Data Mining Applied to Diagnose Diseases Caused by Lymphotropic Virus: a Performance Analysis

F. D. S. Farias, L. V. D. Souza, R. C. M. Sousa, C. A. M. Caldas, and L. F. Gomes, J. C. W. A. Costa

***Abstract*— This paper proposes a new methodology to diagnose the rheumatology manifestations and HTLV-I-Associated Myelopathy/Tropical Spastic Paraparesis, or HAM/TSP, in patients who have Lymphotropic virus of T cells in Humans or HTLV of type I and II. Computational intelligence algorithms are used to classify HTLV patient carriers with or without the presence of rheumatology manifestations and of HAM / TSP. A benchmarking is performed among artificial neural intelligence, naïve bayes, Bayesian networks and decision tree to evaluate the most suitable technique for solving this application issue. The obtained results demonstrate the potential of the methodology on the helping non-specialist doctors to classify the patient with the disease suspicion.**

***Keywords*— Computational Intelligence, Data-Mining, Neural Networks.**

I. INTRODUÇÃO

DIAGNÓSTICO é a metodologia para identificação através do processo de eliminação da natureza de algo. Em áreas do conhecimento como: medicina, engenharia, negócios, entre outras regras de diagnósticos podem ser empregadas. Na medicina, diagnóstico é "O reconhecimento de uma doença ou identificação pelos sinais aparentes e sintomas" ou "A análise da causa psicológica ou biomédica de uma doença ou condição" [1]. Diagnósticos médicos são considerados uma arte, independente de todos os esforços por padrões matemáticos já efetuados. No entanto, pesquisas estão sendo encorajadas pelo avanço na tecnologia de ferramentas computacionais, as quais permitem o desenvolvimento de softwares e técnicas mais robustas que podem vir a ajudar médicos de uma maneira mais precisa em tomadas decisões, ou seja, sem a necessidade direta da consulta a um especialista de determinada área especifica da medicina.

Mineração de dados é a metodologia tecnológica que tem sido mais adotada nas áreas de ciências biomédicas e pesquisa. Mineração de dados é um emergente campo de elevada importância para fornecimento de prognósticos e de classificação de doenças. Para uma aplicação eficiente da mineração de dados são aplicados os conceitos da Extração de Conhecimento em base de dados ou mais conhecido como *Knowledge Discovery in Databases* (KDD), responsável em realizar etapas prévias e posteriores a aplicação da classificação na etapa de mineração.

Atualmente, o desenvolvimento de classificadores para o diagnóstico de doenças tem sido bastante popularizado; na literatura é possível encontrar trabalhos relacionados ao auxílio da classificação de doenças.

Em [2], os autores tentaram identificar padrões específicos do Vírus da Imunodeficiência Humana, ou "*Human Immunodeficiency Vírus*" (HIV), causador da Síndrome da Imunodeficiência Adquirida, mais conhecida como AIDS; e do vírus da Hepatite C (HCV), através do uso de técnicas de aprendizagem de máquina. As simulações resultaram em 92,5% de acerto para a classificação do HIV. Para o HCV, a taxa de acerto foi de 96,25%. Em [3], os autores aplicaram Redes Neurais Artificiais (RNA) e aprendizagem de máquina, sendo proposto um novo modelo classificador de doenças para o mal de Alzheimer e mal de Parkinson, considerando os fatores de risco de maior influência. Os principais fatores de risco citados pelos autores são: idade, genes, diabetes, fumo e acidente vascular cerebral. Em [4], os autores utilizam a rede neural modular para diagnosticar o câncer de pulmão. Com o intuito de melhorar o acerto do diagnóstico, é proposta uma metodologia com quatro módulos, onde cada módulo trabalha com metade dos atributos treinados e testados através de dois modelos: *backpropagation* e função de base radial. A regra probabilística da soma é usada para realizar a integração entre as duas técnicas. Depois do treinamento, o acerto foi 95,75% no treinamento dos dados e 98,22% nos dados testados. O resultado foi determinado experimentalmente para ser melhor do que o alcançado pela rede neural monolítica. Em [5], é proposta a substituição do método tradicional de registro de sintomas e doenças em prontuários de pranchetas não digitais pelo registro dos sintomas de doenças em prontuários digitais. O propósito principal é o desenvolvimento de uma ferramenta computacional para diagnóstico, baseada em registros armazenados digitalmente pela experiência dos médicos. O trabalho discute o potencial da ferramenta de autoclassificação, baseada nas experiências passadas e reajustada a fim de serem acrescentados novos registros. Em [6], o algoritmo Bayesiano foi proposto com o intuito de inferir casos da doença de Alzheimer. O uso do algoritmo de inteligência computacional é motivado com o objetivo de acelerar a identificação da doença para início imediato do

F. D. S. Farias, Universidade Federal do Pará (UFPA), Belém, Pará, Brasil, fabriciosf@ufpa.br
L. V. D. Souza, Universidade Federal do Pará (UFPA), Belém, Pará, Brasil, lvsouza@ufpa.br
R. C. M. Sousa, Universidade Federal do Pará (UFPA), Belém, Pará, Brasil, ritasousa@iec.gov.br
C. A. M. Caldas, Universidade Federal do Pará (UFPA), Belém, Pará, Brasil, cezar_caldas@yahoo.com.br
L. F. Gomes, Universidade Federal do Pará (UFPA), Belém, Pará, Brasil, letícia.gomes@ics.ufpa.br
João C. W. A. Costa, Universidade Federal do Pará (UFPA), Belém, Pará, Brasil, jweyl@ufpa.br

tratamento, provendo uma qualidade de vida melhor para os pacientes e seus familiares. O foco principal do trabalho é desenvolver um modelo de vários critérios para auxiliar nas tomadas de decisão para o diagnóstico da doença. O autor conclui que este algoritmo deveria ser aplicado no diagnóstico da doença de Alzheimer devido a sua exatidão.

Neste artigo é apresentada uma comparação entre classificadores para auxiliar no diagnóstico de duas doenças que ainda não se beneficiam da utilização de algoritmos de inteligência computacional como suporte ao diagnóstico. É proposto um novo modelo de classificação da presença ou ausência da manifestação de doença reumatológica e da manifestação da Paraparesia Espástica Tropical/ Mielopatia associada ao HTLV (PET/MAH), ambas as doenças causadas pela presença do HTLV no individuo infectado. Foi realizado um estudo através da consideração dos pacientes que possuem HTLV e a presença ou ausência das doenças em estudo, além disso, são utilizados os sintomas característicos coletados a partir de um levantamento analítico realizado pelos especialistas da área médica. Ao contrário dos estudos citados, este artigo compara algoritmos clássicos para a classificação de duas doenças e demonstram através de um estudo comparativo dos desempenhos das taxas de acertos e erros, quais algoritmos obtiveram os melhores resultados para o problema proposto. Quatro algoritmos foram empregados e simulados trinta vezes cada um: árvore de decisão, redes bayesianas, naive bayes e a rede neural *backpropagation*.

Este artigo está estruturado da seguinte forma. A seção II apresenta a magnitude, em âmbito global e nacional, das doenças analisadas, assim como a forma atual de diagnóstico das mesmas. A seção III apresenta os passos básicos executados pelo KDD. Na seção IV, os resultados numéricos das simulações com os quatro algoritmos são analisados e discutidos. Finalmente, conclusões e propostas para trabalhos futuros são apresentadas na última seção.

II. Impacto da doença e o formulário de diagnóstico

Estima-se que até 20 milhões de pessoas no mundo possam estar infectadas com HTLV. Dentre as áreas consideradas de alto risco de infecção do vírus HTLV-1, destaca-se o sul do Japão, América do Sul, Ilhas Melanésia, África e Caribe. No entanto, para o HTLV-2 observou-se a predominância em populações indígenas [7]. Na Oceania, verificou-se ocorrência de infecções em Papua Nova Guiné e Austrália em grupos de doadores de sangue. A infecção na América do Sul vem sendo observada em todos os países, com diferentes taxas de prevalência [8]. No Brasil, o vírus foi identificado pela primeira vez, em 1986, em Campo Grande (MS). No entanto, foi em Salvador onde se identificou o maior risco do vírus, apesar de ser encontrado com grande abrangência em diversas regiões geográficas brasileiras. Acredita-se que pelo menos 2,5 milhões de pessoas estão infectadas no país. Na região norte, destaca-se o Amazonas, Amapá e Pará como os estados com o maior número de incidentes na região [9].

Na década de 90 surgiram os primeiros estudos propondo associações entre o HTLV e doenças inflamatórias articulares crônicas, através da avaliação de pacientes portadores do vírus em áreas endêmicas. Verificou-se que a prevalência de doenças reumatológicas, como a Artrite Reumatóide (AR) e síndrome de Sjögren, eram maiores em pacientes soropositivos do que naqueles considerados soronegativos [10]

Atualmente, o primeiro diagnóstico médico é realizado através do preenchimento de um formulário e de uma análise crítica dos sintomas apresentados pelo paciente, a qual está relacionada ao conhecimento médico dos casos anteriores. A partir deste passo, o paciente é enviado para exames laboratoriais para comprovação da suspeita.

A utilização de um algoritmo classificador é uma nova abordagem para auxiliar na análise manual corrente, com a utilização de uma ferramenta computacional, o médico possui um sistema especialista para auxiliar na identificação da suspeita. A maior contribuição está relacionada aos médicos não especialistas do domínio, onde com auxilio do sistema podem identificar a suspeita das doenças, encaminhando o paciente para exames de confirmação, mesmo não possuindo um conhecimento profundo da área específica.

III. Aplicação do KDD

KDD consiste em um conjunto de técnicas baseadas no tratamento da base de dados, e na retirada de conhecimento através da aplicação de uma técnica de inteligência computacional ou estatística no processo de mineração de dados, também conhecido como *data mining*.

Os dados analisados neste trabalho são compostos por 105 pacientes recrutados, os quais estão infectados com o vírus linfotrópico de células T humanas do tipo 1 e 2 (HTLV), podendo manifestar doença reumatológica e Paraparesia Espastica Tropical/ Mielopatia (PET/MAH). Durante as coletas dos dados foram realizadas 105 consultas, e foram aplicados 105 formulários elaborados pelo pesquisador Prof. Dr. Cezar Augusto Muniz Caldas. Durante a fase de aplicação dos formulários foram coletados 94 atributos (informações) por paciente. Após as etapas iniciais do KDD, apenas 25 atributos foram usados como entrada nos quatro algoritmos simulados. A relevância de cada atributo foi estudada nas etapas do KDD [11].

O processo de seleção dos dados relevantes ao modelo classificador foi iniciado através de reuniões com os pesquisadores da área médica, os quais coletaram os dados durante as consultas. Durante essa fase foi criada uma tabela com os principais atributos que de acordo com o especialista, seriam importantes ao se classificar a presença ou ausência das doenças estudadas. Em seguida, os parâmetros selecionados através das reuniões sofreram uma analise estatística afim de verificar a variância de acordo com a presença ou a ausência da doença.

O software WEKA (*Waikato Environment for Knowledge Analysis*) foi usado para simulações envolvendo inteligência artificial. O WEKA [12] provê um ambiente automático de classificação, clusterização, regressão e seleção de características, auxiliando pesquisadores a encontrarem soluções em diferentes ramos que possibilitam a utilização da mineração de dados. O software MATLAB é usado para simulações matemáticas envolvendo ou não inteligência

computacional. Os algoritmos naive bayes, redes bayeanas, e árvore de decisão J48 foram utilizados a partir do WEKA e a rede neural composta pelo algoritmo *backpropagation* foi desenvolvida na linguagem MATLAB.

IV. Resultados

Nesta seção os melhores resultados de cada algoritmo durante as simulações são apresentados e discutidos. Os resultados de acerto em porcentagem, erro médio absoluto e o número de registros classificados incorretamente de todos os algoritmos utilizados são apresentados. Em seguida, o melhor resultado é apresentado detalhadamente. A Fig. 1 mostra a melhor taxa de acerto alcançada por cada um dos algoritmos simulados em uma estala de 0 a 100% durante a tarefa de classificação.

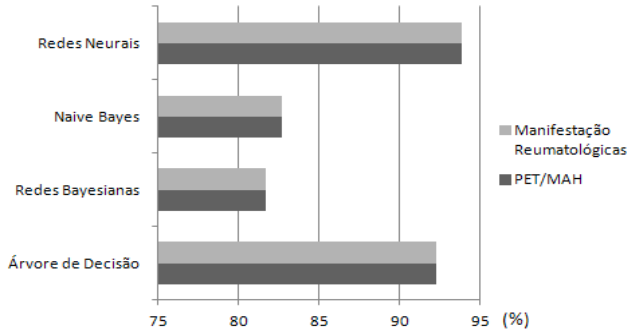


Figura 1. Percentual de acerto para a melhor simulação de cada algoritmo.

A Fig. 1 ilustra que a maior taxa de acerto alcançada foi com a Rede Neural Artificial, atingindo 93,75% dos casos apresentados. Naive bayes e redes bayesianas obtiveram uma taxa de acerto abaixo de 85%; enquanto a árvore de decisão obteve uma taxa de 92,30%. Observando o Erro Médio Absoluto (EMA) naive bayes e redes bayesianas obtiveram uma alta taxa comparando com o EMA da RNA e da árvore de decisão. A Fig. 2 apresenta os EMAs de cada algoritmo.

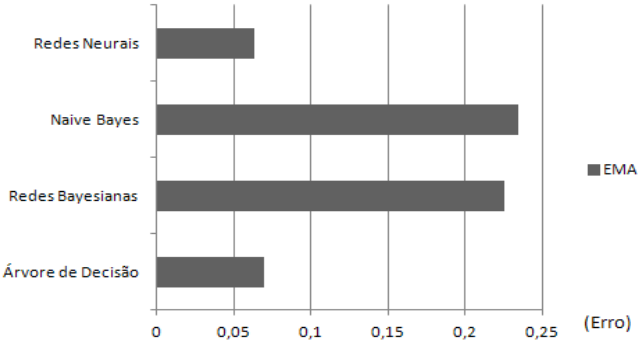


Figura 2. Erro médio absoluto obtido durante a melhor simulação de cada algoritmo.

A última comparação é mostrada no gráfico das classificações incorretas, onde os números dos registros classificados incorretamente são apresentados. A Fig. 3 apresenta o número de registros classificados incorretamente em cada algoritmo.

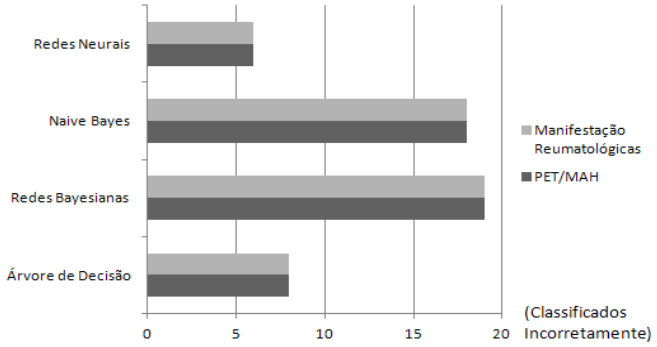


Figura 3. Manifestações reumatológicas e PET/MAH classificados incorretamente.

Classificando somente 12 registros incorretamente, sendo 6 de manifestações e 6 de PET/MAH, a RNA alcança novamente o melhor resultado. Árvore de decisão obteve um bom resultado, errando somente na classificação de 16 registros. Naive bayes classificou incorretamente 36 registros e a redes bayesianas classificou incorretamente 38 registros obtendo o pior resultado dentre os quatro. Em todos os quatros algoritmos simulados, a taxa de classificação incorreta foi de 50% para manifestações reumatólogicas e 50% para PET/MAH.

Dos quatro algoritmos comparados, destaca-se a RNA, a qual obteve uma taxa de erro médio absoluto de 0,0613. O segundo melhor algoritmo analisado foi a árvore de decisão com uma taxa erro médio absoluto de 0,07, seguido da rede bayesiana com 0,2254 de taxa de erro e naive bayes com 0,2347. Para a doença em estudo os algoritmos *backpropagation* e árvore J48 demonstram melhor exatidão para a classificação dos pacientes do que os outros algoritmos comparados. Para este conjunto de dados existe uma maior viabilidade para o uso da RNA devido ao seu maior grau de acerto e baixo grau de erro.

Os resultados demonstram a eficiência da utilização da inteligência computacional para solucionar este tipo de problema proposto. Mesmo com uma base de dados contendo poucos registros, os resultados obtiveram uma alta taxa de acerto.

A Rede Neural Artificial produziu os melhores resultados em todas as comparações obtendo o menor número de falsos positivos. Após 30 simulações, a melhor RNA possui duas camadas: camada escondida e camada de saída. A camada escondida foi configurada com 10 neurônios, enquanto a camada de saída foi configurada com dois. Os dois neurônios representados na camada de saída representam as duas possíveis classes do problema: Manifestações reumatológicas e PET/MAH.

A função de ativação da melhor RNA foi a *sigmoid*. A melhor RNA executou em 45 épocas um desempenho de 0,0613 e obteve seu melhor resultado de validação na época 39 alcançando o erro de 0,000143. A configuração da rede mais adequada contou com uma divisão da base de dados realizada de forma aleatória em 75%, 10% e 15% para treinamento, validação e teste respectivamente. É possível observar que na fase de validação o Erro Médio Quadrático

(EMQ) foi menor do que o na fase de testes, sendo que durante a etapa de treinamento, foi encontrada a maior taxa desse mesmo erro. A Tabela I ilustra o número de amostras e suas divisões, o erro médio quadrático para cada treino, validação, teste e o erro global em porcentagem para cada etapa.

TABELA I
RESULTADOS DA RNA EM FORMA NUMÉRICA.

***	Amostras	EMQ	%E
Treino	78	6,41135e-2	6,41025e-0
Validação	11	1,45298e-4	0
Teste	16	6,28712e-2	6,25000e-0

Para a simulação, foram separadas 78 amostras da base de dados para treinamento, 11 para validação e 16 para o conjunto de teste. A Tabela I apresenta que os EMQ de treino e teste foram maiores em relação ao erro de validação. Além disso, é possível observar a porcentagem de erro, em que, para o treinamento, foi obtida uma precisão de 93,58975% e um erro de aproximadamente 6,41025%; a validação alcançou uma precisão de 100% e um erro de 0% e finalmente a etapa de teste obteve uma precisão de 93,75% e uma taxa de erro de 6,25%.

A Fig. 4 mostra o comportamento do Erro Médio Quadrático para treino, validação e teste em todas as épocas percorridas pelo algoritmo. Como apresentado, a melhor época foi a 39, onde o algoritmo configurou os pesos conforme necessitado para satisfazer as três etapas. É possível observar que o erro de treinamento e teste se mantém sem muitas modificações e o erro de validação aumenta novamente após a melhor época.

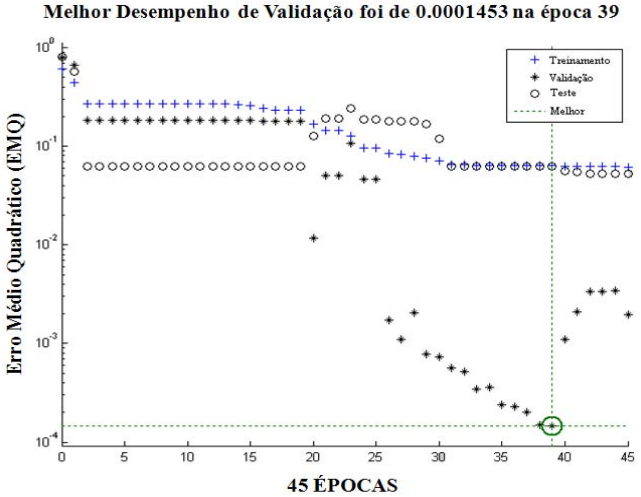


Figura 4. Representação gráfica dos resultados obtidos para as amostras.

V. CONCLUSÃO

Este artigo mostrou a aplicação do classificador baseado nos algoritmos de inteligência computacional para auxiliar no diagnóstico de Manifestações Reumatológicas e PET/MAH em pacientes com HTLV tipo I e II. Os resultados obtidos demonstram a eficiência do modelo classificador e o potencial da futura ferramenta classificadora para ajudar as decisões de diagnósticos de médicos não especialistas, para assim classificar o paciente de acordo com sua suspeita, e encaminhá-lo para exames laboratoriais para a comprovação da suspeita.

A rede neural apresentou os melhores resultados, onde, após as etapas de treinamento, validação e teste alcançando uma exatidão de 93,75% nestes casos. Comparando com os sistemas de diagnósticos desenvolvidos para classificação de outras doenças, conclui-se que os resultados são satisfatórios para o primeiro estudo aplicado em uma pequena base de dados.

Para trabalhos futuros, objetiva-se aumentar o número de pacientes da base de dados com pacientes com e sem HTLV e reduzir a taxa de erro da classificação através da tentativa da utilização de outros algoritmos. Este artigo aplicou as etapas do KDD para a solução de um problema de classificação que atualmente possuía sua solução analítica. O classificador proposto é capaz de auxiliar médicos não especialistas a classificar o paciente com a possível doença e encaminhar-lo para exames laboratoriais para comprovação da suspeita.

REFERÊNCIAS

[1] Gerard Wolff, J., "Medical diagnosis as pattern recognition in a framework of information compression by multiple alignment, unification and search", Decision Support Systems, Vol.42, Issue 2, pp. 608 - 625, 2006.

[2] Ajit, N., Xikun, W., Yang,Z.R."Mining viral protease data to extract cleavage knowledge". Bioinformatics, Vol. 18 Suppl. 1 2002, Pages S5-S13. Received on January 19,2002; revised and accepted on March 28, 2002.

[3] Sandhya, J., Deepa, S., Vibhudendra Simha, G.G."Classification of Alzheimer's Disease and Parkinson's Disease by Using Machine Learning and Neural Network Methods". 2010 Second International Conference on Machine Learning and Computing.

[4] Vazirani, H., Kala, R., Shuka, A., Tiwari, R. "Diagnosis of Breast Cancer by Modular Neural Network". 978-1-4244-5540-9/10/ ©2010 IEEE.

[5] Mane, K.K,. Börner K. "Computational Diagnostics: A Novel Approach to Viewing Medical Data". Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007).

[6] Pinheiro, P. R., de Castro, A.K.A, Pinheiro, M.C.D. "A Multicriteria Model Applied in the Diagnosis of Alzheimer's Disease: A Bayesian Network". 2008 11th IEEE International Conference on Computational Science and Engineering.

[7] Romanelli, L.C.F; Caramelli, P.; Proietti, "A.B.F.C. O vírus linfotrópico de células T humanos tipo 1 (HTLV-1): quando suspeitar da infecção?" .Revista Associação Medica Brasileira, v.56, n.3, p. 340-347, nov-dez, 2010.

[8] Cruz, B.A.; Catalan-Soares, B.; Proietti, F. "Manifestações reumáticas associadas ao vírus lifotrópico humano de células T do tipo I (HTLV-I)". Revista Brasileira de Reumatologia, v.45, n.2, p.71-77, mar-abr.2005.

[9] Colin, D.D. et al. "Prevalência da infecção pelo vírus linfotrópico humano de células Te fatores de risco associados à soropositividade em doadores de sangue da cidade de Rio Branco, AC, Brasil (1998-2001)". Revista da sociedade brasileira de medicina tropical, v.36, n.6, p.677-683, nov-dez, 2003.

[10] Carvalho M.M.N. et al. "Doenças reumátivasauto-imunes em indivíduos infectados pelo HTLV-1". Revista brasileira de reumatologia", v.46, n.5, p.334-339, set-out.2006.

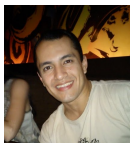
[11] Fayyad, U. Shapiro, G. P. and Smyth, P. "From Data Mining to Knowledge Discovery in Databases". AI MAGAZINE. American Association for Artificial Intelligence, 1996.

[12] I. H. Witten and E. Frank. "Data mining - practical machine learning tools and techniques". St. Louis, Elsevier, 2005, pp. 34.

Fabricio de S. Farias nasceu no Pará/Brasil em 10 de março de 1988. Ele recebeu seu grau de bacharel em Engenharia da Computação pela Universidade Federal do Pará (UFPA) em 2010. Desde 2011 ele é pesquisador no convênio Ericsson-UFPA e membro do LEA (Laboratório de Eletromagnetismo Aplicado) da UFPA. Suas áreas de pesquisa incluem data minig, redes DSL e engenharia de software.

Lamartine V. de Souza recebeu seu grau de bacharel em Engenharia Elétrica em 1999 e seu grau de mestre em Engenharia Elétrica em 2001, ambos pela UFPA. Ele é professor do curso de Engenharia Industrial da UFPA desde 2009, pesquisador no convênio Ericsson-UFPA desde 2005, membro do LEA desde 1995 e GEPAI (Grupo de Estudos e Pesquisas em Aplicações Industriais) da UFPA desde 2009. Suas áreas de pesquisa incluem Processos de Markov, redes multimídia, redes DSL, redes de acesso em telecomunicações.

Rita Catarina Medeiros Sousa recebeu seu grau de bacharel em Medicina em 1993 pela UFPA. Realizou sua residência média em Doenças Infecciosas no Hospital Universitário Barros Barreto da UFPA em 1995, mestrado em Virologia Médica em 1998 e doutorado em Virologia em 2002, ambos no Instituto Pasteur na Universidade de Paris. Ela é professora da UFPA e pesquisadora do Instituto Evandro Chagas. Possui experiência em microbiologia com ênfase em virologia, atuando no seguintes tópicos: gripe, viroses respiratórias e HTLV.



Cezar Augusto Muniz Caldas recebeu seu grau de bacharel em Medicina em 2002 pela UFPA. Realizou sua residência média em Medicina Interna no Hospital Universitário Barros Barreto da UFPA em 2005 e Reumatologia em 2008 pela USP. Recebeu seu grau de doutor em Ciências Médicas em 2010 pela FMUSP.

Letícia Figueiredo Gomes nasceu no Pará/Brasil em 20 de novembro de 1988. Ela é uma estudante de graduação de Medicina pela UFPA. Possui bolsa de pesquisa pelo Centro de Medicina Tropical (CMT) desde 2009, onde está desenvolvendo projetos na área de doenças infecciosas com ênfase em manifestações clínicas por infecção por HTLV.

João Crisóstomo Weyl Albuquerque Costa nasceu no Pará/Brasil em 27 de janeiro de 1959. Ele recebeu seu grau de bacharel em Engenharia Eletricista - Eletrônico em 1981, pós-graduação em Geofísica em 1983, ambos pela UFPA. Recebeu o seu grau de mestre em 1989 pela PUC-RJ e o grau de doutor em Engenharia Elétrica em 1994 pela Unicamp-SP. Desde 1994, ele é professor da Faculdade de Engenharia Elétrica e Computação da UFPA. Atualmente, suas áreas de pesquisa são: modelagem de dispositivos e sistemas ópticos, incluindo redes de acesso. Ele é pesquisador do CNPq desde 1994.