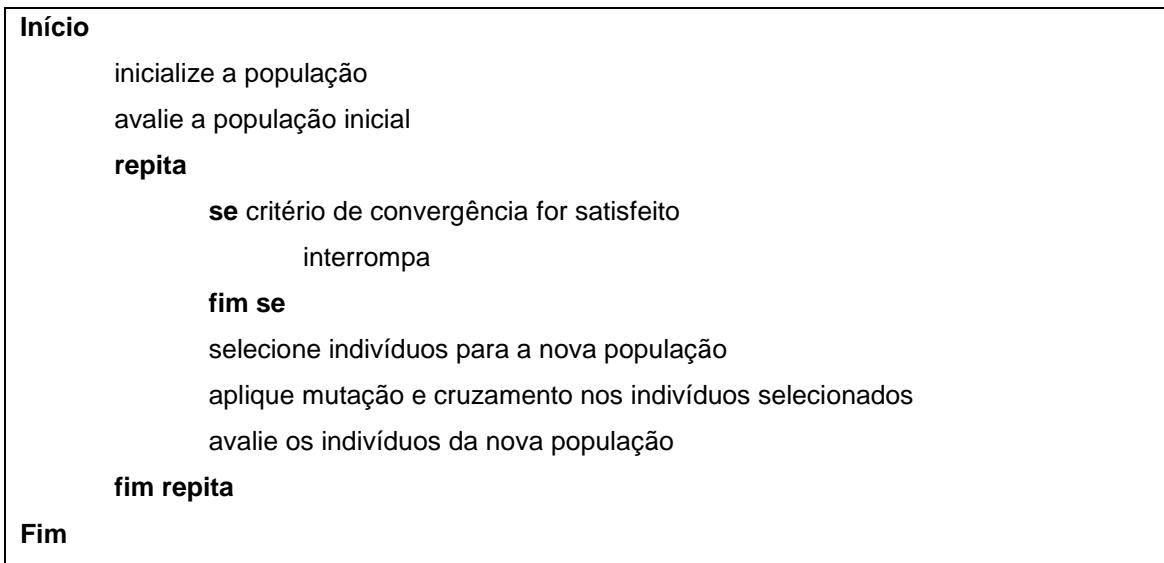


Figura 5: Pseudocódigo do Algoritmo Genético Padrão.

3.3.4 Seleção de Atributos

A seleção de atributos é importante para eliminar informações redundantes ou irrelevantes do espaço de busca, as quais podem gerar ruídos que degradam o desempenho da classificação dos dados (Fukunaga, 1990).

O problema de seleção de atributos pode ser visto como um problema de busca, no qual um subconjunto ótimo de atributos deve ser encontrado. Neste caso, o objetivo dos AGs é criar indivíduos que representem combinações dos atributos de entrada. Nas arquiteturas, que serão utilizadas aqui, um indutor é construído, treinado e avaliado para cada indivíduo do AG. Neste caso os indutores utilizados serão árvores de decisão ou redes neurais artificiais. O desempenho do indutor é retornado

para o AG como o valor da função de avaliação ou *fitness*. A arquitetura utilizada para a seleção dos atributos, selecionada para o sistema, é ilustrada na Figura 6. Esta ilustração exemplifica a interação entre o AG e uma árvore de decisão. Em princípio, tem-se a população inicial que representa soluções aleatórias ao problema em questão. Cada indivíduo da população é composto por um cromossomo, que é representado por uma sequência de *bits*. Cada *bit* representa um gene do cromossomo. Na fase de avaliação, uma árvore de decisão é construída para avaliar cada uma das soluções candidatas, presentes na população. Assim, obtêm-se as respectivas taxas de acerto, que funcionam como função de *fitness*. A função de *fitness* é utilizada na fase de seleção, como um critério para selecionar os indivíduos que irão se reproduzir. A fase de reprodução utiliza os operadores de *crossover* e mutação para gerar novos indivíduos, que irão compor a população atual. Este ciclo se repete até o critério de parada ser satisfeito e a solução final ser encontrada.

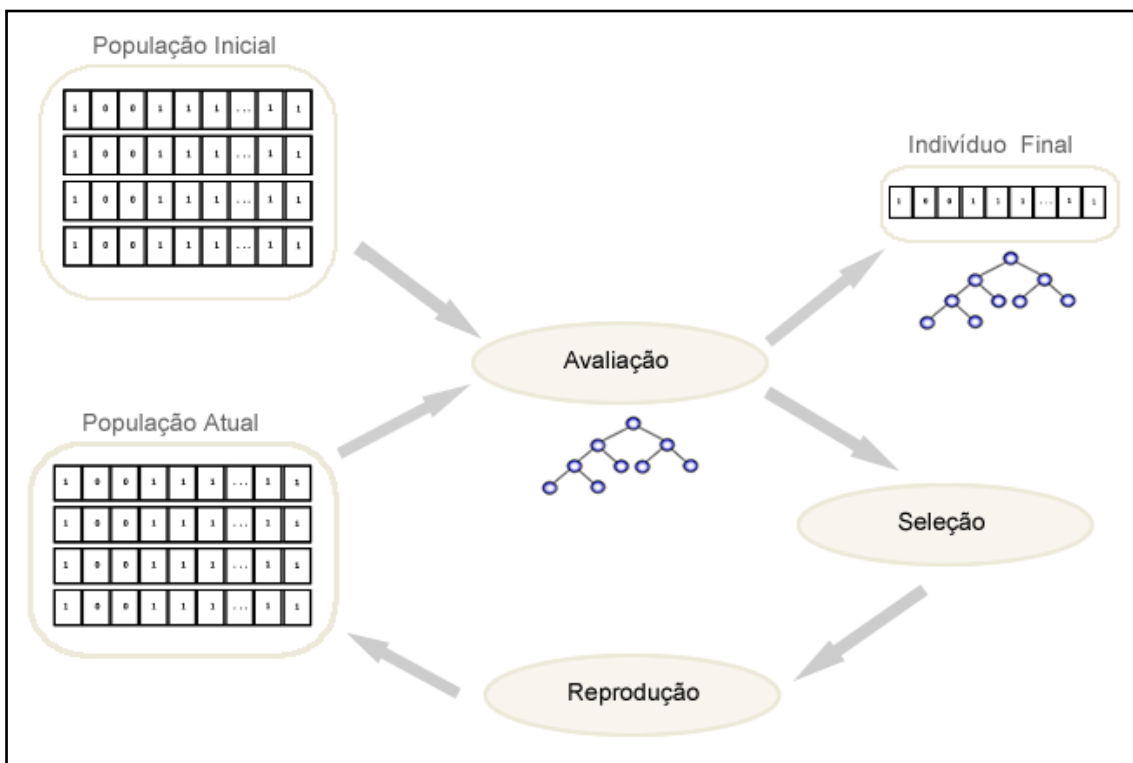


Figura 6: Ilustração da interação entre o algoritmo genético e a árvore de decisão, utilizados na seleção de atributos. A população inicial representa soluções aleatórias ao problema. Na fase de avaliação, uma árvore de decisão é treinada para avaliar cada uma das soluções candidatas. O resultado da avaliação é utilizado na reprodução dos indivíduos que irão compor a população atual. Este ciclo se repete até um critério de parada ser satisfeito. Adaptado de (Favretto et al., 2006).

Existem algumas discussões que argumentam que a representação binária não é a melhor representação possível (Yao, 1999). As vantagens da codificação binária são sua simplicidade de implementação e a vasta quantidade de problemas em que pode ser aplicada. É possível aplicar os operadores clássicos de *crossover* e mutação (Yao, 1999). Porém, quando essa abordagem é aplicada no problema da seleção de genes a representação dos indivíduos fica muito longa. Por exemplo, se um experimento de *microarray* analisa 1000 genes, o tamanho do cromossomo no AG será 1000.

Um cromossomo com codificação binária pode ser observado na Figura 7. Cada posição do cromossomo representa um gene, que por sua vez, representa um atributo do conjunto de treinamento. Quando o valor do gene é igual a 1, o atributo representado pelo respectivo gene estará presente na solução candidata. Caso o valor do gene seja 0, o atributo não fará parte da solução candidata.

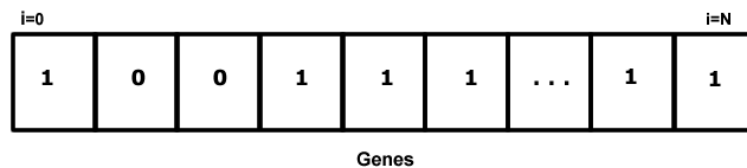


Figura 7: Exemplo de um cromossomo com codificação binária. Cada posição do cromossomo indica um gene que representa um atributo do conjunto de treinamento. Quando o valor do gene é 1, o atributo representado pelo respectivo gene será utilizado na solução candidata. Quando o gene assume o valor 0, o respectivo atributo não será utilizado na solução candidata.

Outra forma de representar os indivíduos do AG é a representação através de números reais. Neste caso, limita-se o tamanho máximo do cromossomo (*setsize*), e este é representado pelos índices de cada gene ou por zero quando nenhum gene for selecionado para aquela posição (Ding et al., 2004). Neste caso, cada posição do cromossomo representa um atributo do conjunto de treinamento, sendo que o valor de cada posição indica qual será o atributo do conjunto de treinamento que será utilizado na solução candidata. Por exemplo, para um *setsize* = 5, um possível cromossomo seria (100, 50, 10, 0, 15) que pode ser observado na Figura 8. Este cromossomo,

representa soluções que irão selecionar no máximo, 5 atributos do conjunto de treinamento e os atributos selecionados seriam os seguintes genes: gene 100, gene 50, gene 10 e o gene 15. A desvantagem deste método é que o tamanho máximo do cromossomo deve ser pré-estabelecido. Porém, elimina-se o problema do tamanho do indivíduo ser muito longo, encontrado na representação binária.

Genes	i=0	i=1	i=2	i=3	i=4
	100	50	10	0	15
	35	12	71	29	43
	1	5	20	84	64

Figura 8: Exemplos de cromossomos com codificação real, utilizando *setsize* = 5. Cada posição do cromossomo representa um atributo do conjunto de treinamento. O valor de cada posição indica o atributo que será utilizado na solução candidata.

Diversas estratégias têm sido propostas para a seleção de genes em dados de expressão gênica (Ding et al., 2004; Ho et al., 2005; Huang et al., 2003). Dois aspectos principais do processo de otimização têm sido tratados: a escolha da função de avaliação a ser utilizada no processo de otimização e a escolha do método de codificação a ser empregado. No presente projeto, será empregada a codificação binária dos cromossomos e a função de avaliação será a taxa de acerto dos indutores gerados para cada indivíduo da população.

3.3.5 Weka

Os algoritmos de aprendizado de máquinas utilizados na implementação da ferramenta de seleção de atributos e classificação de pacientes portadores de AR estão disponíveis no *software* Weka (Witten e Frank, 2005).

O Weka – *Data mining with Open Source Machine Learning Software in Java* - é um *software open source* distribuído sob a licença GPL (*General Public Licence*) que disponibiliza um conjunto de algoritmos de aprendizado de máquinas para mineração de dados. Estes algoritmos disponibilizados podem ser aplicados a um conjunto de dados diretamente ou podem ser incluídos em códigos Java. O Weka oferece diversas ferramentas para pré-processamento, classificação, regressão, agrupamento, regras de associação e visualização dos dados.

O *software* Weka está, atualmente, em sua versão 3.4 e pode ser obtido através do endereço <http://www.cs.waikato.ac.nz/ml/weka/>.

3.4 Modelagem do Sistema

A modelagem do sistema foi feita utilizando a linguagem UML (*Unified Modeling Language*). A UML é uma linguagem apresentada de forma gráfica para visualização, especificação e documentação de sistemas de *software*. Assim, proporciona uma forma-padrão para a preparação de planos de arquitetura de projetos incluindo, também, aspectos conceituais (Booch et al., 2006).

No presente trabalho, a UML está relacionada aos procedimentos que devem ser executados pelos atores do sistema, visto que o sistema utiliza a infra-estrutura já estabelecida do CMS Drupal. A modelagem exemplifica a criação de conteúdos, categorias, cadastro de usuários, cadastro de pacientes, utilização das ferramentas de seleção de atributos e predição da resposta à DMARDs e as outras opções de visualização dos conteúdos que existem no sistema.

3.4.1 Escopo do Sistema

O sistema ARIA é um sistema *Web* restrito apenas a usuários registrados que oferece informações sobre pacientes portadores de AR e ferramentas baseadas em técnicas de IA, a fim de gerar modelos preditivos da resposta dos pacientes ao tratamento utilizando DMARDs.

As funcionalidades oferecidas pelo sistema, para cumprir os requisitos propostos, são descritas nas sessões subseqüentes.

Cadastrar usuários

Novos usuários são cadastrados no sistema pelo administrador. Para isso o administrador acessa a página de Usuários e, então, cria um novo usuário. Os campos **nome do usuário**, **endereço de e-mail** e **senha** são requeridos. Além disso, o administrador designa o papel do novo usuário no sistema. Os possíveis papéis são **médicos** e **colaboradores**. Os usuários possuem diferentes tipos de permissões de acesso. A senha designada é enviada para o e-mail do novo usuário cadastrado. Os usuários cadastrados podem utilizar todas as funcionalidades do sistema.

Criar tipo de conteúdo

Apenas o administrador do sistema possui privilégio para criar novos tipos de conteúdo. Os tipos de conteúdos são as informações que devem ser inseridas no sistema através de campos específicos de entrada de dados. O módulo CCK permite criar novos tipos de conteúdos, que interagem diretamente com o banco de dados MySQL.

Para criar um novo tipo de conteúdo o usuário administrador deve acessar a área administrativa do sistema e selecionar a opção **criar novo tipo de conteúdo**. Em

seguida, o administrador deve informar o nome do conteúdo, a descrição do conteúdo, as informações sobre ajuda para o preenchimento e adicionar os campos necessários para a composição do novo conteúdo.

Cadastrar categorias

O sistema é composto por categorias e subcategorias, que são utilizadas para classificar e organizar os conteúdos. As categorias podem ser específicas para um único tipo de conteúdo ou comum para vários. A categorização do conteúdo é fundamental, pois permite que buscas sejam feitas de forma eficiente, através de buscas por palavras chaves ou através da utilização do módulo *Views*, que permite a realização de buscas relacionando os termos das categorias.

Quando os dados estão sendo cadastrados, alguns dos campos requeridos são também categorias. Isso é possível devido ao módulo *Taxonomy Field*. Este módulo facilita o trabalho de categorização dos dados, pois à medida que os dados são cadastrados, eles são também categorizados automaticamente.

Somente o administrador tem permissão para criar novas categorias e novos termos nestas categorias.

Cadastrar Paciente

Os pacientes são cadastrados no sistema utilizando o tipo de conteúdo *Cadastro de Paciente*. Para isso é necessário acessar o sistema com um usuário cadastrado, em seguida acessar a opção *Cadastrar Paciente* e preencher o formulário de inserção de dados. O formulário de cadastro está dividido em: **Dados Pessoais**, **Parâmetros clínicos**, **Parâmetros Laboratoriais**, **Informações de *Microarray*** e

Classe Resposta ao Tratamento com DMARDs. Após preencher as informações necessárias, o usuário deve enviá-las ao sistema.

Enviar Submissão de Seleção de Atributos

A execução de uma seleção de atributos se dá através do envio de uma solicitação de submissão de seleção de atributos ao sistema. Para isso é necessário acessar o sistema com um usuário cadastrado, em seguida acessar a opção *Seleção de Atributos* no menu *Ferramenta*, preencher o formulário selecionando o tipo de dado a ser utilizado como conjunto de treinamento (dados clínicos ou dados de expressão gênica), selecionar os parâmetros da busca genética, selecionar o algoritmo de classificação a ser utilizado no treinamento (árvore de decisão ou rede neural artificial) e selecionar os parâmetros do algoritmo de classificação. Após preencher as informações necessárias, o usuário deve enviá-las ao sistema e aguardar o processamento dos resultados, para então visualizá-los.

Predição da Resposta à DMARDs

A execução da predição da resposta à DMARDs se dá através do envio de uma solicitação de submissão de predição da resposta à DMARDs ao sistema. É necessário acessar o sistema com um usuário cadastrado, em seguida acessar a opção *Predição da Resposta à DMARDs* no menu *Ferramenta*, preencher o formulário de submissão selecionando um paciente na lista de pacientes e selecionar um modelo na lista de modelos. Após o preenchimento das informações necessárias, o usuário deve enviá-las ao sistema e aguardar o processamento do resultado, para então visualizá-lo.

Visualizar Conteúdo

Para visualizar os conteúdos do sistema é necessário possuir um usuário cadastrado. Existem várias formas de visualização do conteúdo, através de listas de conteúdos, de filtros de conteúdos, através das categorias e através das buscas.

Visualizar Conteúdo Público

Os visitantes são usuários que não são cadastrados no sistema e por isso possuem acesso limitado. O único conteúdo que os visitantes podem acessar são páginas. As páginas são conteúdos estáticos informativos, como a apresentação do sistema, página sobre o grupo e a página de contato.

3.4.2 Diagramas UML

Os diagramas são os meios utilizados para a visualização dos blocos de construção da UML, proveem uma parcial representação do sistema e ajudam a compreender a arquitetura do sistema em desenvolvimento. O diagrama é uma representação gráfica de um conjunto de elementos que permitem a visualização do sistema a partir de diferentes perspectivas. Na UML existem vários tipos de diagramas utilizados na modelagem que especificam modelos a partir dos quais são construídos sistemas executáveis (Booch, Jacobson e Rumbaugh 2006).

Diagrama de Casos de Uso

O Diagrama de Caso de Uso descreve a funcionalidade proposta para o novo sistema. Um Caso de Uso (*Use-case*) descreve a sequência de eventos de um ator que usa um sistema para completar um processo, representando uma unidade da interação entre um usuário (humano ou máquina) e o sistema.

A Visão Global dos Atores e *Use-cases* é um diagrama que representa graficamente os atores do sistema e suas interações com os casos de uso. O ator é qualquer pessoa ou sistema externo que tenha interação com o sistema que está em desenvolvimento e, o caso de uso é um conjunto de funcionalidades de um sistema, representado por fluxos de eventos iniciados por um ator que apresenta um resultado de valor a um ator (Cardoso et al. 2003).

Na Figura 9 é apresentado o Diagrama de Caso de Uso: Visão Global dos Atores e *Use-cases*. Neste diagrama são apresentados os Atores e os Casos de Uso do sistema. O sistema possui quatro atores, **Administrador**, **Médico**, **Colaborador** e **Visitante**. Um quinto ator, denominado **Usuário**, é uma generalização dos atores Administrador, Médico e Colaborador, o que no diagrama é representado por uma seta desses atores para o ator Usuário. Podem ser observados, no diagrama, os casos de uso do sistema, que são: Cadastrar Usuário, Criar Tipo de Conteúdo, Cadastrar Categoria, Cadastrar Paciente, Enviar Submissão de Seleção de Atributos, Enviar Submissão de Predição da Resposta à DMARDs, Alterar Dados, Visualizar Conteúdo e Visualizar Conteúdo Público. Além disso, o diagrama de caso de uso apresenta as interações dos vários atores do sistema com os casos de uso, mostrando quais atores têm permissão para executar quais casos de uso.

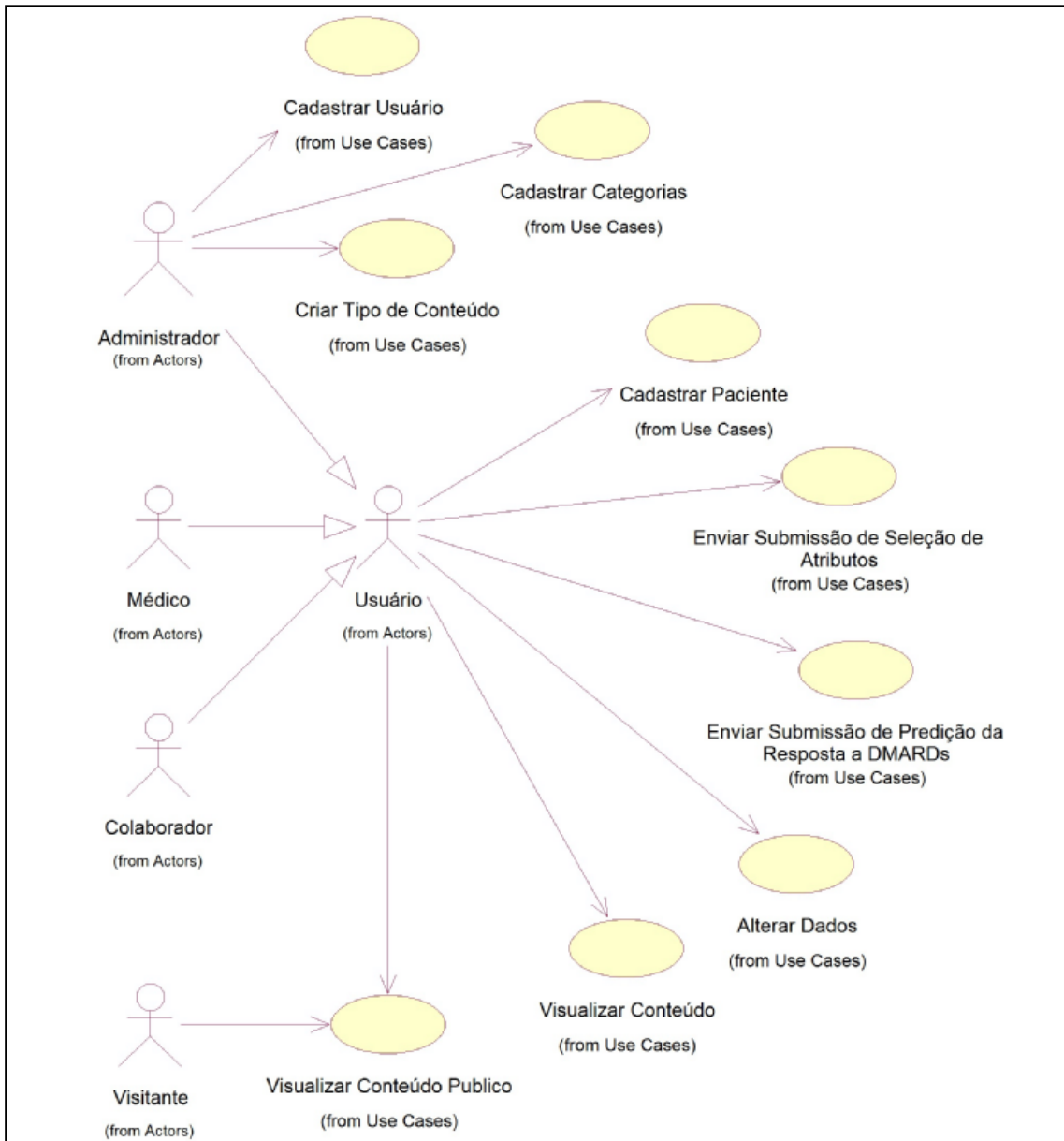


Figura 9: Diagrama de casos de uso: Visão Global dos Atores e Use-Cases. Os atores Administrador, Médico e Colaborador são Usuários do sistema e interagem com as use-cases que cada papel tem permissão de realizar. O ator visitante não é um Usuário cadastrado no sistema, podendo visualizar somente os conteúdos públicos.

Diagramas de Colaboração

A colaboração é representada por meio de diagramas que demonstram o relacionamento das classes no que diz respeito ao sentido das mensagens e ao número de mensagens, dando ênfase à organização dos objetos que participam de uma interação (Cardoso, 2003). O diagrama de colaboração contém seu Fluxo Básico, ou seja, o fluxo padrão de funcionamento de um determinado caso de uso e os Fluxos

Alternativos, que correspondem a quaisquer outros tipos de eventos fora do fluxo padrão, levando-se em consideração a ordem em que se encontram os atores e os estereótipos. Este tipo de diagrama permite observar o fluxo das informações e onde possíveis problemas podem acontecer.

O Diagrama de Colaboração **Cadastrar Usuário** é apresentado na Figura 10. Neste diagrama é possível observar todos os passos necessários para cadastrar um novo usuário no sistema, bem como as informações e mensagens necessárias para realizar essa tarefa. Observa-se no diagrama que somente o ator **Administrador** tem permissão para executar esse caso de uso.

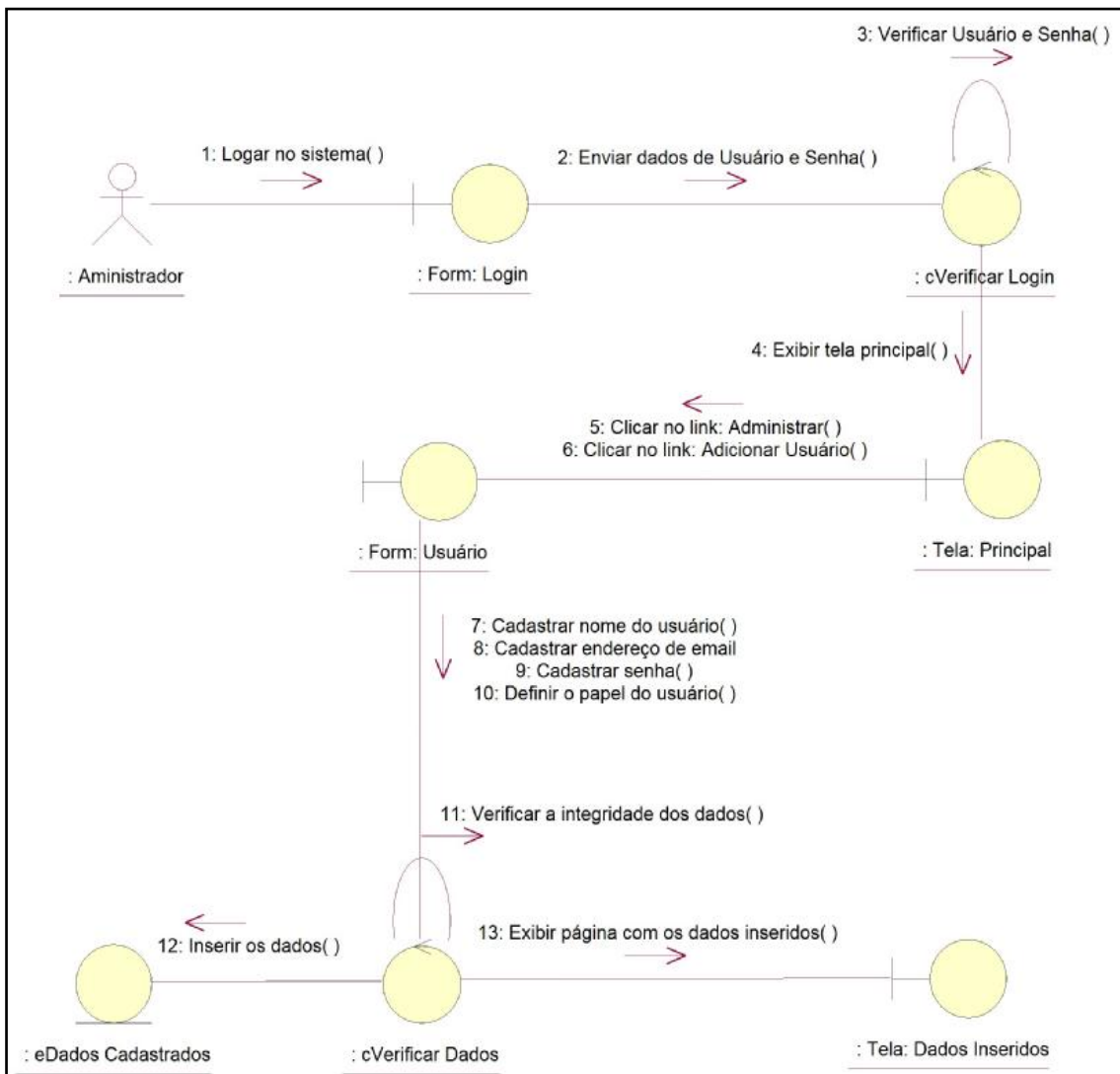


Figura 10: Diagrama de Colaboração do caso de uso Cadastrar Usuário – Fluxo Básico. Representa todos os eventos que o Ator Administrador realiza para criar um novo usuário, descrevendo as entradas de dados necessárias e as mensagens para cada evento.

Na Figura 11 é apresentado o Diagrama de Colaboração **Cadastrar Categoria**. Observa-se no diagrama que somente o ator **Administrador** tem permissão para executar este caso de uso, que apresenta os passos necessários para criar uma nova categoria no sistema, assim como, as mensagens e o fluxo das informações necessárias para executá-lo.

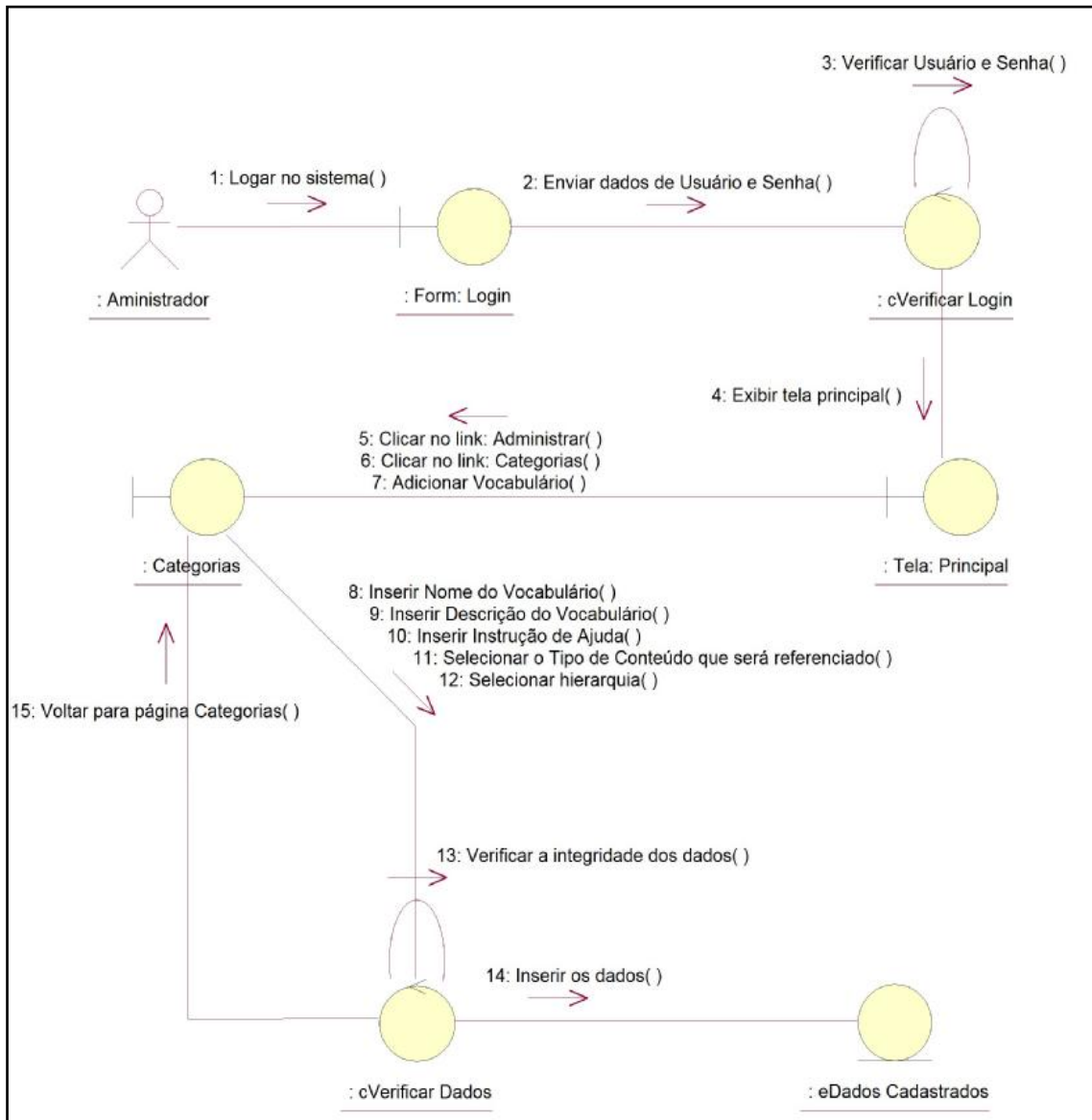


Figura 11: Diagrama de colaboração do caso de uso **Cadastrar Categorias – Fluxo Básico**. Representa os eventos que o ator **Administrador** realiza para criar uma nova categoria, descrevendo as entradas de dados necessárias e as mensagens para cada evento.

O Diagrama de Colaboração **Criar Tipo de Conteúdo** é ilustrado na Figura 12. Neste diagrama pode-se observar o fluxo das informações e mensagens necessárias para a criação de um novo tipo de conteúdo no sistema. Assim como nos casos de usos anteriores, somente o ator **Administrador** tem privilégios para criar um novo tipo de conteúdo.

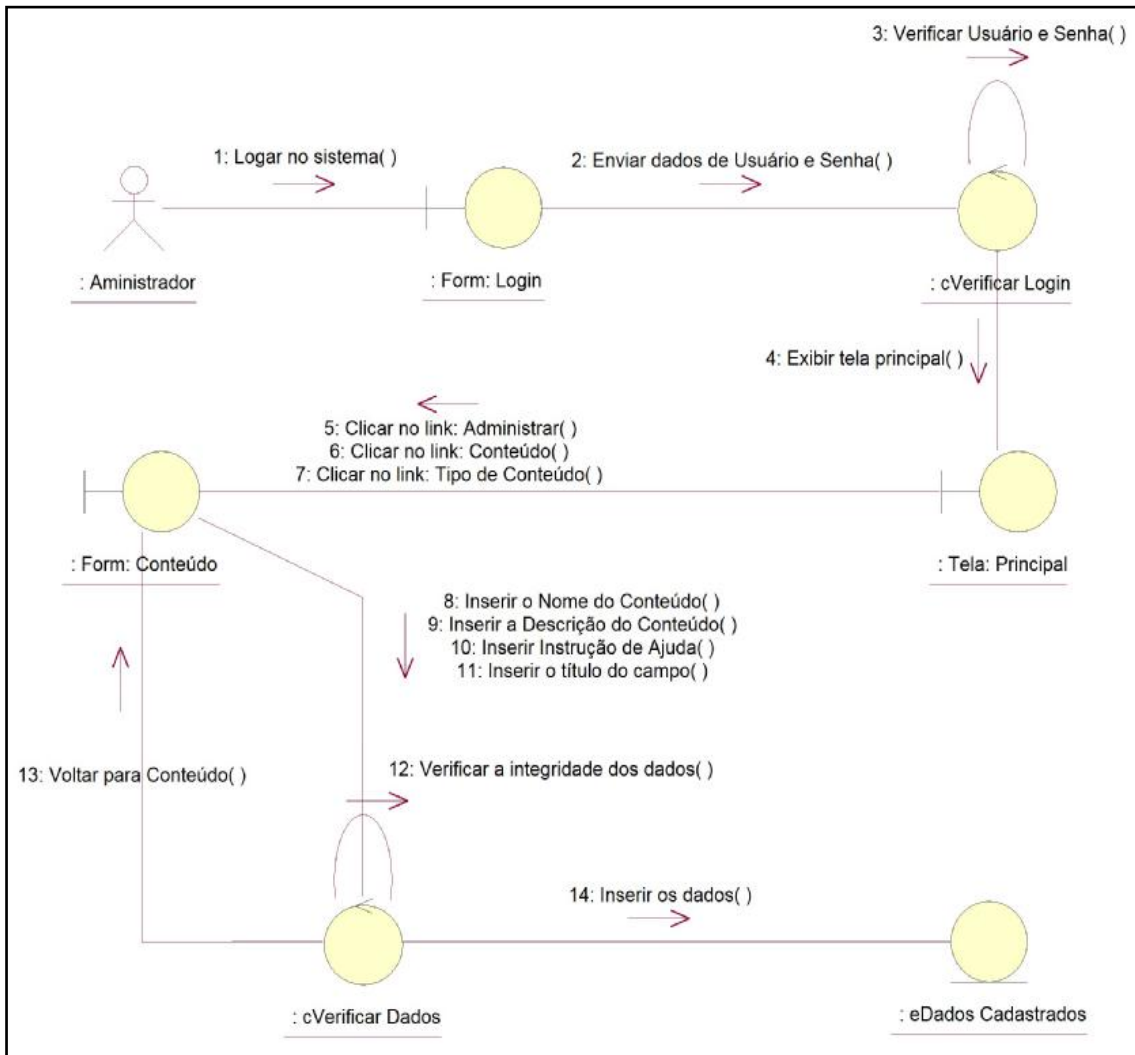


Figura 12: Diagrama de colaboração do caso de uso **Criar Tipo de Conteúdo** – Fluxo Básico. Representa todos os eventos que o ator **Administrador** deve realizar para criar um novo tipo de conteúdo no sistema, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento.

O Diagrama de Colaboração **Cadastrar Paciente** é apresentado na Figura 13. É possível observar no diagrama que o ator **Usuário**, possui privilégios para realizar o cadastro de pacientes no sistema. Isso significa que qualquer um dos atores **Administrador, Médico ou Colaborador**, tem permissão para executar esse caso de uso. Além disso, são apresentados os passos necessários para cadastrar um novo paciente no sistema, assim como as informações e mensagens necessárias para a execução deste caso de uso.

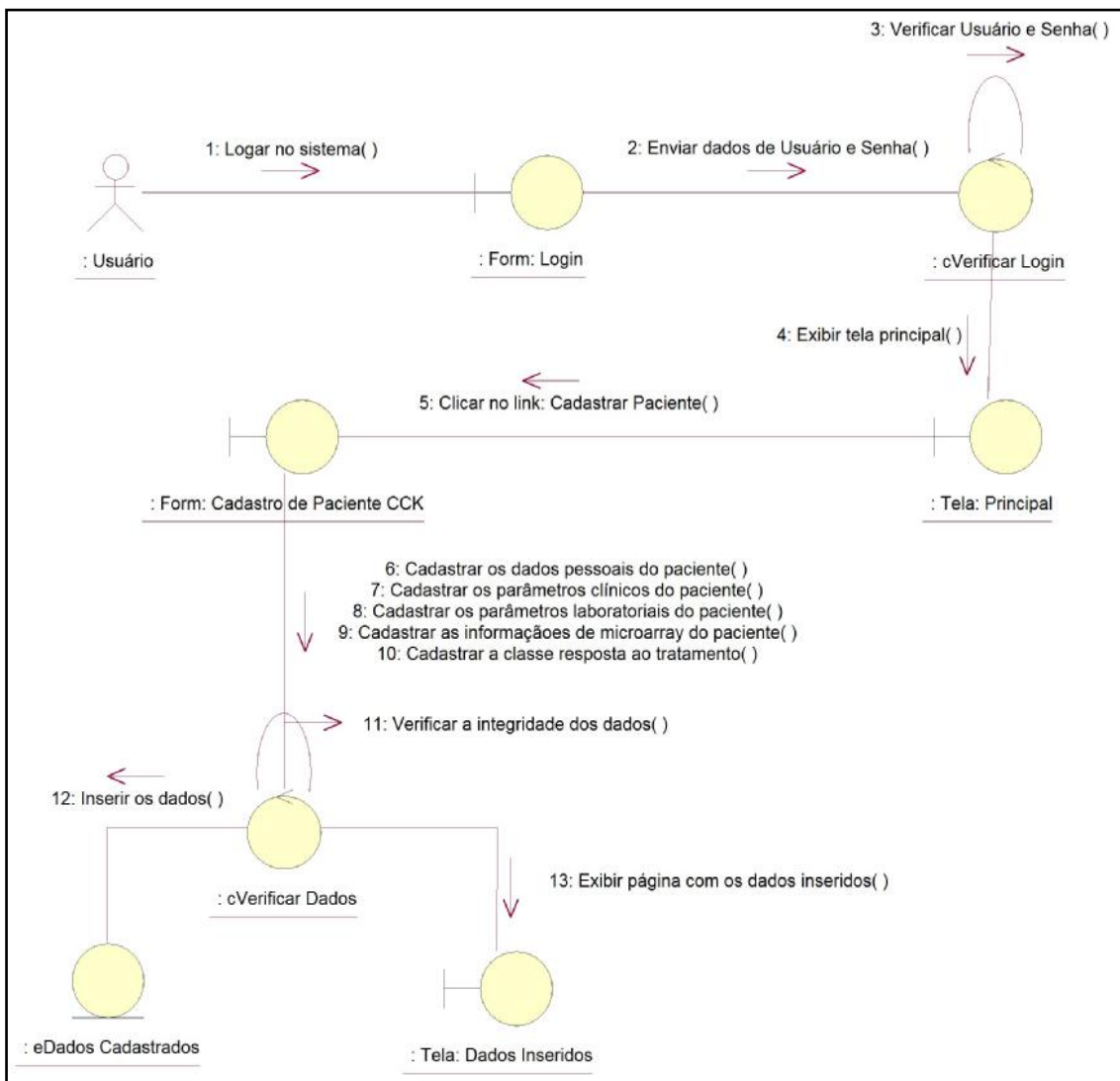


Figura 13: Diagrama de colaboração do caso de uso Cadastrar Paciente – Fluxo Básico. Representa todos os eventos que o ator **Usuário**, ou seja, **Administrador, Médico ou Colaborador**, deve realizar para cadastrar um novo paciente no sistema, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento.

Na Figura 14 é apresentado o Diagrama de Colaboração **Enviar Submissão de Seleção de Atributos**. Neste diagrama observa-se, novamente, que o ator Usuário possui permissão de executar a submissão de seleção de atributos. Dessa forma, qualquer um dos atores **Administrador**, **Médico** ou **Colaborador**, pode executar esse caso de uso. Observam-se também, as informações e mensagens necessárias para enviar uma submissão de seleção de atributos ao sistema.

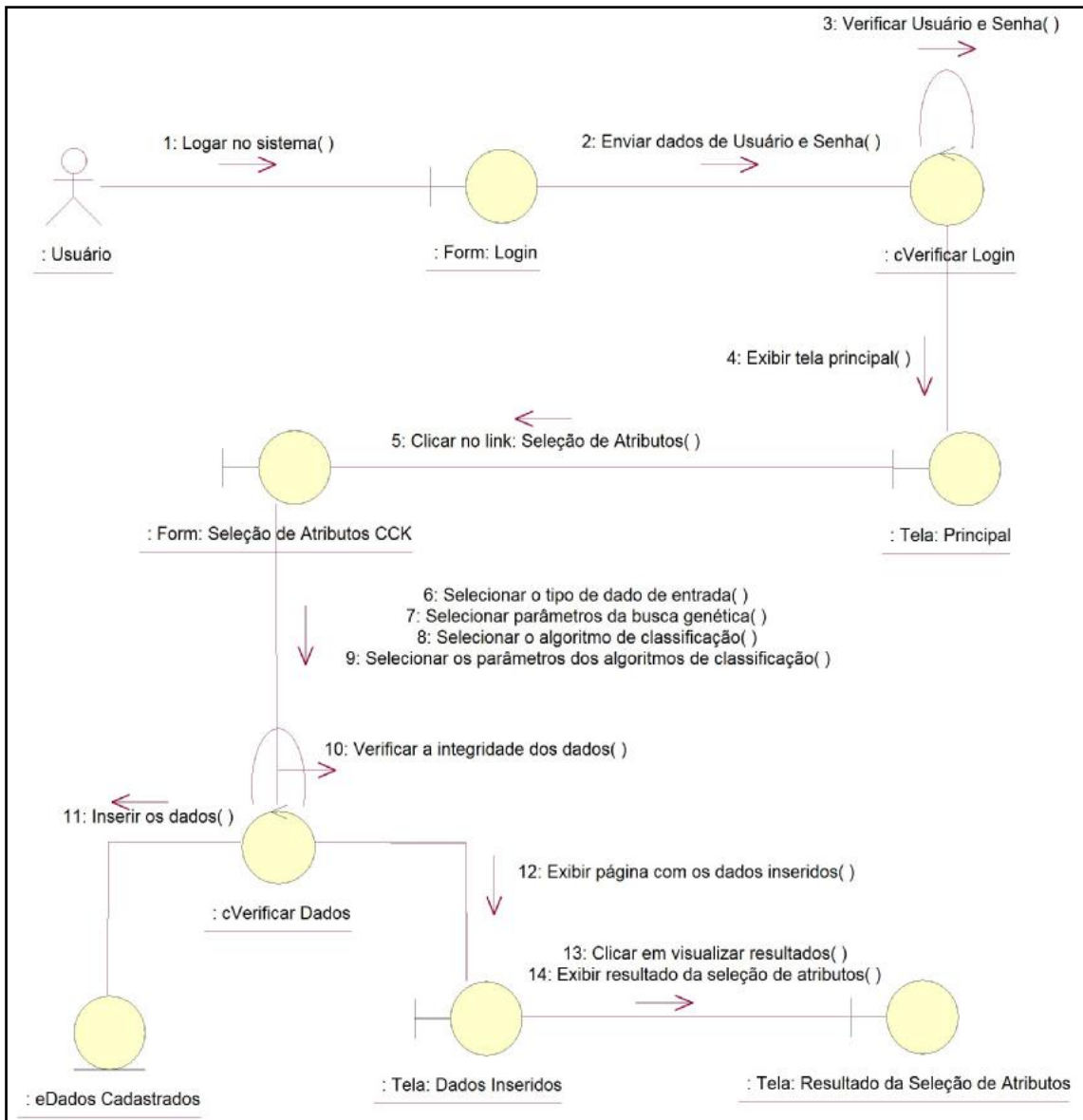


Figura 14: Diagrama de colaboração do caso de uso Enviar Submissão de Seleção de Atributos – Fluxo Básico. Representa todos os eventos que o ator Usuário (Administrador, Médico ou Colaborador) deve realizar para submeter uma seleção de atributos ao sistema, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento.

O Diagrama de Colaboração **Enviar Predição da Resposta à DMARDs** é apresentado na Figura 15. Neste diagrama, observam-se os passos necessários para enviar uma predição da resposta à DMARDs ao sistema, e o fluxo das informações e mensagens necessárias para a execução deste caso de uso. Os atores que possuem privilégio para submeter uma predição da resposta ao sistema são: **Administrador**, **Médico** ou **Colaborador**, representados no diagrama pelo ator **Usuário**.

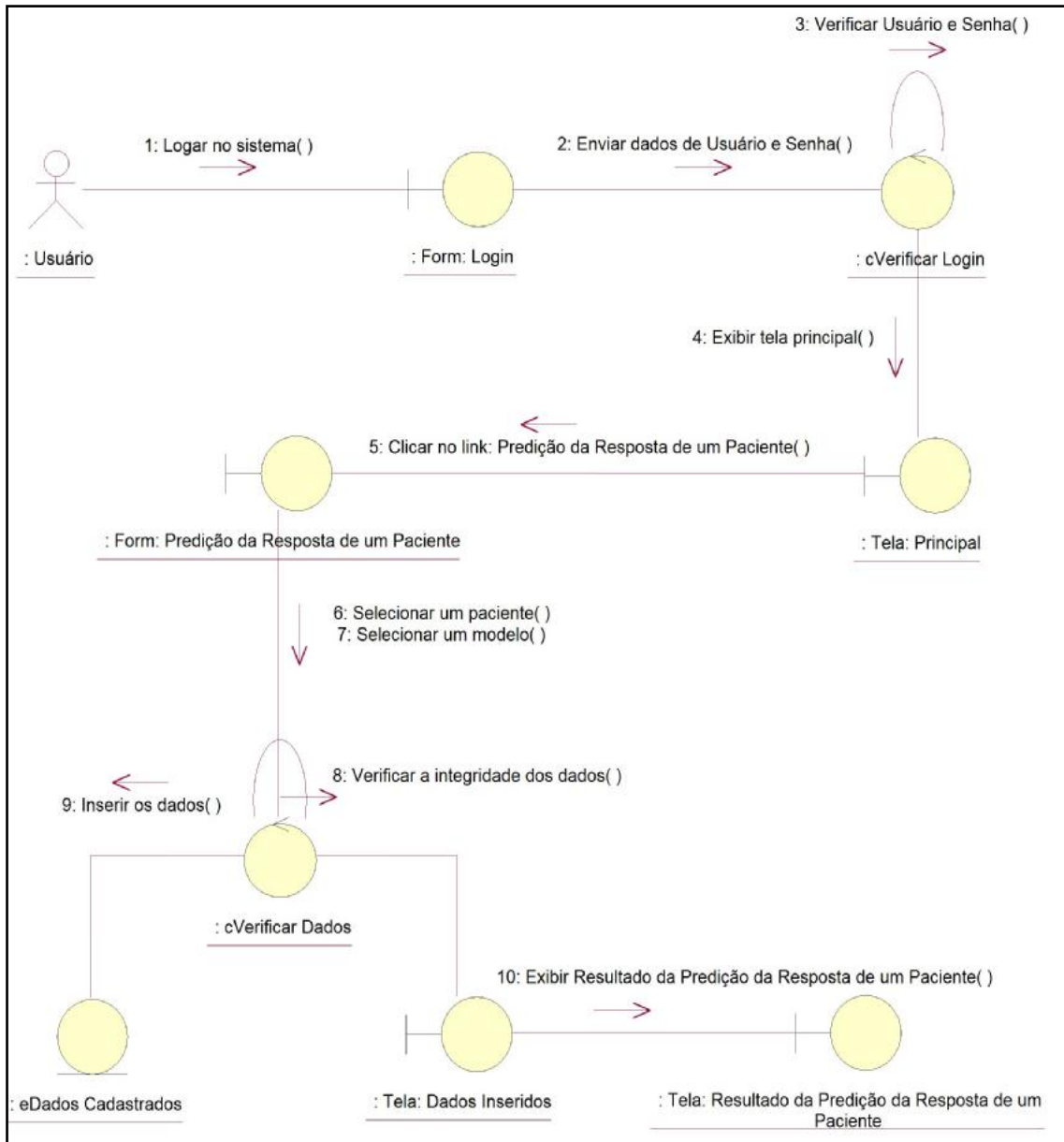


Figura 15: Diagrama de colaboração do caso de uso Enviar Predição da Resposta à DMARDs – Fluxo Básico. Representa todos os eventos que o ator Usuário (Administrador, Médico ou Colaborador) deve realizar para submeter uma predição da resposta à DMARDs ao sistema, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento.

O Diagrama de Colaboração **Alterar Dados** é apresentado na Figura 16. Neste diagrama observam-se os passos necessários para alterar dados previamente cadastrados ao sistema, e o fluxo das informações e mensagens necessárias para a execução deste caso de uso. O ator Usuário possui privilégios para alterar os dados, sendo assim, os atores **Administrador**, **Médico** ou **Colaborador**, têm permissão para executar esse caso de uso.

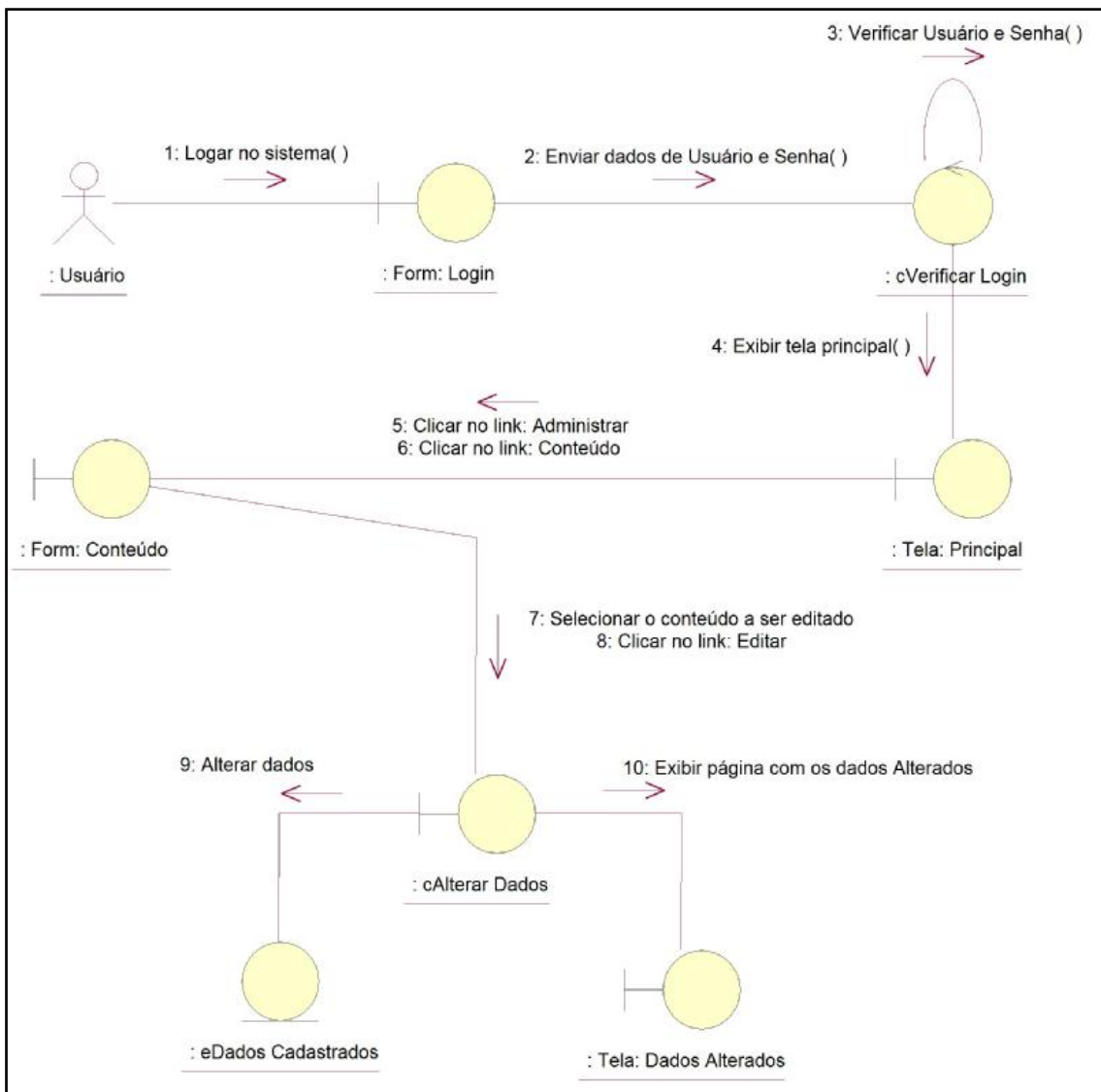


Figura 16: Diagrama de colaboração do caso de uso Alterar Dados – Fluxo Básico. Representa todos os eventos que o ator Usuário (Administrador, Médico ou Colaborador, deve realizar para alterar dados previamente cadastrados no sistema, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento.

Na Figura 17 é apresentado o Diagrama de Colaboração **Visualizar Conteúdo**. Este caso de uso pode ser executado pelo ator Usuário, que representa os atores **Administrador, Médico e Colaborador**. São apresentados os passos necessários para visualizar um conteúdo do sistema, e o fluxo das informações e mensagens trocadas para a execução deste caso de uso.

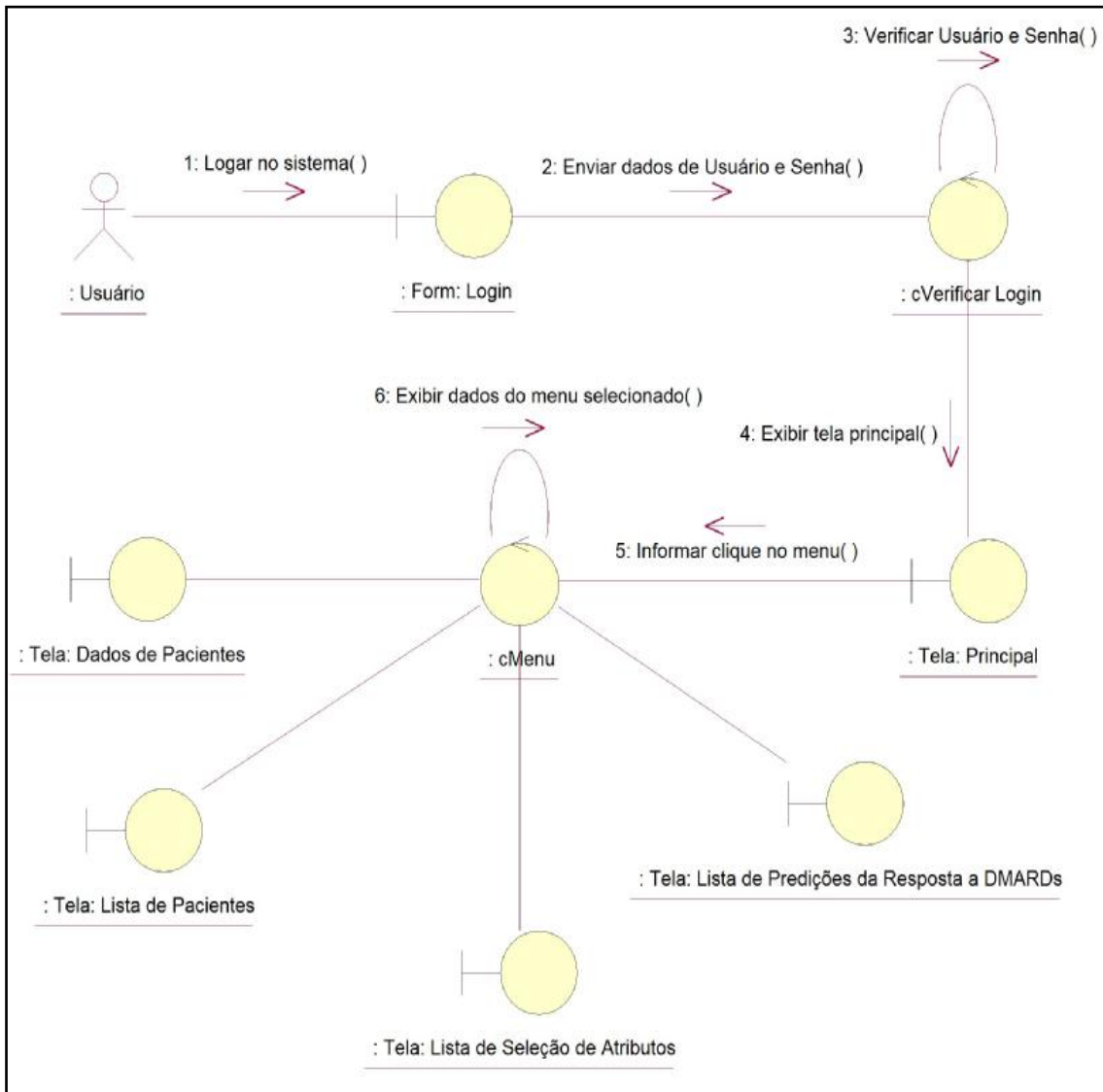


Figura 17: Diagrama de colaboração do caso de uso Visualizar Conteúdo – Fluxo Básico. Representa todos os eventos que o ator Usuário (Administrador, Médico ou Colaborador) deve realizar para visualizar um conteúdo do sistema, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento.

O Diagrama de Colaboração **Visualizar Conteúdo Público** é apresentado na Figura 18. É possível observar no diagrama que o ator **Visitante**, pode visualizar os conteúdos públicos. Visualizar conteúdo público é o único caso de uso que o visitante participa, pois é a única parte do sistema que não requer *Login*.

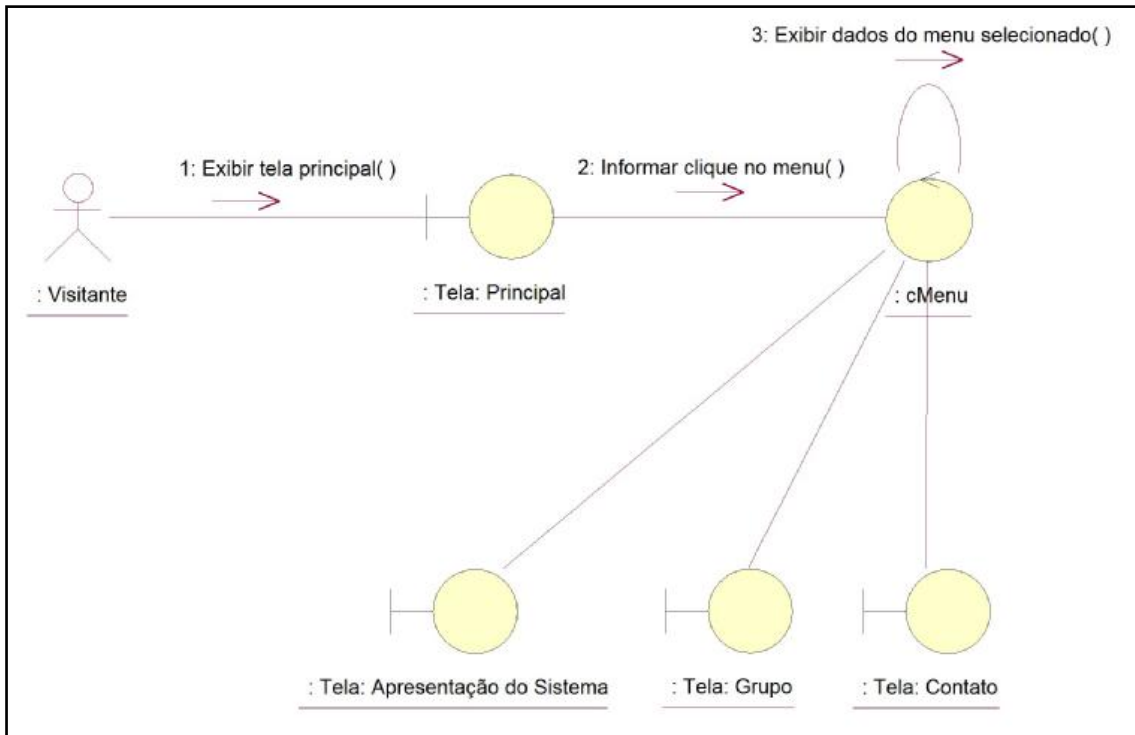


Figura 18: Diagrama de colaboração do caso de uso Visualizar Conteúdo Público – Fluxo Básico. Representa os eventos que o ator Visitante deve realizar para visualizar um conteúdo público, descrevendo as entradas de dados necessárias e o fluxo das mensagens para cada evento. Visualizar Conteúdo Público é o único caso de uso que não requer um usuário cadastrado no sistema.

RESULTADOS E DISCUSSÃO

4. RESULTADOS E DISCUSSÃO

4.1 Caracterização da Amostra de Pacientes

Os dados dos pacientes participantes do estudo foram cedidos ao projeto pelos Professores Dr. Paulo Louzada Jr. e Dr. Geraldo Aleixo S. Passos Jr. As características dos pacientes cadastrados no sistema são apresentadas na Tabela 4. Cento e vinte e seis pacientes participaram do projeto, sendo 94 mulheres e 32 homens. Os pacientes foram divididos em dois grupos. O grupo de treinamento é composto por 106 pacientes e o grupo de teste por 20. Os pacientes do grupo de treinamento apresentam média de idade $54,16 \pm 11,03$ anos, tempo de doença médio de $11,16 \pm 8,84$ anos e é composto por 74,56% de mulheres. O grupo de pacientes de testes apresenta média de idade $53,78 \pm 7,65$ anos, tempo de doença médio de $9 \pm 5,66$ anos e é composto por 75% de mulheres.

Tabela 4: Características dos Pacientes*.

Atributos	Pacientes de Treinamento	Pacientes de Teste
Tempo de doença (anos)	11,16 \pm 8,84	9 \pm 5,66
Idade (anos)	54,16 \pm 11,03	53,78 \pm 7,65
VHS (mm/hora)	32,01 \pm 16,3	43,13 \pm 22,54
Sexo (% Feminino)	74,56	75,00
Tabagistas (% Fumantes)	41,51	50,00
Nível de anti-CCP (% Positivos)	68,87	60,00
(% Negativos)	30,19	15,00
Nível de PCR (% Positivos)	54,72	70,00
(% Negativos)	25,47	10,00
Fator Reumatóide (% Negativos)	23,58	30,00
(% Positivos)	22,64	20,00
(% Título Elevado)	53,77	45,00
DAS 28 (% Alta)	35,84	80,00
(% Moderada)	33,01	10,00
(% Baixa)	19,81	0,00
(% Remissão)	9,43	10,00
Epítipo (%SE/SE)	15,09	15,00
(%SE/X)	35,84	35,00
(%X/X)	34,90	30,00

* Os valores dos atributos Tempo de doença, Idade e VHS estão expressos em relação à Média \pm Desvio Padrão.

4.2 Desenvolvimento do Sistema

O desenvolvimento de sistemas que armazenam e disponibilizam ferramentas para a análise de dados tem se tornado cada vez mais importante, tanto pela agilidade da análise de grandes quantidades de dados como pela facilidade de acesso a estes sistemas. Tais sistemas são utilizados nas mais diversas áreas, inclusive na área médica e biológica.

Um sistema de apoio ao diagnóstico, disponibilizado na Internet, provendo acesso a informações de forma dinâmica e interativa entre os usuários, para auxiliar a análise de dados clínicos e de expressão gênica de pacientes portadores de AR foi desenvolvido para interligar a clínica e a pesquisa, buscando associar relações de diferentes parâmetros analisados e a resposta dos pacientes ao tratamento com DMARDs.

O sistema desenvolvido denomina-se Sistema ARIA. As principais funcionalidades disponibilizadas são: Cadastro de Pacientes, Seleção de Atributos e Predição da Resposta à DMARDs. Uma visão geral dessas funcionalidades presentes no sistema pode ser visualizada na Figura 19. Um paciente cadastrado no sistema pode fazer parte do grupo de treinamento ou do grupo de teste. Caso faça parte do grupo de treinamento os dados do paciente serão utilizados pela seleção de atributos, no treinamento de modelos. Caso contrário, os dados do paciente serão utilizados para a predição da resposta à DMARDs.

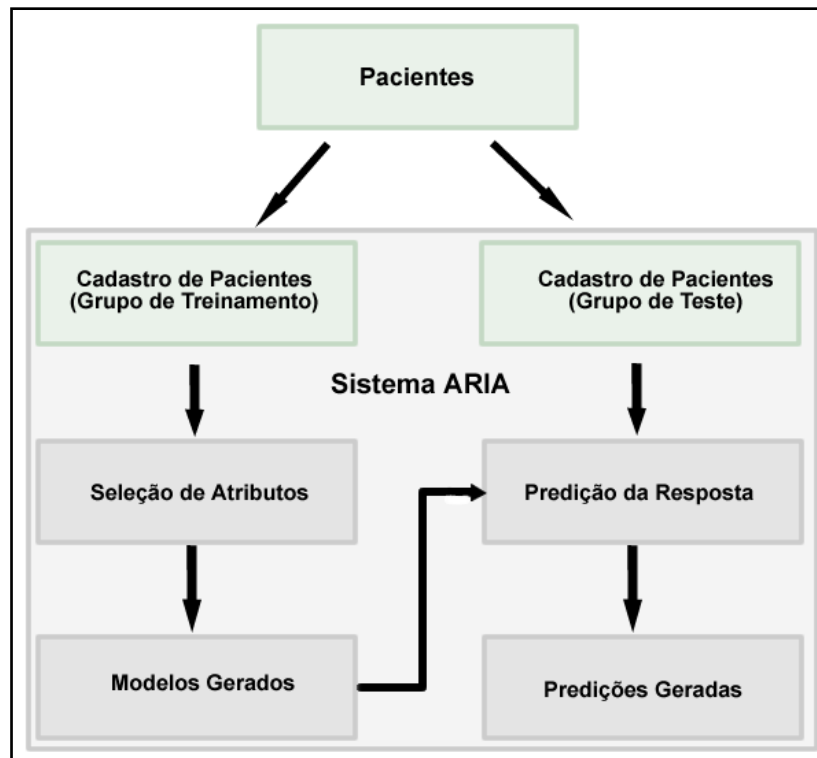


Figura 19: Visão geral das funcionalidades do Sistema ARIA. As principais funcionalidades são: Cadastro de Pacientes, Seleção de Atributos e Predição da Resposta à DMARDs.

O Sistema ARIA está disponível na Internet através do endereço <http://gbi.fmrp.usp.br/artrite>. Pesquisadores colaboradores do projeto e médicos diretamente ligados ao projeto têm permissão para acessá-lo. O acesso aos conteúdos do sistema é feito através de *Login*, fornecendo um nome de usuário e uma senha, previamente cadastrados pelo administrador do sistema. A página principal do sistema é mostrada na Figura 20, na qual é possível observar a apresentação do projeto, informações sobre o grupo de pesquisa (GBI), formulário de contato e o formulário para *Login* no sistema.

ARIA

Apresentação Grupo Contato

English Português

Login do usuário

Nome de usuário: *

Senha: *

Login

Solicitar uma nova senha

Sistema ARIA - Artrite Reumatóide e Inteligência Artificial

A Bioinformática em conjunto com Técnicas de Inteligência Artificial (IA) têm sido aplicadas a diversos problemas biológicos. No presente projeto, as Técnicas de IA, como Redes Neurais Artificiais, Árvores de Decisão e Algoritmos Genéticos, em conjunto com o CMS Drupal serão aplicadas para a construção de um sistema de auxílio ao diagnóstico de pacientes portadores de Artrite Reumatóide (AR).

» [Leia mais](#)

Apoio

FAPESP

Grupo de Bioinformática
Depto. de Genética - FMRP/USP

Universidade de São Paulo - Faculdade de Medicina de Ribeirão Preto - Departamento de Genética - Av. Bandeirantes, 3900 Ribeirão Preto, São Paulo - CEP: 14040-900

Figura 20: Página principal do sistema ARIA. Nesta página, pode ser observada a apresentação do projeto, informações sobre o grupo de pesquisa, o formulário de contato e o formulário de *Login* no sistema.

Para facilitar a navegação e utilização do sistema, ele foi dividido em três partes: Pacientes, Ferramentas e Resultados. Em cada uma das partes citadas estão disponíveis diferentes tipos de conteúdos e formas de acessá-los. Esta estrutura é ilustrada na Figura 21, na qual se observa o bloco **Pesquisar**, utilizado para buscas por palavras-chave, o bloco **Paciente**, que contém as opções de Cadastrar Paciente, Lista de Pacientes, Lista de Pacientes A-Z, Lista de Pacientes com Expressão Gênica e o Filtro de Pacientes. Todas essas listas são diferentes formas de visualizar as informações dos pacientes cadastrados. O bloco **Ferramenta** contém as duas ferramentas baseadas em técnicas de IA disponíveis no sistema, a ferramenta de Seleção de Atributos e a ferramenta de Predição da Resposta à DMARDs. O bloco **Resultados** oferece diversas formas de visualização dos resultados gerados com as ferramentas do sistema. Os modelos gerados através da seleção de atributos podem ser visualizados através da Lista de Seleção de Atributos, da Lista de Modelos AD, Lista de Modelos MLP e a Lista de Modelos *Microarray*. As predições da resposta à DMARDs, realizadas através do sistema, são visualizadas através da Lista de

Predições. O bloco **Usuário** disponibiliza informações sobre a conta do usuário, os conteúdos possíveis de serem criados pelo usuário em questão, assim como as contribuições recentes feitas pelo usuário, e a opção Sair, para efetuar o *Logout* do sistema.

ARIA Apresentação Grupo Contato English Português

Pesquisar a

Pacientes b

- Cadastrar Paciente
- Lista de Pacientes
- Lista de Pacientes A-Z
- Lista de Pacientes com Expressão
- Filtro de Pacientes

Ferramenta c

- Seleção de atributos
- Predição da Resposta de um Paciente

Resultados d

- Lista de Seleção de Atributos
 - Modelos AD
 - Modelos MLP
 - Modelos microarray
- Lista de Predições

gpersinoti e

- Minha conta
- Criar conteúdo
- Contribuições recentes
- Sair

Sistema ARIA - Artrite Reumatóide e Inteligência Artificial

A Bioinformática em conjunto com Técnicas de Inteligência Artificial (IA) têm sido aplicadas a diversos problemas biológicos. No presente projeto, as Técnicas de IA, como Redes Neurais Artificiais, Árvores de Decisão e Algoritmos Genéticos, em conjunto com o CMS Drupal serão aplicadas para a construção de um sistema de auxílio ao diagnóstico de pacientes portadores de Artrite Reumatóide (AR).

> [Leia mais](#)

Apoio

FAPESP

Grupo de Bioinformática
Deplo. de Genética · FMRP/USP

Universidade de São Paulo - Faculdade de Medicina de Ribeirão Preto - Departamento de Genética - Av. Bandeirantes, 3900 Ribeirão Preto, São Paulo - CEP: 14040-900

Figura 21: Front Page do Sistema ARIA. Apresenta os diferentes tipos de conteúdos e diversas formas de acesso e visualização. a. Pesquisar. b. Pacientes. c. Ferramenta. d. Resultados. e. Usuário.

4.2.1 Pacientes

O bloco **Pacientes** trata do tipo de conteúdo **Cadastro de Pacientes**, fornecendo o formulário para o cadastro dos pacientes e as listas e filtros disponíveis para visualizar esse tipo de conteúdo.

Cadastrar Pacientes

O cadastro dos pacientes no sistema é feito de maneira estruturada através do preenchimento de um formulário, apresentado nas Figuras 22 e 23. Este formulário é composto por cinco grupos de dados: Dados Pessoais, Parâmetros Clínicos, Parâmetros Laboratoriais, Informações de *Microarray* e Classe Resposta ao Tratamento.

No grupo **Dados Pessoais** os seguintes itens são solicitados: Nome completo, Registro no Hospital das Clínicas (HC), Data de nascimento, Sexo, Data de preenchimento da ficha de coleta de dados, Etnia e Data da última consulta no HC. Os campos Nome completo, Registro no HC, Data de Nascimento e Sexo são de preenchimento obrigatório. No grupo de **Parâmetros Clínicos** devem ser cadastrados: Tempo de diagnóstico em anos, Tempo decorrido até o início do tratamento, Número de critérios diagnósticos do *American College of Rheumatology* (ACR), Número de articulações acometidas, Edema persistente de articulações interfalangeanas proximais, Tabagismo, Presença de manifestações extra-articulares, Questionário de qualidade de vida (HAQ), Atividade da doença (DAS 28), Classe funcional, Medicamentos utilizados e Determinação da gravidade. Já o grupo de **Parâmetros Laboratoriais** é composto por: Elevação das provas de atividade inflamatória, Nível de Fator Reumatóide, Anticorpos anti-Peptídeo Cíclico Citrulinado (anti-CCP), Nível da proteína C reativa (PCR), Velocidade de Hemossedimentação (VHS), Fator anti-nuclear (FAN), Presença de epítipo compartilhado, Presença de DR de proteção,

Tipagem HLA DR e Erosões radiológicas. O grupo de **Informações de Microarray** contém os seguintes itens: se o paciente possui expressão gênica determinada e os arquivos de dados brutos provenientes dos experimentos de *microarray* realizados. O grupo **Classe Resposta ao Tratamento** é composto pela classe resposta ao tratamento com DMARDs. Esse campo somente é preenchido quando os dados do paciente que está sendo cadastrado serão utilizados para o treinamento da ferramenta de Seleção de Atributos. Caso contrário, esse campo poderá ser deixado em branco, e o sistema poderá ser utilizado para prever a resposta do paciente em questão, através da ferramenta de Predição da Resposta à DMARDs. A resposta cadastrada no sistema é a resposta apresentada pelos pacientes três meses a um ano após a data de inclusão no banco de dados. Os pacientes foram avaliados, após este período e caso o paciente tenha recebido uma dose de metotrexato de 15 a 20 mg/m²/semana, acompanhada ou não de leflunomida, e não houve melhora, este paciente é considerado não respondedor à DMARDs. Caso contrário, o paciente é considerado respondedor à DMARDs.

Todos os usuários cadastrados no sistema possuem permissão para cadastrar um paciente. Um usuário cadastrado como médico ou colaborador só poderá editar ou apagar um cadastro de paciente, caso tenha inserido o registro em questão. Já o usuário administrador tem permissão para inserir, editar e apagar qualquer cadastro de paciente.

Estão cadastrados, até o momento, 106 pacientes para o treinamento da ferramenta de Seleção de Atributos. Vinte e cinco dos 106 pacientes possuem informações de expressão gênica cadastradas. Vinte pacientes estão cadastrados para testes, utilizando a ferramenta de Predição da Resposta à DMARDs.

Enviar Cadastro de Pacientes

▼ Dados Pessoais a

Nome completo: *

Registro no HC: *

Data de Nascimento:

Sexo: *

Feminino

Masculino

Data de Preenchimento:

Data do preenchimento da ficha de coleta de dados

Etnia:

Data da última consulta:

Data da última consulta de acordo com o prontuário do HC

▼ Parâmetros Clínicos b

Diagnóstico em: (ano):

Tempo até o início do tratamento:

Tempo transcorrido entre o início da doença e o início do tratamento

Número de critérios ACR:

Número de articulações acometidas:

Edema persistente de articulações interfalangeanas proximais:

Sim

Não

?

Tabagismo:

Fumante

Não Fumante

?

Presença de manifestações extra-articulares:

Não

Sim

?

Questionário de qualidade de vida HAQ:

Atividade da Doença - DAS 28:

Classe funcional:

Medicamentos:

- anti-TNF
- Azatioprina
- Ciclosporina
- Ciclosporina EV
- Ciclosporina VO
- Cloramb
- Cloroquina
- Corticosteróides AL
- Corticosteróides IS
- Corticosteróides Pulso
- D-penicilamina
- Leflunomide
- Metotrexato
- Sal Ouro
- SSA

Determinação da Gravidade:

Leve/Moderada

Grave

?

Figura 22: Formulário de Cadastro de Paciente do Sistema ARIA. a. Dados Pessoais. b. Parâmetros Clínicos devem ser preenchidos.

▼ Parâmetros Laboratoriais c

Elevação das provas de atividade inflamatória :

Sim
 Não
 ?

Nível de Fator Reumatóide:

Negativo
 Positivo
 Título elevado
 ?

Anticorpos anti-CCP:

Positivo
 Negativo
 ?

Nível da proteína C reativa:

Positivo
 Negativo
 ?

Velocidade de Hemossedimentação:

0

Fator anti-nuclear:

Negativo
 Positivo
 ?

Presença do epítipo compartilhado:

...

Presença de DR de proteção:

Não
 Sim
 ?

HLA DR:

Erosões radiológicas :

Sim
 Não
 ?

Erosões radiológicas nos 12 primeiros meses de doença

▼ Informações Microarray d

Possui Expressão gênica:

Não
 Sim

▼ Arquivo bruto L

Changes made to the attachments are not permanent until you save this post.

Anexar novo arquivo:

Allowed extensions: txt tav xls zip

▼ Arquivo bruto R

Changes made to the attachments are not permanent until you save this post.

Anexar novo arquivo:

Allowed extensions: txt tav xls zip

▼ Classe Resposta ao Tratamento e

Classe Resposta ao tratamento com DMARDs:

...

Figura 23: Continuação do Formulário de Cadastro de Pacientes. c. Parâmetros Laboratoriais. d. Informações de *Microarray*. e. Classe Resposta ao Tratamento.

Categorias

Quando um paciente é cadastrado no sistema, as informações inseridas são automaticamente categorizadas para que futuras buscas possam ser executadas de forma rápida e dinâmica. Esta funcionalidade é implementada através do módulo *Taxonomy Fields*. Este módulo permite que campos do formulário de cadastro de pacientes sejam automaticamente categorizados.

Os campos do formulário de cadastro de pacientes (categorias) são os seguintes: Anticorpos anti-CCP, Classe funcional, Atividade da doença (DAS 28), Determinação da Gravidade, Edema persistente de articulações interfalangeanas proximais, Elevação de atividade inflamatória, Erosões Radiológicas, Etnia, Fator anti-nuclear (FAN), Haploidia, Medicamentos, Nível de fator reumatóide, Nível de Proteína C Reativa, Possui Expressão Gênica Determinada, Presença de DR de proteção, Presença de Manifestações extra-articulares, Sexo, Tabagismo e a Classe Resposta ao Tratamento.

As categorias criadas e o módulo *Views* são utilizados para a criação de listas e filtros, disponibilizando, assim, diversas maneiras de acesso e visualização dos conteúdos do sistema.

Listas e Filtros

Após o cadastro dos pacientes no sistema é necessário oferecer uma forma de visualizar e acessar os conteúdos inseridos e para isso são criados listas e filtros. As listas e os filtros foram criados utilizando o módulo *Views* e as categorias.

O sistema ARIA oferece ao usuário quatro modos de visualização do conteúdo cadastro de pacientes. Entre eles estão a Lista de Pacientes, Lista de Pacientes A-Z, Lista de Pacientes com Expressão Gênica e o Filtro de Pacientes.

A **Lista de Pacientes** é uma forma de visualizar todos os pacientes cadastrados no sistema e é apresentada na Figura 24. Nesta listagem, o título do Paciente é um *link* para a página completa contendo todas as informações sobre o paciente em questão (Figura 24) e, logo abaixo do título, é apresentado um quadro resumo sobre cada paciente, contendo o Registro no HC, a Data de Nascimento, Sexo e a Classe Resposta ao Tratamento com DMARDs.

Lista de Pacientes

ID 8 - Paciente 435588 a

Registro no HC: 435588 b
Data de Nascimento: 19/04/1937
Sexo: Feminino
Classe Resposta ao tratamento com DMARDs: Não Respondeu ao Tratamento com DMARDs

ID 9 - Paciente 41446

Registro no HC: 41446
Data de Nascimento: 14/02/1942
Sexo: Masculino
Classe Resposta ao tratamento com DMARDs: Não Respondeu ao Tratamento com DMARDs

ID 10 - Paciente 362812

Registro no HC: 362812
Data de Nascimento: 18/07/1969
Sexo: Feminino
Classe Resposta ao tratamento com DMARDs: Não Respondeu ao Tratamento com DMARDs

ID 11 - Paciente 731284

Registro no HC: 731284
Data de Nascimento: 07/01/1975
Sexo: Feminino
Classe Resposta ao tratamento com DMARDs: Respondeu ao Tratamento com DMARDs

Figura 24: Lista de Pacientes do Sistema ARIA. a. *Link* para a página contendo todas as informações do paciente (ver Figura 24). b. Quadro resumo sobre o paciente.

Na Figura 25, é ilustrada a maneira que um registro de paciente é apresentado ao usuário através do sistema. No sistema esse tipo de informação é armazenada como um *node*.

ID 8 - Paciente 435588

▼ Dados Pessoais

Registro no HC: 435588
Data de Nascimento: 19/04/1937
Sexo: Feminino
Etnia: Caucasião

▼ Parâmetros Clínicos

Diagnóstico em: (ano): 1969
Número de critérios ACR: 7
Tabagismo: Não Fumante
Presença de manifestações extra-articulares: Sim
Atividade da Doença - DAS 28: Moderada
Classe funcional: Classe III
Medicamentos:
Corticosteróides AL
Corticosteróides IS
Cloroquina
Metotrexato
Derterminação da Gravidade: Grave

▼ Parâmetros Laboratoriais

Nível de Fator Reumatóide: Positivo
Anticorpos anti-CCP: Positivo
Nível da proteína C reativa: Positivo
Fator anti-nuclear: Negativo
Velocidade de Hemossedimentação: 45
Presença do epítipo compartilhado: SE/SE
Presença de DR de proteção: Não
HLA DR: 04./04
Erosões radiológicas : Sim

▼ Informações Microarray

Possui Expressão gênica: Não

▼ Classe Resposta ao Tratamento

Classe Resposta ao tratamento com DMARDs: Não Respondeu ao Tratamento com DMARDs

Figura 25: Visualização de um registro de um paciente.

A **Lista de Pacientes A-Z** é uma lista de pacientes organizada pelo nome do paciente em ordem alfabética. Esta lista é sub-dividida em duas outras listagens. A primeira apresenta, em forma de tabela, todos os pacientes que já possuem classe resposta ao tratamento, contendo a Identificação do Paciente, como um *link* para o seu registro, o nome do paciente e a resposta que o paciente apresentou ao tratamento com DMARDs. Esta lista apresenta os pacientes utilizados para o treinamento da ferramenta de seleção de atributos. A outra listagem apresenta, em forma de tabela, todos os pacientes que não possuem classe resposta ao tratamento, contendo apenas a Identificação do Paciente e o Nome do Paciente. Esta lista apresenta os pacientes que terão a resposta à DMARDs predita através da ferramenta de predição da resposta à DMARDs disponível no sistema.

O **Filtro de Pacientes** é um tipo de busca específica que utiliza as categorias para filtrar quais pacientes serão mostrados no resultado da listagem. Este filtro é exemplificado na Figura 26, na qual os filtros selecionados foram: nível de fator reumatóide **negativo**, se paciente possui expressão gênica determinada através de *microarrays* igual a **sim** e atividade da doença **alta**. O resultado dessa busca apresentou dois pacientes ID 96 e ID 104.

Todas as categorias podem ser utilizadas para pesquisas. O bloco Pesquisa é utilizado para efetuar buscas por palavras-chave. Caso essa busca não retorne um resultado satisfatório é possível utilizar a Busca Avançada, na qual todas as categorias estão disponíveis para buscas.

Filtro de Pacientes

Medicamentos	Nível de Fator Reumatóide	Expressão Gênica (microarray)	Atividade da Doença DAS 28
<div style="border: 1px solid #ccc; padding: 5px;"> Is All Of ▾ <All> anti-TNF Azatioprina Ciclosporina Ciclosporina EV Ciclosporina VO Cloramb Cloroquina </div>	<div style="border: 1px solid #ccc; padding: 5px;"> Is All Of ▾ <All> Negativo Positivo Título elevado ? </div>	<div style="border: 1px solid #ccc; padding: 5px;"> Is All Of ▾ <All> Não Sim </div>	<div style="border: 1px solid #ccc; padding: 5px;"> Is All Of ▾ <All> Alta Moderada Baixa Remissão clínica ? </div>
<input type="button" value="Enviar"/>			

ID 96 - Paciente 446264

Registro no HC: 446264
 Data de Nascimento: 22/03/1968
 Sexo: Feminino

Classe Resposta ao tratamento com DMARDs: Não Respondeu ao Tratamento com DMARDs

ID 104 - Paciente 649057

Registro no HC: 649057
 Data de Nascimento: 03/10/1957
 Sexo: Feminino

Classe Resposta ao tratamento com DMARDs: Não Respondeu ao Tratamento com DMARDs

Figura 26: Filtro de Pacientes do Sistema ARIA. A pesquisa realizada neste exemplo apresenta os pacientes que possuem as seguintes características: Nível de fator reumatóide Negativo, Possui expressão gênica determinada através de *microarray* Sim e a Atividade da doença Alta.

A **Lista de Pacientes com Expressão Gênica** é outra forma de visualização dos pacientes. Neste caso são listados apenas os pacientes que possuem informações sobre a expressão gênica. Esta lista é apresentada em forma de tabela, contendo as seguintes informações a respeito dos pacientes: Identificação do Paciente, Hábito Tabagista, Haplótipo do Epítipo Compartilhado, Nível de Fator Reumatóide, Nível da Proteína C Reativa, Nível de anticorpos anti-CCP e Atividade da Doença. Esta listagem pode ser observada na Figura 27.

Lista de Pacientes com Dados de Expressão Gênica						
Paciente	Tabagismo	Epítipo	Fator Reumatóide	Proteína C Reativa	Anticorpos anti-CCP	DAS 28
ID 10 - Paciente 362812	Não Fumante	X/X	Positivo	Negativo	Negativo	Moderada
ID 11 - Paciente 731284	Não Fumante	X/X	Positivo	Negativo	Positivo	Baixa
ID 12 - Paciente 705219	Não Fumante	X/X	Positivo	?	Positivo	Moderada
ID 14 - Paciente 285642	Não Fumante	?	Título elevado	Positivo	Positivo	Alta
ID 111 - Paciente 126798	Não Fumante	?	Título elevado	Positivo	Positivo	Alta
ID 18 - Paciente 23402	Não Fumante	X/X	Título elevado	?	Positivo	Moderada
ID 32 - Paciente 71035	Não Fumante	X/X	Título elevado	Positivo	Negativo	Alta
ID 40 - Paciente 546981	Fumante	SE/X	Título elevado	Negativo	Positivo	Remissão clínica
ID 46 - Paciente 115666	Fumante	SE/X	Negativo	Positivo	Negativo	Remissão clínica
ID 110 - Paciente 279283	?	?	Positivo	Positivo	?	Moderada
ID 49 - Paciente 749363	Não Fumante	SE/SE	Título elevado	?	Positivo	Alta
ID 52 - Paciente 423737	Fumante	SE/SE	Título elevado	Positivo	Positivo	Alta
ID 112 - Paciente 103306	Não Fumante	?	Título elevado	Negativo	Positivo	Alta
ID 113 - Paciente 283596	Fumante	?	Título elevado	Positivo	Positivo	Moderada
ID 57 - Paciente 729029	Fumante	SE/SE	Negativo	?	Negativo	?
ID 62 - Paciente 172988	Não Fumante	SE/X	Positivo	Positivo	Positivo	Moderada
ID 63 - Paciente 475636	Não Fumante	?	Título elevado	Positivo	Positivo	Alta
ID 113 - Paciente 283596	Fumante	?	Título elevado	Positivo	Positivo	Moderada
ID 57 - Paciente 729029	Fumante	SE/SE	Negativo	?	Negativo	?
ID 62 - Paciente 172988	Não Fumante	SE/X	Positivo	Positivo	Positivo	Moderada
ID 63 - Paciente 475636	Não Fumante	?	Título elevado	Positivo	Positivo	Alta
ID 64 - Paciente 735945	Não Fumante	SE/SE	Título elevado	Positivo	Positivo	Alta
ID 77 - Paciente 302102	Fumante	X/X	Positivo	Positivo	Positivo	Alta
ID 114 - Paciente 506542	?	?	Negativo	Positivo	Negativo	Moderada
ID 93 - Paciente 715835	Fumante	SE/X	Título elevado	Positivo	Positivo	Alta
ID 96 - Paciente 446264	Não Fumante	SE/X	Negativo	Negativo	Negativo	Alta
ID 99 - Paciente 72915	Não Fumante	SE/SE	Título elevado	Positivo	Positivo	Baixa
ID 104 - Paciente 649057	Fumante	X/X	Negativo	Positivo	Negativo	Alta
ID 105 - Paciente 729349	Não Fumante	X/X	Título elevado	Positivo	Positivo	Alta

Figura 27: Lista de Pacientes com Expressão Gênica do Sistema ARIA. As informações apresentadas nesta lista são: ID do Paciente, Tabagismo, Haplótipo do Epítipo Compartilhado, Nível de Fator Reumatóide, Nível da Proteína C Reativa, Nível de Anticorpos anti-CCP e Atividade da Doença.

4.2.2 Ferramentas

As ferramentas disponibilizadas no Sistema ARIA são a **Seleção de Atributos** e a **Predição da Resposta à DMARDs**. Essas ferramentas são, ambas, baseadas em técnicas de inteligência artificiais, como árvores de decisão, redes neurais artificiais e algoritmos genéticos. Estas ferramentas serão discutidas nas sessões a seguir.

Seleção de Atributos

A ferramenta de seleção de atributos é responsável pelo treinamento de modelos que serão utilizados na predição da resposta à DMARDs e pela seleção de um subconjunto de atributos que seja suficientemente bom para a classificação dos pacientes. Para execução da seleção de atributos são utilizados os dados dos pacientes do grupo de treinamento, dados clínicos ou dados de expressão gênica. Para realizar a busca pelo subconjunto de atributos que será selecionado, a ferramenta utiliza a busca genética. A busca genética emprega um algoritmo genético como método de busca. Para a avaliação do subconjunto de atributos, é utilizado um algoritmo de indução, que neste caso, pode ser uma rede neural artificial MLP ou uma árvore de decisão. A precisão do indutor é utilizada pelo algoritmo genético, como função de avaliação. Ao final da execução da busca pelo subconjunto de atributos selecionados, uma avaliação do modelo final é feita utilizando os dados de treinamento e algoritmo de indução utilizado na fase de avaliação dos subconjuntos, durante a busca genética. Na Figura 28, é apresentado um diagrama mostrando o funcionamento da ferramenta de seleção de atributos.

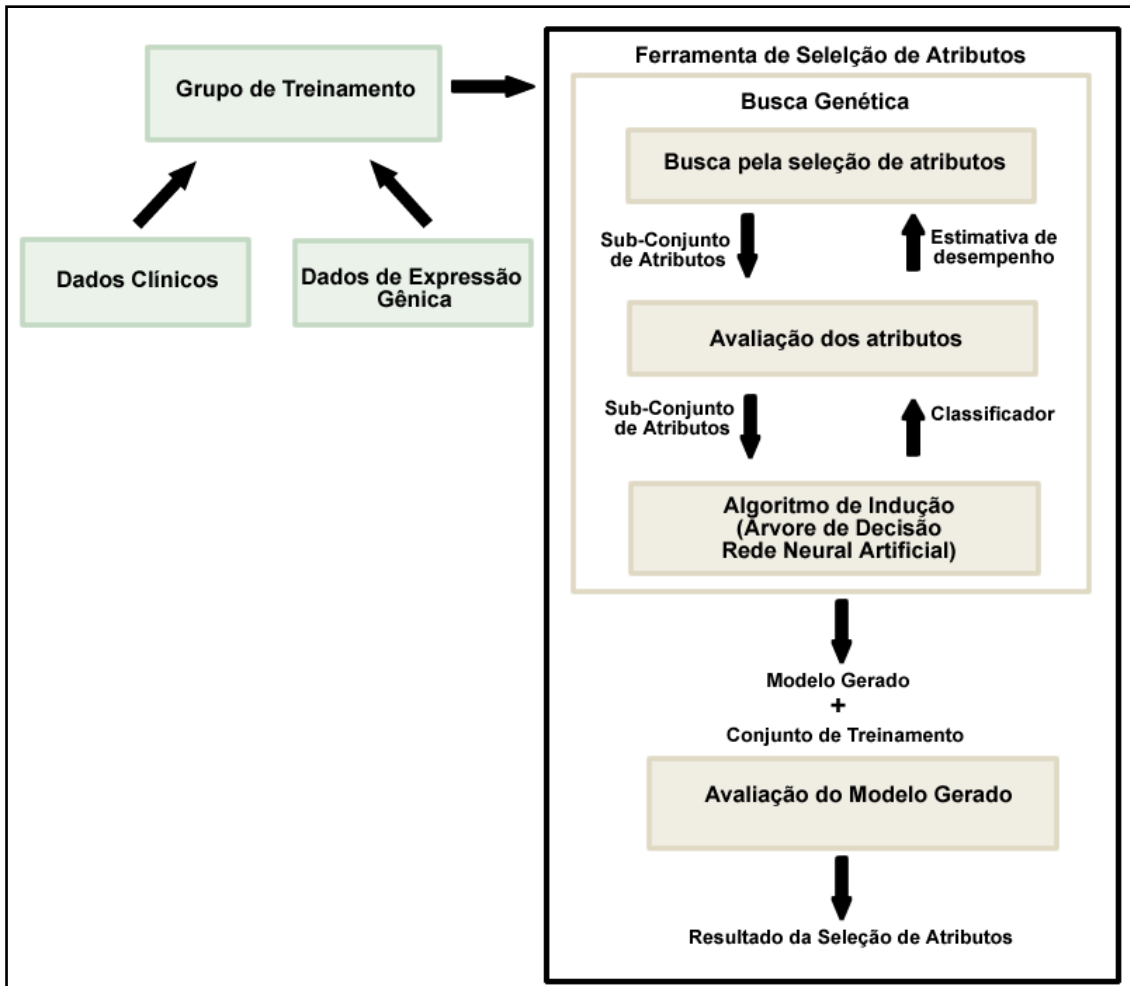


Figura 28: Visão geral do funcionamento da ferramenta de seleção de atributos disponível no sistema ARIA.

Para a execução da ferramenta, primeiramente, os dados dos pacientes são recuperados do banco de dados através da linguagem SQL e é gerado um arquivo arff. O formato arff é um formato específico nativo ao Weka. Este arquivo é composto por um cabeçalho e pelos dados. O cabeçalho contém o nome do conjunto de dados e uma lista de todos os atributos. A lista de atributos deve definir o tipo do atributo ou os valores que ele pode assumir. Para definir o nome do conjunto de dados utiliza-se a sintaxe *@relation nome_do_conjunto_de_dados*. E para definir a lista de atributos utiliza-se a sintaxe *@attribute nome_do_atributo tipo_do_atributo* ou *@attribute nome_do_atributo valor1, valor2, ... , valorN*. Os dados são definidos através da

sintaxe `@data` seguida dos dados separados por vírgula. Um exemplo de um arquivo `arff` é apresentado na Figura 29.

```
@relation artrite

@attribute Idade numeric
@attribute DAS28 {Alta,Moderada,Baixa,RemissaoClinica}
@attribute Sexo {Masculino,Feminino}
@attribute antiCCP {Positivo,Negativo}
@attribute PCR {Positivo,Negativo}
@attribute Classe {Respondeu,NaoRespondeu}

@data

71,Moderada,Feminino,Positivo,Positivo,NaoRespondeu
66,Alta,Masculino,Negativo,Positivo,NaoRespondeu
39,Moderada,Feminino,Negativo,Negativo,NaoRespondeu
33,Baixa,Feminino,Positivo,Negativo,Respondeu
30,Moderada,Feminino,Positivo,?,NaoRespondeu
68,Alta,Feminino,Negativo,Positivo,NaoRespondeu
64,Alta,Feminino,Positivo,Positivo,NaoRespondeu
55,Alta,Masculino,Positivo,?,NaoRespondeu
69,Moderada,Feminino,Positivo,Positivo,Respondeu
51,Moderada,Feminino,Negativo,Positivo,NaoRespondeu
61,Moderada,Feminino,Positivo,?,NaoRespondeu
50,Moderada,Masculino,Positivo,Positivo,Respondeu
62,Moderada,Masculino,Positivo,Negativo,NaoRespondeu
44,Moderada,Feminino,Positivo,Negativo,NaoRespondeu
66,Baixa,Feminino,Positivo,Negativo,Respondeu
```

Figura 29: Exemplo de um arquivo no formato `arff`. `@relation` define o nome do conjunto de dados. `@attribute` é utilizado para definir a lista de atributos. `@data` define os dados do conjunto.

O código da ferramenta é escrito em Java e utiliza o pacote `weka.jar` no qual estão disponíveis todas as classes disponibilizadas no *software* Weka. Estas classes são utilizadas na ferramenta para realizar a seleção de atributos.

A ferramenta de seleção de atributos é apresentada na Figura 30. Ela é composta por duas partes. Na primeira delas, o usuário deve selecionar qual tipo de dado de treinamento será utilizado na seleção de atributos, os dados clínicos ou os dados de expressão gênica. Em seguida, são apresentados os parâmetros *default* da busca genética, são eles: taxa de mutação, taxa de *crossover*, número máximo de gerações e tamanho da população. Fica a critério do usuário alterá-los. A segunda parte da ferramenta envolve a escolha do algoritmo de classificação que será utilizado

na seleção de atributos. Novamente, fica a critério do usuário alterar os parâmetros dos classificadores. Os parâmetros possíveis de serem alterados,

Submissão de Seleção de Atributos

A seleção de atributos visa reduzir o número de atributos, encontrando um subconjunto de atributos com melhor acurácia de classificação.

Dados
Selecionar qual tipo de dados será utilizado na seleção de atributos

▼ Tipos de dados a

Dados Clínicos

Dados de Microarray

Busca Genética

▶ Parâmetros da Busca Genética b

Algoritmos de Classificação
O Algoritmos de classificação desempenhará o papel de buscar uma função que permita associar corretamente cada registro do banco de dados a um único rótulo, denominado classe.

Árvore de Decisão c

▶ Parâmetros da Arvore de Decisão d

Rede Neural Artificial

▶ Parâmetros da Rede Neural Artificial

Figura 30: Ferramenta de Seleção de Atributos do Sistema ARIA. a. Selecionar o tipo de dado de treinamento a ser utilizado para a seleção de atributos. b. Parâmetros da busca genética. c. Selecionar o algoritmo de classificação a ser utilizado na seleção de atributos. d. Parâmetros dos algoritmos de classificação.

Uma vez selecionados os parâmetros necessários, a seleção de atributos será executada no servidor e o resultado será apresentado ao usuário. Na Figura 31 pode ser observado o resultado da seleção de atributos que é apresentado ao usuário. A primeira informação apresentada é o número de atributos selecionados e, logo abaixo, são apresentados quais os atributos foram selecionados. Em seguida, é apresentada a estrutura da árvore de decisão gerada. Essa informação somente é disponibilizada quando o classificador selecionado for árvore de decisão indutiva. Após a ilustração da árvore, encontra-se a avaliação do classificador, que é feita utilizando-se o método de

validação cruzada (*cross-validation*). Este método divide o conjunto de treinamento em subconjuntos, treina o classificador usando $n-1$ conjuntos e o testa no conjunto que não foi utilizado no treinamento. A validação utilizada foi o *10-fold cross validation* que subdivide o conjunto de treinamento em 10 subconjuntos. Em seguida, é apresentada a Matriz de Confusão gerada, mostrando o número de classificações corretas *versus* as classificações preditas para cada classe, sobre o conjunto de treinamento. Com os dados apresentados na matriz é possível calcular outras métricas de avaliação como a sensibilidade e especificidade, que são apresentadas no final do resultado da seleção de atributos. No exemplo mostrado na Figura 31, foram selecionados como dados de treinamento os dados clínicos e, como classificador, uma árvore de decisão.

Caso o classificador selecionado para a execução da ferramenta de seleção de atributos seja uma rede neural artificial, o resultado da seleção de atributos é semelhante ao resultado quando o classificador utilizado é uma árvore de decisão, exceto pela descrição da árvore que não é apresentada. Um exemplo de um resultado de uma seleção de atributos, utilizando como classificador uma rede neural artificial é apresentado na Figura 32. Neste exemplo, mostrado na figura, o conjunto de treinamento utilizado é composto por dados clínicos.

Resultado da seleção de atributos

Resultado pronto

Numero de atributos selecionados: 3

Atributos Selecionados

4: VHS
5: DAS28
16: PCR

a

Arvore de decisao gerada

DAS28 = Alta: NaoRespondeu (38.73)
DAS28 = Moderada: NaoRespondeu (35.67/11.0)
DAS28 = Baixa
| PCR = Positivo: Respondeu (12.96/0.93)
| PCR = Negativo
| | VHS <= 15: NaoRespondeu (3.08)
| | VHS > 15: Respondeu (5.37/0.39)
DAS28 = RemissaoClinica: Respondeu (10.19/1.19)

b

Avaliacao do classificador

Instancias corretamente classificadas	93	87.74 %
Instancias incorretamente classificadas	13	12.26 %
Estatistica Kappa	0.71	
Erro Absoluto Medio	0.19	
Erro Medio Quadratico	0.3	
Erro Absoluto Relativo	40.98	
Erro Relativo Quadratico	40.98	
Numero total de instancias	106	

c

Matriz de Confusao

a	b	<-- Classificado como
67	2	a = Respondeu ao Tratamento com DMARDs
11	26	b = Nao Respondeu ao Tratamento com DMARDs

d

Sensibilidade: 97.1%

Especificidade: 70.27%

e

Figura 31: Resultado da Seleção de Atributos utilizando como classificador uma árvore de decisão. a. Apresenta os atributos selecionados. b. Apresenta a árvore de decisão gerada. c. Apresenta a avaliação do classificador nos dados de treinamento após a seleção de atributos. d. Apresenta a matriz de confusão do classificador. e. Sensibilidade e Especificidade do classificador.

Resultado da seleção de atributos

Resultado pronto

Numero de atributos selecionados: 10

Atributos Selecionados

1: Tempo_de_doenca
 2: Idade
 3: Numero_de_criterios_acr
 4: VHS
 5: DAS28
 7: Etnia
 9: Manifestacoes_extra_articulares
 12: FR
 14: Erosoes
 18: DRprotecao

a

Avaliacao do classificador

Instancias corretamente classificadas	102	96.226 %
Instancias incorretamente classificadas	4	3.774 %
Estatistica Kappa	0.916	
Erro Absoluto Medio	0.092	
Erro Medio Quadratico	0.189	
Erro Absoluto Relativo	20.274 %	
Erro Relativo Quadratico	39.58 %	
Numero total de instancias	106	

b

Matriz de Confusao

a	b	<-- Classificado como
34	3	a = Respondeu ao Tratamento com DMARDs
1	68	b = Nao Respondeu ao Tratamento com DMARDs

c

Sensibilidade: 91.892 %

Especificidade: 98.551 %

d

Figura 32: Resultado da Seleção de Atributos, utilizando como classificador uma rede neural artificial. a. Apresenta os atributos selecionados. b. Apresenta a avaliação do classificador nos dados de treinamento após a seleção de atributos. c. Apresenta a matriz de confusão do classificador. d. Sensibilidade e Especificidade do classificador.

Predição da Resposta à DMARDs

A ferramenta de Predição da Resposta à DMARDs utiliza os modelos previamente treinados através da ferramenta de Seleção de Atributos para prever a resposta de um novo paciente. Os pacientes disponibilizados para teste da ferramenta não possuem a informação sobre a classe resposta, quando são cadastrados no sistema.

Para executar a predição da resposta à DMARDs é necessário escolher um paciente e um modelo previamente treinado através do sistema. A interface da ferramenta é apresentada na Figura 33.

Predição da Resposta de um Paciente

Predição da resposta de um paciente ao tratamento com DMARDs (Drogas Anti-Reumáticas Modificadoras da Doença) através de um modelo previamente treinado.

Predição da Resposta a DMARDs

Selecionar um paciente para a predição da resposta a DMARDs através do sistema ARIA.

Paciente: ID 173 - Paciente 453367 ▾

Selecionar o modelo para ser utilizado na predicao da resposta do paciente.

Modelo: Submissão 210 - 2009-01-23T15:29:55 ▾

Avançar

Figura 33: Ferramenta de Predição da Resposta de um Paciente à DMARDs. O usuário deve selecionar um paciente e um modelo para submeter uma predição.

Após a seleção dos parâmetros necessários, a predição será executada e o resultado será exibido ao usuário. O resultado de uma predição pode ser visualizado na Figura 34.



Figura 34: Resultado da Predição da Resposta de um Paciente à DMARDs.

4.2.3 Resultados

Cada vez que uma das ferramentas do sistema é executada, o resultado gerado pela ferramenta é armazenado no banco de dados para que futuras consultas possam ser efetuadas. Os resultados armazenados podem ser acessados através do bloco **Resultados**, que disponibiliza a Lista de Seleção de Atributos, Lista de Modelos AD, Lista de Modelos MLP, Lista de Modelos de *microarray* e a Lista de Predições.

A **Lista de Seleção de Atributos** apresenta todas as seleções de atributos submetidas ao sistema ARIA, nas quais os dados de treinamento selecionados foram dados clínicos. Essa lista apresenta a identificação da submissão, a data da submissão, o tipo de dados e o algoritmo de classificação utilizado na seleção de atributos. Em seguida, é apresentado o resultado da submissão de seleção de atributos, contendo o número de atributos selecionados, quais os atributos selecionados e o arquivo de resultado, que apresenta todas as informações sobre o resultado da seleção de atributos em questão. A Lista de Seleção de Atributos é mostrada na Figura 35.

Lista de Submissões de Seleção de Atributos

Lista todas as seleções de atributos submetidas ao sistema.

Submissão 229 - 27/01/2009 11:51

Seleção de Atributos

Tipo de dados: clínicos

Algoritmo de classificação: ad

Resultado

Número de atributos selecionados: 3

Atributos selecionados:

VHS

DAS28

PCR

Arquivo de Resultado: 229_2009-01-27BRST11:51:.txt

Submissão 228 - 26/01/2009 15:16

Seleção de Atributos

Tipo de dados: clínicos

Algoritmo de classificação: ad

Resultado

Número de atributos selecionados: 3

Atributos selecionados:

VHS

DAS28

PCR

Arquivo de Resultado: 228_2009-01-26BRST15:16:.txt

Figura 35: Lista de submissões de seleção de Atributos. A lista de seleção de atributos apresenta as informações da seleção, como a data de submissão, tipo de dados, algoritmo de classificação, e o resultado da seleção, contendo o número de atributos, os atributos selecionados e o arquivo de resultado, que contém a avaliação e a matriz de confusão do classificador gerado.

A **Lista de Modelos AD** apresenta, de forma mais sucinta, através de tabela, as seleções de Atributos realizadas por meio do sistema ARIA, nas quais os dados de treinamento são os dados clínicos e o algoritmo de árvore de decisão foi utilizado como classificador. Esta listagem é mostrada na Figura 36 e apresenta a Identificação da Seleção de Atributos, o Número de Atributos Selecionados, os Atributos Selecionados e o Arquivo de Resultado.

Lista de Modelos AD Treinados			
Lista os modelos treinados para a seleção de atributos, nos quais foi utilizado como classificador uma Árvore de Decisão.			
ID da Seleção de Atributos	Número de Atributos	Atributos Selecionados	Resultado
Submissão 229 - 27/01/2009 11:51	3	VHS DAS28 PCR	229_2009-01-27BRST11:51.txt
Submissão 228 - 26/01/2009 15:16	3	VHS DAS28 PCR	228_2009-01-26BRST15:16.txt
Submissão 223 - 24/01/2009 17:21	3	VHS DAS28 PCR	223_2009-01-24T17:21:37.txt
Submissão 162 - 05/11/2008 - 23:39	1	DAS28	162_2008-11-05T23-39-45.txt
Submissão 159 - 05/11/2008 - 23:16	2	DAS28 Gravidade	159_2008-11-05T23-16-57.txt
Submissão 157 - 11/05/2008 - 22:42	2	DAS28 EpitopoCompartilhado	157_2008-11-05T22-42-51.txt
Submissão 146 - 31/10/2008 - 03:16	9	Tempo_de_doenca Idade VHS DAS28 Manifestacoes_extra_articulares Gravidade antiCCP FAN DRprotecao	146_2008-10-31T05:16:23.txt

Figura 36: Lista de Seleção de Atributos realizadas através do sistema ARIA, nas quais Árvores de Decisão foram utilizadas como classificadores e os dados de treinamento foram dados clínicos.

A **Lista de Modelos MLP** apresenta, em forma de tabela, as seleções de Atributos realizadas através do sistema ARIA, nas quais os dados de treinamento são os dados clínicos e redes neurais artificiais MLP foram utilizadas como algoritmo de classificação. Esta listagem é mostrada na Figura 37 e apresenta a Identificação da Seleção de Atributos, o Número de Atributos Selecionados, os Atributos Selecionados e o Arquivo de Resultado.

Lista de Modelos MLP Treinados			
Lista os modelos treinados para a seleção de atributos, nos quais foi utilizado como classificador uma rede neural MLP.			
ID da Seleção de Atributos	Número de Atributos	Atributos Seleccionados	Resultado
Submissão 209 - 23/01/2009 10:34	11	Numero_de_criterios_acr DAS28 Sexo Tabagismo Manifestacoes_extra_articulares ClasseFuncional Gravidade FR EpitopoCompartilhado PCR FAN	209_2009-01-23T10:34:04.txt
Submissão 208 - 23/01/2009 10:18	2	VHS DAS28	208_2009-01-23T10:18:20.txt
Submissão 207 - 23/01/ 2009 10:10	6	Tempo_de_doenca Idade DAS28 Manifestacoes_extra_articulares Gravidade DRprotecao	207_2009-01-23T10:10:17.txt
Submissão 206 - 23/01/2009 10:04	1	DAS28	206_2009-01-23T10:04:50.txt
Submissão 200 - 10/12/2008 10:00	8	Tempo_de_doenca Idade DAS28 Tabagismo Manifestacoes_extra_articulares ClasseFuncional Gravidade DRprotecao	200_2008-12-10T10:00:32.txt
Submissão 161 - 05/11/2008 - 23:34	3	DAS28 Sexo antiCCP	161_2008-11-05T23-34-02.txt
Submissão 133 - 10/10/2008 - 20:06	12	Tempo_de_doenca Idade Numero_de_criterios_acr VHS DAS28 Etnia Manifestacoes_extra_articulares Gravidade antiCCP PCR FAN DRprotecao	133_2008-10-10T22:06:00.txt

Figura 37: Lista de Seleção de Atributos realizadas através do sistema ARIA, nas quais Redes Neurais Artificiais, MLP foram utilizadas como classificadores e como dados de treinamento, os dados clínicos.

A **Lista de Modelos de *microarray*** apresenta as seleções de Atributos realizadas através do sistema ARIA, nas quais os dados de treinamento são os dados de *microarray*. Esta listagem é mostrada na Figura 38 e apresenta a Identificação da Seleção de Atributos, o Número de Atributos Seleccionados e o Arquivo de Resultado. É possível visualizar os atributos selecionados clicando-se na Identificação da Seleção de Atributos ou no Arquivo de Resultado. Os atributos não são apresentados na listagem, pois normalmente os dados de *microarray* apresentam muitos genes (atributos) o que dificulta a visualização em forma de tabelas.

Lista de Seleção de Atributos microarray			
Lista as seleções de atributos realizadas utilizando dados de microarray.			
ID da Seleção de Atributos	Algoritmo de Classificação	Número de Atributos Seleccionados	Arquivo de Resultado
Submissão 224 - 25/01-2009 15:50	ad	1912	224_2009-01-25T15:50:02.txt

Figura 38: Lista de Seleção de Atributos realizadas através do Sistema ARIA, nas quais os dados de treinamento foram dados de expressão gênica.

Diversas listas que mostram as seleções de atributos foram criadas a fim de prover melhores formas de visualizar cada um dos tipos de seleção de atributos realizados através do sistema, para facilitar a comparação e análise dos diversos modelos gerados pela ferramenta de seleção de atributos.

A **Lista de Predições** apresenta todas as predições realizadas através do sistema ARIA que estão armazenadas no banco de dados. Esta lista pode ser visualizada na Figura 39 e apresenta o Identificador da Predição, o Identificador do Paciente, o Identificador do Modelo e a Predição gerada pela ferramenta de Predição da Resposta à DMARDs. Tanto o identificador do paciente como o do modelo são *links* para as respectivas páginas contendo as informações completas desses conteúdos.

Lista das Predições Realizadas			
Lista todas as predições realizadas através do sistema.			
ID Predição	Paciente	Modelo	Predição
Predicao: 213 - Paciente: 173	ID 173 - Paciente 453367	Submissão 158 - 11/05/2008 - 23:09	Respondeu ao Tratamento com DMARDs
Predicao: 212 - Paciente: 173	ID 173 - Paciente 453367	Submissão 210 - 2009-01-23T15:29:55	Não Respondeu ao Tratamento com DMARDs
Predicao: 211 - Paciente: 173	ID 173 - Paciente 453367	Submissão 210 - 2009-01-23T15:29:55	Não Respondeu ao Tratamento com DMARDs
Predicao: 205 - Paciente: 173	ID 173 - Paciente 453367	Submissão 198 - 2008-12-08T11:34:14	Não Respondeu ao Tratamento com DMARDs
Predicao: 204 - Paciente: 173	ID 173 - Paciente 453367	Submissão 203 - 2009-01-14T15:51:34	Respondeu ao Tratamento com DMARDs
Predicao: 196 - Paciente: 173	ID 173 - Paciente 453367	Submissão 162 - 05/11/2008 - 23:39	Não Respondeu ao Tratamento com DMARDs
Predicao: 193 - Paciente: 173	ID 173 - Paciente 453367	Submissão 162 - 05/11/2008 - 23:39	Não Respondeu ao Tratamento com DMARDs
Predicao: 172 - Paciente: 11	ID 11 - Paciente 731284	Submissão 162 - 05/11/2008 - 23:39	Respondeu ao Tratamento com DMARDs
Predicao: 165 - Paciente: 11	ID 11 - Paciente 731284	Submissão 133 - 10/10/2008 - 20:06	Respondeu ao Tratamento com DMARDs
Predição: 144 - Paciente: 8	ID 8 - Paciente 435588	Submissão 132 - 10/10/2008 - 20:24	Não Respondeu ao Tratamento com DMARDs

Figura 39: Lista de Predições realizadas através do sistema ARIA. a. Identificador da Predição. b. Identificador do Paciente, link para a página contendo as informações do paciente. c. Modelo utilizado na predição, link para a página contendo as informações do modelo. d. Predição gerada.

4.3 Análise dos Resultados

Diversos modelos foram treinados a fim de se encontrar um modelo preditivo adequado aos dados dos pacientes portadores de AR. Para se comparar os diversos modelos gerados foram usadas diversas métricas de avaliação, além da precisão do classificador. A precisão é a relação entre o número de instâncias corretamente classificadas e o número total de instâncias. Mesmo sendo a métrica mais utilizada, outras métricas são apresentadas como: Desvio Padrão, Estatística Kappa, Erro Absoluto Médio, Erro Médio Quadrático, Sensibilidade e Especificidade.

Na Tabela 5 são apresentadas as métricas de avaliação dos modelos gerados utilizando como classificador, Árvores de Decisão e dados clínicos como dados de treinamento. As métricas foram calculadas a partir do método de validação cruzada.

Tabela 5: Métricas de avaliação dos modelos gerados utilizando como classificadores, Árvores de Decisão e dados clínicos como dados de treinamento.

Métricas de Avaliação	Padrão*		Submissão 258		Submissão 259 e 257	
Instâncias corretamente classificadas	87	82,08%	92	86,79 %	93	87.74 %
Instâncias incorretamente classificadas	19	17,92%	14	13,21 %	13	12.26 %
Desvio Padrão	11,20%		9,43%		9,60%	
Estatística Kappa	0,61		0,69		0,71	
Erro Absoluto Médio	0,21		0,19		0,18	
Erro Médio Quadrático	0,39		0,31		0,30	
Sensibilidade	78,38		67,56 %		70,27 %	
Especificidade	84,05		97,10 %		97,10 %	
Número total de instâncias			106			

* Avaliação da árvore de decisão gerada sem a execução da seleção de atributos.

A árvore de decisão gerada sem a execução da seleção de atributos (padrão) apresentou precisão de 82,02%, sensibilidade de 78,38% e 84,05% de especificidade. Quando comparada com os modelos gerados após a seleção de atributos o modelo padrão foi o que apresentou a maior sensibilidade (78,38%), porém apresentou precisão um pouco mais baixa que os demais modelos.

As submissões 259 e 257 apresentaram resultados idênticos na avaliação. Ambas selecionaram quatro atributos: velocidade de hemossedimentação (VHS), índice de atividade da doença (DAS 28), gravidade da doença e nível da proteína C reativa (PCR). Os modelos apresentam precisão próxima a 87%, especificidade de 97,10% e sensibilidade de 70,27%.

A árvore de decisão gerada pelas submissões 257 e 259 é apresentada na Figura 40. Observando a figura, nota-se que o atributo gravidade não foi utilizado na construção da árvore.

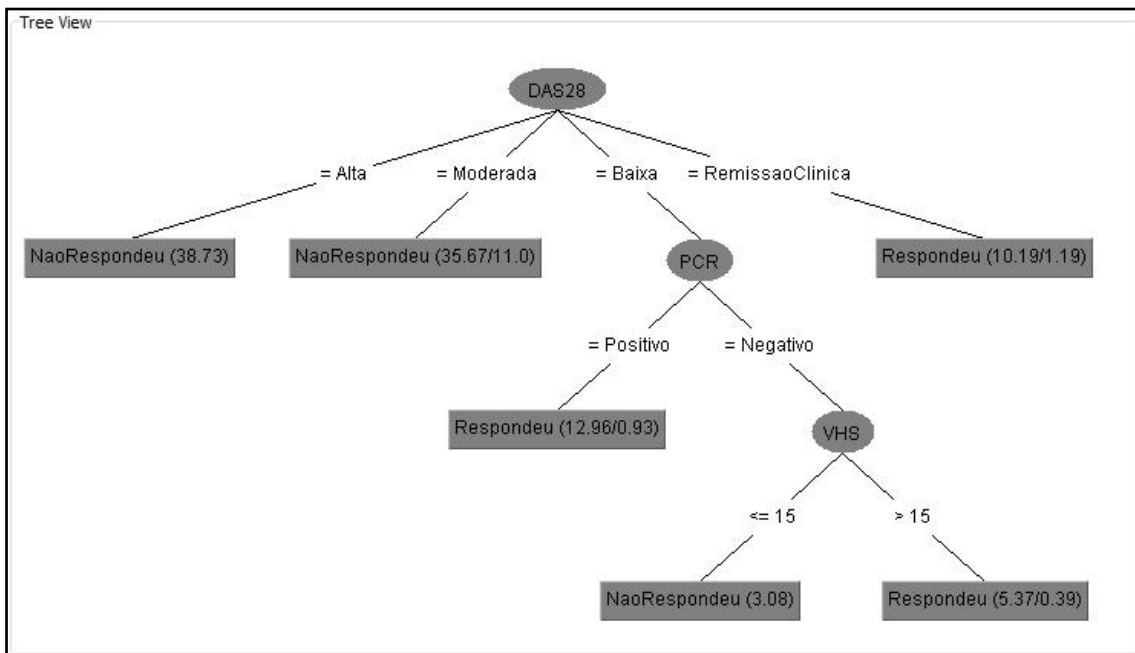


Figura 40: Árvore de decisão resultante das Submissões 257 e 259. Foram selecionados, nestas submissões, os atributos: VHS, DAS 28, gravidade da doença e PCR. Apenas o atributo gravidade não foi utilizado na construção da árvore.

A submissão 258 teve como resultado a seleção de cinco atributos. Foram eles: número de critérios ACR, velocidade de hemossedimentação, índice de atividade da doença (DAS 28), presença de manifestações extra-articulares e nível de proteína C reativa. Este modelo apresentou precisão próxima a 87%, sensibilidade de 67,56% e especificidade de 90,10%. A árvore de decisão gerada pela Submissão 258 é apresentada na Figura 41, na qual se observa que apenas os atributos DAS 28 e VHS foram utilizados para a construção da árvore.

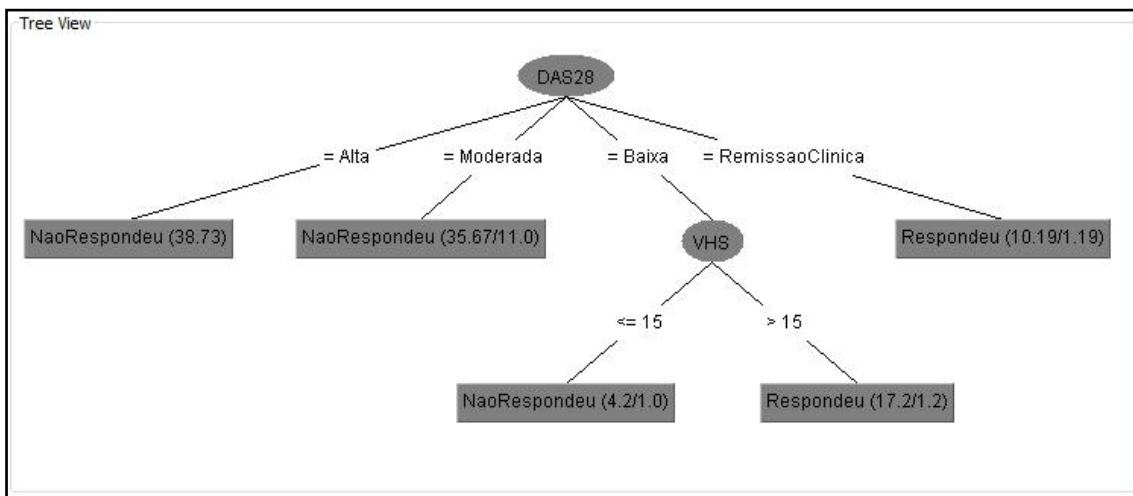


Figura 41: Árvore de decisão gerada pela Submissão 258. Esta submissão selecionou os atributos: número de critérios ACR, VHS, DAS 28, manifestações extra-articulares e PCR, mas apenas DAS 28 e VHS foram utilizados na construção da árvore.

Os modelos gerados no treinamento da ferramenta de seleção de atributos, utilizando, como classificadores, árvores de decisão e como dados de treinamento, dados clínicos foram novamente avaliados, agora utilizando o conjunto de testes.

Na Tabela 6 são apresentadas as métricas de avaliação, no conjunto de pacientes de teste, dos modelos gerados utilizando, utilizando como classificadores, árvores de decisão. Os três modelos apresentados obtiveram exatamente o mesmo desempenho no conjunto de teste, apresentando precisão de 90%.

Tabela 6: Métricas de avaliação no conjunto de teste dos modelos gerados utilizando como classificadores, Árvores de Decisão.

Métricas de Avaliação	Árvores de Decisão	
Instâncias Corretamente Classificadas	18	90,00%
Instâncias Incorretamente Classificadas	2	20,00 %
Desvio Padrão	20,10%	
Estatística Kappa	0	
Erro Médio Absoluto	0,10	
Erro Médio Quadrático	0,24	
Número Total de Instâncias	20	

Outros modelos foram gerados, utilizando agora as redes neurais artificiais como classificadores e os dados clínicos como dados de treinamento. Esses modelos foram avaliados utilizando o conjunto de treinamento, e as métricas de avaliação são apresentadas na Tabela 7. As métricas foram calculadas, novamente, a partir do método de validação cruzada. São apresentados, quatro modelos distintos. Um modelo gerado antes da seleção de atributos, denominado padrão e três modelos gerados após a seleção de atributos, Submissões 260, 263 e 271.

O modelo de rede neural construído antes da seleção de atributos teve como resultado precisão, especificidade e sensibilidade inferiores aos obtidos pelos demais modelos, construídos após a seleção. O modelo Submissão 260 teve como resultado a seleção de apenas dois atributos, velocidade de hemossedimentação (VHS) e índice de atividade da doença (DAS 28). Este modelo apresentou precisão de 85,85%. O modelo 263 apresentou como resultado a seleção de seis atributos, tempo de doença, idade, índice de atividade da doença, presença de manifestações extra-articulares, gravidade da doença e presença de HLA DR de proteção. Este modelo obteve precisão de 90,56%. O modelo Submissão 271 teve como resultado a seleção de 10 atributos, tempo de doença, idade, número de critérios ACR, velocidade de hemossedimentação, índice de atividade da doença DAS 28, etnia, presença de manifestações extra-articulares, fator reumatóide, presença de erosões radiológicas, e

a presença de DR de proteção. Esse modelo apresentou precisão de 96,23%, que foi a precisão mais alta alcançada. Essa taxa foi alcançada na validação cruzada, ou seja, no treinamento do modelo.

Tabela 7: Métricas de avaliação dos modelos gerados utilizando como classificadores, Redes Neurais Artificiais MLP e dados clínicos como dados de treinamento.

Métricas de Avaliação	Padrão*		Submissão 260		Submissão 263		Submissão 271	
Instâncias corretamente classificadas	82	77,36%	91	85,85%	96	90,56%	102	96,23%
Instâncias incorretamente classificadas	24	22,64%	15	14,15%	10	9,43%	4	3,77%
Desvio Padrão	12,17%		9,80%		10,21%		11,71%	
Estatística Kappa	0,50		0,67		0,79		0,92	
Erro Absoluto Médio	0,25		0,20		0,15		0,09	
Erro Médio Quadrático	0,45		0,31		0,26		0,19	
Sensibilidade	67,56%		70,27 %		83,78 %		91,89%	
Especificidade	82,60%		94,20 %		94,20 %		98,55%	
Número total de instâncias	106							

* Avaliação do modelo MLP sem a execução da seleção de atributos.

Os modelos gerados no treinamento da ferramenta de seleção de atributos, utilizando redes neurais e dados clínicos foram novamente avaliados, agora utilizando o conjunto de pacientes de testes.

Na Tabela 8 são apresentadas as métricas de avaliação no conjunto de testes dos modelos gerados utilizando como classificadores as Redes Neurais Artificiais. Os modelos apresentaram precisão variando de 80% a 90%.

Tabela 8: Métricas de avaliação no conjunto de teste dos modelos gerados utilizando como classificador, Redes Neurais Artificiais.

Métricas de Avaliação	Submissão 260		Submissão 263		Submissão 271	
Instâncias Corretamente Classificadas	18	90,00%	16	80,00%	18	90,00%
Instâncias Incorretamente Classificadas	2	10,00%	4	20,00%	2	10,00%
Desvio Padrão	20,10%		20,10%		20,10%	
Estatística Kappa	0		0,11		0	
Erro Médio Absoluto	0,07		0,14		0,10	
Erro Médio Quadrático	0,21		0,31		0,31	
Número Total de Instâncias	20					

Os modelos gerados utilizando RN foram modelos distintos entre si, sendo que houve um modelo que selecionou apenas dois atributos e outro que apresentou a seleção dez atributos. O modelo que selecionou mais atributos foi, no entanto, o que apresentou valor de precisão mais alto, no conjunto de treinamento. Uma desvantagem na utilização de modelos de RN é o tempo computacional gasto para o treinamento da rede, que variou de 6 minutos a 38 minutos, enquanto o tempo gasto para o treinamento das árvores de decisão, não ultrapassou 12 segundos. O tempo computacional foi mensurado executando o treinamento dos modelos no servidor em que o sistema está hospedado (HP ProLiant ML Pentium Xeon 2.0). Além de menor tempo computacional necessário para o treinamento, as árvores de decisão fornecem a representação simbólica do classificador induzido, que na área médica, tem um peso importante, pois se apresentam de forma mais clara e objetiva, podendo ser facilmente transcritos em regras, e tais regras podem ser verificadas na prática clínica.

Foram gerados dois modelos utilizando dados de expressão gênica. O primeiro deles utiliza como classificador uma árvore de decisão e o segundo, utiliza uma rede neural artificial. Estes modelos foram treinados e avaliados no grupo de pacientes que possui dados de expressão gênica provenientes de experimentos de *microarray* e as métricas resultantes da avaliação são apresentadas na Tabela 9.

Tabela 9: Métricas de avaliação dos modelos gerados utilizando dados de expressão gênica como dados de treinamento.

Métricas de Avaliação	Submissão 276	Submissão 277
Instâncias Corretamente Classificadas	45 97,82%	33 71,14%
Instâncias Incorretamente Classificadas	1 2,17%	13 28,66%
Desvio Padrão	17,78%	21,76%
Estatística Kappa	0,95	0,44
Erro Médio Absoluto	0,04	0,36
Erro Médio Quadrático	0,13	0,42
Sensibilidade	100%	45,83%
Especificidade	95,45%	100 %
Número Total de instâncias	46	

* Modelo gerado sem a execução da seleção de atributos

O modelo gerado pela submissão 276 teve como resultado a seleção de 1912 genes. Este modelo apresentou precisão de 97,82%, sensibilidade de 100% e especificidade de 95,45%. O modelo gerado pela submissão 277, teve como resultado a seleção de 1921 genes e apresentou precisão de 71,14%, sensibilidade de 45,83% e especificidade de 100%. Este modelo, não apresenta precisão tão alta quanto o modelo gerado pela submissão 276, mas foi capaz de classificar corretamente todos pacientes do conjunto de treinamento não respondedores ao tratamento com DMARDs.

Os modelos gerados pelo Sistema ARIA levam em consideração características clínicas ou características de expressão gênica. Isso foi definido, pois os modelos baseados em características clínicas levam em consideração informações de rotina na AR, que são acessíveis aos médicos na prática clínica. Já os modelos baseados em dados de *microarray*, levam em consideração informações de expressão gênica que não são tão acessíveis quanto às informações clínicas e assim, são mais difíceis de serem verificados, uma vez que, os experimentos são, atualmente, muito custosos. Mesmo assim, os modelos baseados em expressão gênica não foram descartados, pois podem ser refinados e, futuramente, utilizados para, por exemplo, pesquisas de marcadores moleculares para a AR.

A construção de modelos preditivos para a Artrite Reumatóide é foco de outros trabalhos. Um modelo proposto para a predição do prognóstico de pacientes com artrite indiferenciada foi sugerido por van der Helm-van Mil e colaboradores (Van Der Helm-Van Mil et al., 2007). Este modelo propõe o cálculo de um índice de predição baseado em 9 características clínicas: sexo, idade, localização dos sintomas, rigidez matinal, número de articulações dolorosas e número de articulações inchadas, nível da proteína C reativa, fator reumatóide positivo e anticorpos anti-CCP positivos. O valor do índice de predição pode variar de 0 a 14, sendo que quanto maior o valor do índice maior é a chance do paciente evoluir para artrite reumatóide.

Outro trabalho proposto por Combe e colaboradores (Combe et al., 2001) teve como objetivo identificar fatores preditivos para o dano radiológico e para a progressão do dano em pacientes portadores de artrite reumatóide. Este trabalho avaliou as seguintes características: idade, sexo, índice de massa corpórea, tempo de doença, duração da rigidez matinal, dor medida pela escala visual analógica, número de articulações dolorosas e edemaciadas, índice articular de Ritchie, índice de atividade da doença (DAS 28), presença de nódulos e manifestações extra-articulares, questionário de qualidade de vida (HAQ), velocidade de hemossedimentação, nível de proteína C reativa, nível de fator reumatóide, anticorpos anti-queratina, anticorpos anti-RA33, anticorpos anti-HSP90, anticorpos anti-calpastatina, YKL-40, anticorpos anti-nulcear, HLA-DRB1 e DQB1 genotipagem, índice erosões radiológicas, índice de espaçamento das articulações e índice de dano radiológico total. Como resultado, o índice de dano radiológico total foi identificado como o melhor fator para predição do dano e sua progressão.

Esses trabalhos citados anteriormente levam em consideração muitas das características que foram avaliadas no presente projeto, mostrando a importância de construção de modelos preditivos que levam em consideração informações clínicas acessíveis, para que os modelos possam ser validados pela clínica.

No presente projeto, a principal contribuição da metodologia empregada é a possibilidade de aplicar mais de um algoritmo e assim, gerar diferentes modelos, oferecendo flexibilidade aos médicos na geração, escolha e avaliação do melhor modelo para os dados dos pacientes em questão.

CONCLUSÕES

5. CONCLUSÕES

Sistemas computacionais devidamente projetados para a área biológica e médica são extremamente importantes e cada vez mais necessários devido ao volume de dados gerados pelas novas tecnologias disponíveis. Sistemas *Web* são ideais, pois são muito abrangentes devido à independência de plataforma e a popularização do acesso da Internet atualmente alcançado. Além disso, os sistemas *Web* agregam os bancos de dados que são importantes para o armazenamento das informações de forma organizada e estruturada.

O sistema ARIA foi desenvolvido e está disponível para o acesso de pesquisadores e médicos colaboradores do projeto através do endereço <http://gbi.fmrp.usp.br/artrite>. O sistema oferece informações sobre pacientes portadores de AR e as duas ferramentas baseadas em técnicas de IA, a ferramenta de seleção de atributos e a ferramenta de predição da resposta à DMARDs. As ferramentas estão disponíveis e funcionais.

Os modelos gerados pelo sistema são factíveis com a realidade da artrite reumatóide. A característica que foi selecionada pela maioria dos modelos foi o índice de atividade da doença. Isso se deve ao fato de que o índice DAS 28 leva em conta outras características, como o número de articulações dolorosas e edemaciadas, o VHS e a escala visual analógica que mede o estado de saúde geral do paciente. Os modelos gerados utilizando árvores de decisão foram considerados mais adequados que os modelos gerados utilizando redes neurais artificiais, pois selecionam, em média, menos atributos, utilizam menos tempo computacional para treinamento e, principalmente, por apresentarem a representação simbólica do modelo gerado. A representação simbólica foi considerada importante, principalmente na área médica, pois fornece, de forma objetiva, uma visualização da relação dos atributos selecionados e a resposta dos pacientes. O modelo criado utilizando redes neurais que apresentou um número maior de atributos selecionados (10 atributos) foi, no

entanto, o modelo que atingiu a precisão mais alta no treinamento. Quanto ao teste, os modelos tanto de árvores como de redes apresentaram taxas próximas a 90% de precisão.

O sistema ARIA, assim como qualquer sistema, poderá sofrer alterações decorrentes de atualizações e manutenções necessárias, a fim de crescer, de modo flexível, podendo abranger outras funcionalidades e outros tipos de dados, para complementá-lo. Uma das alterações que será feita é aumentar o número de pacientes para o treinamento dos modelos gerados pela ferramenta de seleção de atributos. Isso irá fazer com que as predições possam ser mais confiáveis. Outra alteração que visa aumentar a confiabilidade do sistema é a validação dos modelos em um número maior de pacientes de teste.

O sistema ARIA não será disponibilizado ao público, porém espera-se que este sistema seja de grande utilidade aos pesquisadores e médicos envolvidos no projeto, por se tratar de um ambiente estruturado, organizado que oferece uma nova abordagem de análise da resposta dos pacientes ao tratamento com DMARDs.

***REFERÊNCIAS
BIBLIOGRÁFICAS***

6. REFERÊNCIAS BIBLIOGRÁFICAS

ALAMANOS, Y. , DROSOS, A. A. Epidemiology of adult rheumatoid arthritis. **Autoimmunity Reviews**, v.4, n.3, p.130-136. 2005.

ALETAHA, D., BREEDVELD, F. C. , SMOLEN, J. S. The Need for New Classification Criteria for Rheumatoid Arthritis. **Arthritis & Rheumatism**, v.52, n.11, p.3333-3336. 2005.

AMERICAN COLLEGE OF RHEUMATOLOGY SUBCOMMITTEE ON RHEUMATOID ARTHRITIS GUIDELINES. Guidelines for the Management of Rheumatoid Arthritis. **Arthritis & Rheumatism**, v.46, n.2, p.328-346. 2002.

ARNETT, F. C., EDWORTHY, S. M., BLOCH, D. A., MCSHANE, D. J., FRIES, J. F. , COOPER, N. S. The American Rheumatism Association 1987 revised criteria for classification of rheumatoid arthritis. **Arthritis & Rheumatism**, v.31, p.315-324. 1988.

BALDI, P. , BRUNNAK, S. **Bioinformatics: The Machine Learning Approach**. Cambridge. MIT Press. 1998.

BARANAUSKAS, J. A. **Extração de conhecimento por múltiplos indutores**. Tese (Doutorado). Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2001. 183 p.

BÉRTOLO, M. B., COSTALLAT, L. T. L., PERSOLI, L. B. , COSTA, F. Alelos HLA-DRB1 e o prognóstico da Artrite Reumatóide em pacientes brasileiros. **Revista Brasileira de Reumatologia**, v.41, n.3, p.151-156. 2001.

BOOCH, G., JACOBSON, I. , RUMBAUGH, J. **UML, Guia do Usuário tradução**. Rio de Janeiro. Campus. 2006.

BRASIL. Protocolo Clínico e Diretrizes Terapêuticas - Artrite Reumatóide Cloroquina, Hidroxicloroquina, Sulfassalazina, Metotrexato, Ciclosporina, Leflunomida, Agentes anti-Citocinas. T. E. I. E. Secretaria De Ciência 2006.

CARDOSO, C. **UML na prática: do problema ao sistema**. Rio de Janeiro. Editora Ciência Moderna Ltda. 2003.

CHOY, E. H. , PANAYI, G. S. Cytokine pathways and joint inflammation in rheumatoid arthritis. **N Engl J Med**, v.344, n.12, p.907-916. 2001.

COMBE, B., DOUGADOS, M., GOUPILLE, P., CANTAGREL, A., ELIAOU, J. F., SIBILIA, J., MEYER, O., SANY, J., DAURE`S, J. P. , DUBOIS, A. Prognostic Factors for Radiographic Damage in Early Rheumatoid Arthritis. **Arthritis & Rheumatism**, v.44, n.8, p.1736-1743. 2001.

COOPER, N. J. Economic burden of rheumatoid arthritis: a systematic review. **Rheumatology**, v.39, n.1, p.28-33. 2000.

CRAVEN, M. W. , SHAVLIK, J. W. Machine Learning Approaches to gene recognition. **IEEE Expert**, v.9, n.2, p.2-10. 1994.

D'HAESELEER, P., LIANG, S. , SOMOGYI, R. **Gene expression data analysis and modeling**. Pacific Symposium on Biocomputing. Hawaii, USA. 1999.

DAVIDSON, A. , DIAMOND, B. Autoimmune diseases. **N Engl J Med**, v.345, n.5, p.340-350. 2001.

DELGADO-VEGA, A. M. , ANAYA, J.-M. Meta-analysis of HLA-DRB1 polymorphism in Latin American patients with rheumatoid arthritis. **Autoimmunity Reviews**, v.6, n.6, p.402-408. 2007.

DING, S., LIU, J. , YANG, Q. **A Genetic algorithm applied to optimal gene subset selection**. IEEE Congress on Evolutionary Computation. 2004.

DORLAND, W. A. N., ANDERSON, D. M., ALBERT, D. M., BEHRMAN, R. E. , BARASH, P. G. **Dicionário Médico Ilustrado Dorland**. São Paulo. Manole. 1999.

DOUGLASS, R. T., LITTLE, M. , SMITH, J. W. **Building Online Communities with Drupal, phpBB, and WordPress**. Apress. 2006.

FAVRETTO, F. O., TINÓS, R. , CARVALHO, A. C. P. F. L. **Selection of Sensors in an Artificial Tongue via Genetic Algorithms**. I Workshop on Computational Intelligence (WCI) and Brazilian Symposium on Neural Networks (SBRN'06). Ribeirão Preto, SP, Brazil. 2006.

FELDMANN, M., BRENNAN, F. M. , MAINI, R. N. Rheumatoid arthritis. **Cell**, v.85, n.3, p.307-310. 1996.

FERRARI, F. A. **Crie Banco de Dados em MySQL**. São Paulo. Digeratti Books. 2007.

FUKUNAGA, K. **Introduction to Statistical Pattern Recognition**. Academic Press. 1990.

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLIER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D. , LANDER, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. **Science**, v.286, n.5439, p.531-537. 1999.

GONDRO, C. , KINGHORN, B. P. A simple genetic algorithm for multiple sequence alignment. **Genet Mol Res**, v.6, n.4, p.964-982. 2007.

GUELICH, S. , GUNDAVAN, S. **Programação CGI com Perl (O Reilly)**. Rio de Janeiro. Editora Ciência Moderna Ltda. 2001.

HAYKIN, S. **Redes Neurais Artificiais: Princípios e Práticas**. Bookman. 2001.

HECHT-NIELSEN, R. **Theory of the Backpropagation Neural Network**. International Joint Conference on Neural Networks (IJCNN). Washington, DC, USA. 1989.

HO, S. Y., LEE, C. C., CHEN, H. M. , HUANG, H. L. **Efficient Gene Selection for Classification of Microarray Data**. IEEE Congress on Evolutionary Computation. Edinburgh, UK. 2005.

HOCHBERG, M. C., CHANG, R. W., DWOSH, I., LINDSEY, S., PINCUS, T. , WOLFE, F. The American College of Rheumatology 1991 revised criteria for the classification of global functional status in rheumatoid arthritis. **Arthritis & Rheumatism**, v.35, n.5, p.498-502. 1992.

HUANG, J., SHIMIZU, H. , SHIOYA, S. Clustering gene expression pattern and extracting relationship in gene network based on artificial neural networks. **J Biosci Bioeng**, v.96, n.5, p.421-428. 2003.

JAWAHEER, D. , GREGERSEN, P. K. Rheumatoid arthritis: The genetic components. **Rheumatic Disease Clinics of North America**, v.28, n.1, p.1-15. 2002.

KAROPKA, T., FLUCK, J., MEVISSSEN, H. T. , GLASS, A. The Autoimmune Disease Database: a dynamically compiled literature-derived database. **BMC Bioinformatics**, v.7, p.325. 2006.

LEECH, S. Molecular mimicry in autoimmune disease. **Arch Dis Child**, v.79, n.5, p.448-451. 1998.

LINDEN, R. **Algoritmos Genéticos**. Rio de Janeiro. Brasport. 2006.

LORENA, A. C. , CARVALHO, A. C. P. F. L. **Utilização de Técnicas Inteligentes em Bioinformática**. Relatório Técnico. São Carlos. 2003.

MEIDANIS, J. , SETÚBAL, J. C. **Introduction to Computational Molecular Biology**. Boston. PWS. 1997.

MELONI, J. C. **Fundamentos de PHP**. Rio de Janeiro. Editora Ciência Moderna Ltda. 2000.

MERCER, D. **Drupal - Creating Blogs, Forums, Portals and Community Websites**. Packt Publishing. 2006.

MITCHELL, M. **An introduction to genetic algorithms**. Cambridge. MIT Press. 1996.

MITCHELL, T. **Machine Learning**. McGraw-Hill. 1997.

NGUYEN, H. D., YOSHIHARA, I., YAMAMORI, K. , YASUNAGA, M. Aligning multiple protein sequences by parallel hybrid genetic algorithm. **Genome Inform**, v.13, p.123-132. 2002.

NOTREDAME, C. , HIGGINS, D. G. SAGA: sequence alignment by genetic algorithm. **Nucleic Acids Res**, v.24, n.8, p.1515-1524. 1996.

NOTREDAME, C., O'BRIEN, E. A. , HIGGINS, D. G. RAGA: RNA sequence alignment by genetic algorithm. **Nucleic Acids Res**, v.25, n.22, p.4570-4580. 1997.

O'DELL, J. R. Therapeutic strategies for rheumatoid arthritis. **N Engl J Med**, v.350, n.25, p.2591-2602. 2004.

QUACKENBUSH, J. Computational analysis of microarray data. **Nat Rev Genet**, v.2, n.6, p.418-427. 2001.

QUINLAN, R. Induction of Decision Trees. **Machine Learning**, v.1, p.81-106. 1986.

RICH, E. , KNIGHT, K. **Artificial Intelligence**. McGraw-Hill. 1991.

SALEEM, B., COX, S. R. , EMERY, P. Biomarkers: Strategies to predict outcome of rheumatoid arthritis. **Drug Discovery Today: Therapeutic Strategies**, v.3, n.1, p.11-16. 2006.

SENNA, E. R., DE BARROS, A. L., SILVA, E. O., COSTA, I. F., PEREIRA, L. V., CICONELLI, R. M. , FERRAZ, M. B. Prevalence of rheumatic diseases in Brazil: a study using the COPCORD approach. **J Rheumatol**, v.31, n.3, p.594-597. 2004.

SOUTO, M. C. P., LORENA, A. C., DELBEM, A. C. B. , CARVALHO, A. C. P. F. L. **Técnicas de Aprendizado de Máquinas para Problemas em Biologia Molecular**. . Renata Vieira. (Org.). II Jornadas de Atualização em Inteligência Artificial, SBC. 2003.

VAN DER HELM-VAN MIL, A. H., LE CESSIE, S., VAN DONGEN, H., BREEDVELD, F. C., TOES, R. E. , HUIZINGA, T. W. A prediction rule for disease outcome in patients with recent-onset undifferentiated arthritis: how to guide individual treatment decisions. **Arthritis & Rheumatism**, v.56, n.2, p.433-440. 2007.

VYBORNÝ, C. J. , GIGER, M. L. Computer vision and artificial intelligence in mammography. **AJR Am J Roentgenol**, v.162, n.3, p.699-708. 1994.

WALL, L., CHRISTIANSEN, T. , ORWANT, J. **Programação Perl**. Rio de Janeiro. Campus. 2001.

WITTEN, I. H. , FRANK, E. **Data Mining: Practical machine learning tools and techniques**. San Francisco. Morgan Kaufmann. 2005.

YAO, X. **Evolving Artificial Neural Networks**. IEEE Congress on Evolutionary Computation. 1999.

ANEXOS

7. ANEXOS

ANEXO A - Ficha de Coleta de Dados

Nome do Paciente: _____

Data do Preenchimento: _____ Registro no HC: _____

Sexo: Masculino Feminino Idade: _____ Etnia: _____

Data da última consulta registrada no prontuário: _____

Parâmetros clínicos

Tempo de doença (em anos): _____ Data do diagnóstico: _____

Tempo entre o início da doença e o início do tratamento (em anos): _____

Número de critérios da ACR: _____ Número de articulações acometidas: _____

Edema persistente de articulações interfalangeanas proximais: sim não

Tabagismo: sim não Presença de manifestações extra-articulares: sim não

Questionário de qualidade de vida (HAQ): _____

Escore de atividade de doença (DAS 28): _____

Parâmetros laboratoriais

Elevação das provas de atividade inflamatória nos três primeiros meses da doença ou no início do tratamento com DMARDs: sim não

Fator reumatóide > 100 UI/ml: sim não

Anticorpos anti-CCP positivo: sim não

Presença do epítipo compartilhado: sim não

Erosões radiológicas nos primeiros 12 meses de doença: sim não

Determinação da gravidade de doença: _____

ANEXO B - Critérios de Classificação da Artrite Reumatóide

Critérios estabelecidos pelo Colégio Americano de Reumatologia para o diagnóstico de Artrite Reumatóide (Arnett et al., 1988).

Para fins de classificação, pelo menos, quatro dos sete critérios devem estar presentes. Critérios de 1 a 4 devem estar presentes por pelo menos 6 semanas.

1. Rigidez matinal: rigidez matinal nas articulações, com duração de no mínimo uma hora até a máxima melhora.
2. Artrite em três ou mais áreas articulares: no mínimo três áreas articulares simultaneamente edemaciadas observadas por um médico. As 14 possíveis áreas são: interfalangeanas proximais, metacarpofalangeanas, punhos, cotovelos, joelhos, tornozelos e metatarsofalangeanas.
3. Artrite nas mãos: no mínimo uma área edemaciada nas articulações interfalangeanas proximais, metacarpofalangeanas ou punhos, observada por um médico.
4. Artrite simétrica: envolvimento simultâneo das mesmas articulações (definidas no critério número 2) em ambos os lados do corpo. Envolvimento das articulações interfalangeanas proximais e metacarpofalangeanas é aceitável sem simetria absoluta.
5. Nódulos reumatóides: nódulos subcutâneos em proeminências ósseas, superfícies extensoras, ou em regiões periarticulares, observados pelo médico.
6. Fator reumatóide: demonstração de valor anormal do fator reumatóide sérico por um método que o resultado seja positivo no máximo em 5% dos controles saudáveis.
7. Alterações radiográficas: alterações radiográficas típicas da artrite reumatóide no Raio-X de mãos e punhos na incidência pósterio-anterior, a qual inclui erosões ou osteopenia periarticular inequívoca.

ANEXO C - Classes Funcionais da Artrite Reumatóide

Critérios de classificação das classes funcionais da Artrite Reumatóide, estabelecidos pelo Colégio Americano de Reumatologia (Hochberg et al., 1992).

Classe Funcional I: apresenta capacidade completa para realizar atividades cotidianas, como atividades profissionais, higiene pessoal e lazer.

Classe Funcional II: capaz de realizar atividades profissionais e higiene pessoal, mas apresenta limitações para realizar atividades de lazer e esporte.

Classe Funcional III: apresenta limitações para atividades cotidianas e profissionais, exceto para higiene pessoal e alimentação.

Classe Funcional IV: Paciente apresenta limitações para todas as atividades, cotidianas, profissionais e higiene pessoal.

ANEXO D – Índice de Atividade da Doença (DAS 28)

O Índice de Atividade da Doença DAS 28 (*Disease Activity Score*) é um índice validado em estudos clínicos usado para medir atividade da doença em pacientes portadores de Artrite Reumatóide.

Para o cálculo do DAS 28 são usados os seguintes parâmetros clínicos: número de articulações dolorosas, número de articulações edemaciadas, o valor da velocidade de hemossedimentação (VHS) e uma avaliação global de saúde por escala visual analógica (EVA). As articulações avaliadas são: ombros, cotovelos, punhos, metacarpofalangeanas, interfalangeanas proximais e joelhos. Com estes dados, é realizado o seguinte cálculo:

$$DAS28 = (0,56 * \sqrt{n \text{ articulações dolorosas}}) + (0,28 * \sqrt{n \text{ articulações edemaciadas}}) \\ + (0,78 * \ln VHS) + (0,14 * EVA)$$

Equação 1: Cálculo do índice de atividade de doença DAS 28.

O valor do DAS 28 acima de 5,1 indica alta atividade de doença, enquanto o valor do DAS 28 menor que 3,2 indica baixa atividade de doença. O valor do DAS 28 entre 5,1 e 3,2 indica atividade da doença moderada. E o valor do DAS 28 menor que 2,6 é considerado remissão clínica.

ANEXO E - Parecer do Comitê de Ética em Pesquisa



HOSPITAL DAS CLÍNICAS DA FACULDADE DE MEDICINA
DE RIBEIRÃO PRETO DA UNIVERSIDADE DE SÃO PAULO

www.hcrp.fmrp.usp.br



Ribeirão Preto, 26 de setembro de 2007

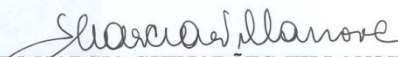
Ofício nº 3347/2007
CEP/MGV

Prezada Professora,

O trabalho intitulado **“AUXÍLIO AO DIAGNÓSTICO DE ARTRITE REUMATÓIDE ATRAVÉS DE TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL”**, foi analisado pelo Comitê de Ética em Pesquisa, em sua 255ª Reunião Ordinária realizada em 24/09/2007, e enquadrado na categoria: **APROVADO**, bem como o **Termo de Consentimento Livre e Esclarecido**, de acordo com o Processo HCRP nº 4880/2007.

Lembramos que devem ser apresentados a este CEP, o Relatório Parcial e o Relatório Final da pesquisa.

Atenciosamente.


DR^a MARCIA GUIMARÃES VILLANOVA
Vice-Coordenadora do Comitê de Ética em
Pesquisa do HCRP e da FMRP-USP

Ilustríssima Senhora
PROF^a DR^a SILVANA GIULIATTI
GABRIELA FÉLIX DOS SANTOS (Aluna)
Depto. de Genética – FMRP-USP

Comitê de Ética em Pesquisa HC e FMRP-USP - Campus Universitário

FWA – 0000 2733; IRB – 0000 2186

Fone (16) 3602-2228 - E-mail : cep@hcrp.fmrp.usp.br

Monte Alegre 14048-900 Ribeirão Preto SP

Livros Grátis

(<http://www.livrosgratis.com.br>)

Milhares de Livros para Download:

[Baixar livros de Administração](#)

[Baixar livros de Agronomia](#)

[Baixar livros de Arquitetura](#)

[Baixar livros de Artes](#)

[Baixar livros de Astronomia](#)

[Baixar livros de Biologia Geral](#)

[Baixar livros de Ciência da Computação](#)

[Baixar livros de Ciência da Informação](#)

[Baixar livros de Ciência Política](#)

[Baixar livros de Ciências da Saúde](#)

[Baixar livros de Comunicação](#)

[Baixar livros do Conselho Nacional de Educação - CNE](#)

[Baixar livros de Defesa civil](#)

[Baixar livros de Direito](#)

[Baixar livros de Direitos humanos](#)

[Baixar livros de Economia](#)

[Baixar livros de Economia Doméstica](#)

[Baixar livros de Educação](#)

[Baixar livros de Educação - Trânsito](#)

[Baixar livros de Educação Física](#)

[Baixar livros de Engenharia Aeroespacial](#)

[Baixar livros de Farmácia](#)

[Baixar livros de Filosofia](#)

[Baixar livros de Física](#)

[Baixar livros de Geociências](#)

[Baixar livros de Geografia](#)

[Baixar livros de História](#)

[Baixar livros de Línguas](#)

[Baixar livros de Literatura](#)
[Baixar livros de Literatura de Cordel](#)
[Baixar livros de Literatura Infantil](#)
[Baixar livros de Matemática](#)
[Baixar livros de Medicina](#)
[Baixar livros de Medicina Veterinária](#)
[Baixar livros de Meio Ambiente](#)
[Baixar livros de Meteorologia](#)
[Baixar Monografias e TCC](#)
[Baixar livros Multidisciplinar](#)
[Baixar livros de Música](#)
[Baixar livros de Psicologia](#)
[Baixar livros de Química](#)
[Baixar livros de Saúde Coletiva](#)
[Baixar livros de Serviço Social](#)
[Baixar livros de Sociologia](#)
[Baixar livros de Teologia](#)
[Baixar livros de Trabalho](#)
[Baixar livros de Turismo](#)