Ferramenta para diagnóstico de diabetes através de Perceptron Multicamadas e Lógica Fuzzy

Matheus Oliveira
Universidade Tecnológica Federal do Paraná
Dois Vizinhos - PR, Brasil
mttvittorelli@gmail.com

Marcio Silva
Universidade Tecnológica Federal do Paraná
Dois Vizinhos - PR, Brasil
marcio.silva@ifpb.edu.br

Thiago Amaral
Universidade Tecnológica Federal do Paraná
Dois Vizinhos - PR, Brasil
thiago.magalhaes@univasf.edu.br

Wagner Schmitz
Universidade Tecnológica Federal do Paraná
Dois Vizinhos - PR, Brasil
wagnerms@Outlook.com

***Resumo* — O diabetes é uma doença crônica e multifatorial que afeta um grande número de pessoas. A detecção precoce reduz a morbidade e mortalidade e diversas ferramentas de inteligência artificial têm sido aplicadas para análises de impactos e efeitos da doença, das quais se destacam o Perceptron Multicamadas (MLP) e a Lógica Fuzzy. Apesar dos diversos estudos citados, é importante frisar que existe um gap na literatura sobre a aplicação de trabalhos para analisar o diabetes na ótica do machine learning e da lógica fuzzy, o que faz esse Projeto Integrador ser pertinente. Dessa, forma o objetivo deste trabalho é aplicar o Perceptron Multicamadas e Lógica Fuzzy para analisar um dataset público com dados reais de pacientes diabéticos a fim de melhorar o diagnóstico da doença. Este trabalho chegou a uma acurácia de 73% usando o MLP. No caso da Lógica Fuzzy, a xyz. Assim, sugerimos para trabalhos futuros uma ampliação nas amostras para se aumentar a taxa de acurácia e robustez do modelo. Além disso, recomenda-se fazer o uso de outras arquiteturas de machine learning a fim de comparar as acurácias. Espera-se que este modelo melhore a precisão diagnóstica e contribua para o avanço da prática clínica, aumentando a taxa de cura do diabetes.**

Palavras-chave — MLP, Lógica Fuzzy, Diabetes, Saúde

I. INTRODUÇÃO

O diabetes é uma doença crônica e multifatorial que afeta um grande número de pessoas, sendo considerado um dos problemas de saúde mais desafiadores de se controlar. É caracterizado por um distúrbio metabólico que resulta na deficiência total ou parcial de insulina produzida pelo pâncreas, bem como na diminuição da sua eficácia nos tecidos do corpo. Essa condição compromete o metabolismo de lipídeos, glicídios, proteínas, água, vitaminas e minerais [1]. A diabetes pode levar a complicações graves, como problemas cardiovasculares, neuropatia, retinopatia e doença renal. É fundamental para os pacientes diabéticos adotarem um estilo de vida saudável, que inclua uma dieta equilibrada, exercícios físicos regulares e o uso adequado de medicamentos para controlar a glicemia. Além disso, a educação sobre a doença e o monitoramento regular são essenciais para um bom gerenciamento do diabetes e prevenção de suas complicações.

O diabetes mellitus tipo 2 é uma epidemia global e representa cerca de 90% de todos os casos de diabetes [2, 4]. Em 2010, estima-se que 285 milhões de pessoas com mais de 20 anos de idade viviam com diabetes em todo o mundo, e esse número pode chegar a 439 milhões até 2030 [3, 4]. É preocupante observar que aproximadamente metade dos indivíduos com diabetes desconhecem que possuem a doença, evidenciando a necessidade de maior conscientização e rastreamento da população [4]. Esforços devem ser direcionados para a prevenção, diagnóstico precoce e tratamento adequado do diabetes tipo 2, a fim de reduzir o impacto dessa condição na saúde pública. A detecção precoce reduz a morbidade e mortalidade [6].

Dessa forma, considerando a necessidade de identificar e prevenir a diabetes o mais cedo possível, diversas ferramentas de inteligência artificial têm sido aplicadas para o entendimento dos efeitos da doença, das quais se destaca o Perceptron Multicamadas (MLP) e a Lógica Fuzzy. Por exemplo, o trabalho de [5] que realizaram uma comparação abrangente de vários algoritmos de classificação para a detecção precoce do diabetes. Eles coletaram e pré-processaram um conjunto de dados de registros de pacientes, contendo informações pessoais, problemas médicos associados e medicamentos. O conjunto de dados foi dividido em conjuntos de treinamento e teste, e usado para treinar e avaliar vários algoritmos de classificação populares. Os resultados do estudo revelaram que os algoritmos de Perceptron Multicamadas (MLP), Gradient Boost Machine (GBM) e Random Forest (RF) tiveram o melhor desempenho geral, seguidos de perto pelas Máquinas de Vetores de Suporte (SVM). Essas descobertas demonstram o potencial desses algoritmos para uso na detecção precoce do diabetes e sugerem que mais pesquisas são necessárias para aprimorar e otimizar esses modelos para uso clínico.

Outro importante trabalho, foi realizado por [6]. Esses autores propuseram um modelo de agrupamento de características baseado em cluster não supervisionado para identificação precoce do diabetes, utilizando um conjunto de dados de código aberto contendo informações de 520 pacientes diabéticos. No conjunto de dados baseado em cluster e no conjunto de dados completo, a máxima precisão (acurácia) é de 99,57% e 99,03%, respectivamente. A melhor precisão, recall, erro quadrado médio mínimo (MSE), erro quadrado médio máximo (MSE) e pontuação F1 de 1,000 são obtidos a partir do perceptron multicamadas (MLP), Random Forest (RF) e k-Nearest Neighbors (KNN), 0,984 a partir do random forest (RF)

e support vector machine (SVM), 0,010 a partir do RF, 0,067 a partir do KNN e 99,20% a partir do RF, respectivamente.

Outro trabalho recente na área, foi realizado por [7]. Esses autores usaram o datased PIMA Indian diabetes da Universidade da Califórnia/Irvine (UCI) para fins experimentais. O estudo foi realizado em três estágios: (1) uma técnica de correlação foi desenvolvida para seleção de características; (2) a técnica AdaBoost foi implementada nas características selecionadas para classificação; e (3) uma técnica de empilhamento inovadora com perceptron multicamadas, máquina de vetores de suporte e regressão logística (MLP, SVM e LR, respectivamente) foi projetada e desenvolvida para as características selecionadas. A técnica de empilhamento proposta integrou os modelos inteligentes e resultou em uma melhoria no desempenho do modelo, superando assim o problema de gerar múltiplas decisões por meio do AdaBoost. A técnica de empilhamento proposta superou outros modelos em comparação com o AdaBoost em termos de métricas de desempenho.

Apesar dos diversos estudos citados, é importante frisar que existe um gap na literatura sobre a aplicação de trabalhos para analisar o diabetes na ótica do machine learning e da lógica fuzzy, o que faz esse Projeto Integrador ser pertinente. Dessa, forma o objetivo deste trabalho é aplicar o Perceptron Multicamadas e Lógica Fuzzy para analisar um dataset público com dados reais de pacientes diabéticos a fim de melhorar o diagnóstico da doença.

A próxima seção irá mostra o método proposto, e em seguida serão exibidos os resultados e discussões deste trabalho. Por último, serão delineadas as conclusões.

II. MÉTODO

As linguagens utilizadas neste Projeto Integrador foi o Python e R para a codificação das arquiteturas do MLP e e do sistema Fuzzy e a biblioteca usadas foi pandas e o scikit-learn. A biblioteca Pandas é uma biblioteca de código aberto para manipulação e análise de dados em Python. Ela fornece estruturas de dados flexíveis e eficientes, como o DataFrame, que permite realizar operações sofisticadas de limpeza, transformação e exploração de dados. O Pandas é amplamente utilizado em tarefas de ciência de dados e análise de dados, pois oferece recursos poderosos para lidar com conjuntos de dados estruturados, o que é o caso de nosso dataset.

O scikit-learn oferece uma ampla gama de algoritmos e ferramentas para tarefas de aprendizado supervisionado e não supervisionado, como classificação, regressão, agrupamento e seleção de características. O scikit-learn é projetado para ser fácil de usar e possui uma sintaxe consistente que facilita o desenvolvimento e a avaliação de modelos de aprendizado de máquina. O Google Colab foi usado para realização das simulações com o MLP. Ele fornece tempos de execução do Python 2 e 3 pré-configurados com as bibliotecas essenciais de aprendizado de máquina e inteligência artificial, como o scikit-learn.

A. Base de Dados

A base de dados foi obtida no Kaggle. Este conjunto de dados foi criado pelo National Institute of Diabetes and Digestive and Kidney Diseases. O objetivo do conjunto de dados é prever, de forma diagnóstica, se um paciente tem ou não diabetes, com base em determinadas medições diagnósticas incluídas no conjunto de dados. Várias restrições foram aplicadas à seleção dessas instâncias em um banco de dados maior. Em particular, todas as pacientes são mulheres com pelo menos 21 anos de idade e de origem indígena Pima.

O dataset escolhido possui 9 colunas (Gravidez, Glicose, Pressão Sanguínea, Espessura da Pele, Insulina, IMC, Função de Pedigree do Diabetes, Idade e Resultado. O dataset possui 768 dados de pacientes. Os dados foram divididos em 70% para o treinamento e 30% para o teste.

B. MLP e Lógica Fuzzy

Os perceptrons de múltiplas camadas (MLP's) são um tipo de redes neurais artificiais (RNA) utilizadas em aprendizagem de máquina para viabilizar o reconhecimento de padrões em bases de dados, auxiliando às tomadas de decisões em diversas áreas de conhecimento. Sua popularidade se dá por permitir resolver problemas complexos, utilizando-se de algoritmos que possibilitam aprender a partir dos erros gerados na propagação do sinal de entrada dos dados (vetor de entrada) ao longo das camadas que compõem a estrutura da rede neural artificial (fig.1). A aprendizagem da rede acontece a partir da interação dos neurônios cujo processamento ocorre através da propagação dos sinais mediante funções matemáticas conhecidas como funções de ativação que apresentam seus pesos sinápticos.

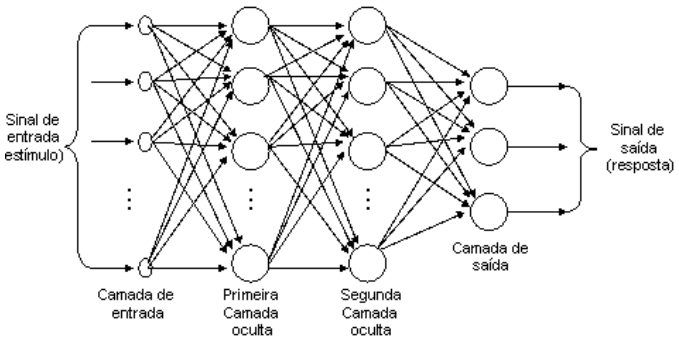

Fig. 1 – Arquitetura de uma rede neural artificial do tipo MLP com duas camadas ocultas. Fonte: [14].

O MLP traz ainda a capacidade de retornar esses sinais de volta a sua origem, promovendo a chamada "Aprendizagem por retropropagação" (*error back-propagation*), que consiste em fazer os sinais percorrerem por essa mesma estrutura realizando ajuste dos pesos sinápticos e corrigindo os erros identificados. Essa dinâmica de aprendizagem possibilita aprender com os erros, de maneira que a resposta real da rede se mova para mais perto da resposta desejada.

Para treinar o dataset adotado, o MLP foi estruturado considerando o método de otimização Adam, do inglês, *Adaptive Moment Estimation* cuja contribuição consiste em ajustar os pesos da rede neural com base nas estimativas

adaptativas de primeira ordem (gradiente) e segunda ordem (momento), permitindo que o algoritmo se adapte de forma eficiente a diferentes taxas de aprendizado e superfícies de erro não uniformes.

Considerou também para a ativação do sinal nas camadas ocultas a função ReLU por ser computacionalmente mais eficiente visto que envolve apenas uma operação de comparação representada pela relação f(x) = max(0, x), isto é, se o valor de x for negativo, a função ReLU retorna zero; caso contrário, retorna o próprio valor de x. Assim, tendo sido definidos os parâmetros do MLP, a base de treino rodou com 1000 iterações em busca de padrões e correções de erros.

Outro modelo também adotado no trabalho considera os princípios da lógica Fuzzy, outro ramo de estudos da ciência de dados, expressa por um conjunto de variáveis linguísticas, que possibilita desenvolver análises mais qualitativas trazendo aspectos comuns do cotidiano relativo às incertezas e imprecisões do mundo real e, com isso, aproximando-se da forma como o ser humano raciocina. Para [11], a força da lógica difusa provém de sua habilidade de extrair conclusões e gerar respostas baseadas em informações vagas, ambíguas, qualitativas, incompletas ou imprecisas.

Trata-se de uma extensão da teoria tradicional dos conjuntos, sendo que, em vez de um raciocínio preciso, a lógica difusa é sobre o raciocínio aproximado, isto é, traz a ideia de que o elemento não precisa pertencer a um conjunto por completo, mas poderá apresentar um grau de pertinência variando num intervalo de valores.

De outra forma, significa dizer que em vez de atribuir uma classificação binária a um objeto ou uma proposição, a lógica Fuzzy permite que um objeto tenha graus de pertinência a diferentes categorias ou que uma proposição seja avaliada como parcialmente verdadeira e parcialmente falsa.

Nesses termos, um conjunto fuzzy pode ser representado por proposições (afirmações) imprecisas compostas por variáveis e valores linguísticos que carecem de um grau de pertinência para que possa indicar o quão verdadeira são tais afirmações. Desse modo, o formato geral de uma proposição fuzzy apresenta o seguinte padrão: “Se <antecedente> Então <consequente>”.

Esse sistema se processa em 3 principais etapas: Fuzzificação, Inferência e Defuzificação (fig. 2). Os parâmetros de entrada do sistema são mapeados em variáveis linguísticas que são utilizadas na definição de regras para o processamento de variáveis de saída (também representadas por variáveis linguísticas). Após essa etapa de inferência, faz-se necessário transformar o valor de saída em dado numérico, procedimento este conhecido como Defuzificação.

Em vista do dataset adotado no presente trabalho, definiu-se três variáveis linguísticas, sendo duas de entrada – Glicose (Glucose) e Pressão Arterial (BloodPressure) – e uma variável de saída – Diabetes. O propósito desse modelo utilizando a lógica Fuzzy é o de prever a probabilidade de um paciente apresentar tendência a ter diabetes a partir da análise das taxas de glicose e pressão arterial, considerando a definição de regras de inferência que consideram as funções de pertinência para cada uma das variáveis.

D. Parâmetros dos testes (MLP e Fuzzy)

Para o caso do modelo implementado utilizando o algoritmo MLP, foram considerados os seguintes parâmetros:

I. Classificador: MLP
II. Camadas ocultas: 6
III. Função de ativação nas camadas escondidas: Rectified Linear Units (ReLU)
IV. Solver: ADAM
V. Alpha = 0.009
VI. Máximo de Interações: 1000

Já para a aplicação utilizando o sistema de inferência Fuzzy foram adotados os seguintes parâmetros:

I. Método de inferência: Mamdani
II. Variáveis e valores linguísticos:
   - Input: Glucose (Baixo, Médio e Alto) e BloodPressure (Baixo, Médio e Alto)
   - Output: Diabetes (Sem Diabetes e Com Diabetes)
III. Funções de Pertinência e Domínio do conjunto Fuzzy: Glucose (Triangular [0, 200]), BloodPressure (Trapezoidal [0, 150]) e Diabetes (Trapezoidal [0, 100])

Ressalta-se que o ideal seria utilizar os oito atributos completos da base de dados para realizar o processo de defuzificação, entretanto, devido à complexidade técnica na área da saúde, foi desenvolvido um exemplo utilizando apenas dois deles – glicose e pressão sanguínea – para simular o processo que seria realizado com todos os dados e cálculos.

E. Métricas de avaliação

No modelo que considerou o algoritmo MLP, foram utilizadas as métricas oriundas da matriz de confusão para o conjunto de teste, a saber:

I. Precision – relação entre o TP (número de verdadeiros positivos) e o FP (número de falsos positivos).
II. Recall – relação entre o TP (número de verdadeiros positivos) e FN (número de falsos negativos).
III. F1-score – O F1-score pode ser interpretado como uma média ponderada do precision e do recall, em que um f1- score alcança seu melhor valor em 1 e o pior escore em 0 [10].

Para o modelo considerando o sistema Fuzzy, buscou-se definir as funções de pertinência e regras de inferências considerando as variáveis linguísticas descritas anteriormente com o propósito de identificar a probabilidade de um paciente apresentar, ou não, quadro de diabetes.

III. RESULTADOS E DISCUSSÕES

Os resultados mostraram que o MLP foi capaz de aprender rapidamente a distinguir o diagnóstico do diabetes. A taxa de erro convergiu rapidamente gerando uma acurácia de aproximadamente 74%.

Também foi possível observar a partir das análises trazidas pela Matriz de Confusão da Figura 2, o número de registros considerados Verdadeiros Positivos (121); os Verdadeiros Negativos (49); os Falsos Positivos (30) e os Falsos Negativos (31). A Tabela 1 mostra as métricas obtidas: Precision, Recall e F1-Score. A precisão para a identificação do paciente normal foi maior quando comparada à identificação do paciente com diabetes (80% e 62%, respectivamente). Já o recall mostrou um valor maior para as imagens normais (80%) quando comparadas ao paciente com diabetes (61%). Por último, o f1-score mostra justamente que o classificador foi melhor para distinguir o paciente normal, pois o resultado foi mais próximo a 1.

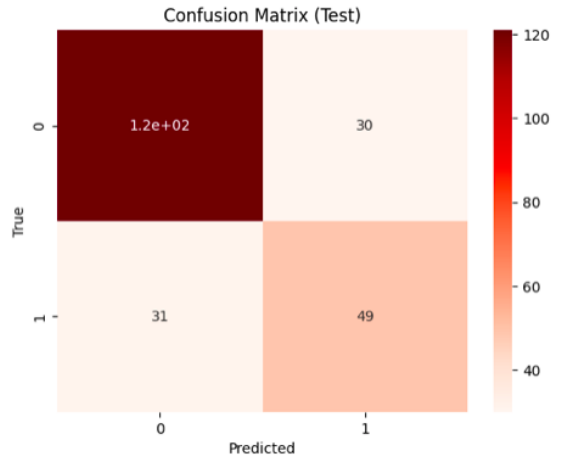

Fig. 2: Matriz de Confusão para o conjunto de teste

TABELA 1:Métricas para o conjunto de teste (Precision, Recall e F1-Score)

	Precision	**Recall**	**F1-Score**
Normal (0)	0.80	0.80	0.80
Diabetes (1)	0.62	0.61	0.62
Acurácia			0.74
Macro Avg	0.71	0.71	0.71
Weighted Avg	0.74	0.74	0.74

Os resultados demonstram que o modelo criado usando o MLP pode ser usado para ajudar no diagnóstico do diabetes. Os resultados encontrados na nossa pesquisa estão um pouco abaixo daqueles encontrados por [6] que chegaram a uma acurácia próxima de 99%, demonstrando a necessidade de testar outros algoritmos para identificar melhoria no desempenho da previsão.

No tocante aos resultados alcançados com o sistema de inferências Fuzzy foi possível gerar as funções de pertinência e regras de inferência considerando as variáveis linguísticas Glicose, Pressão Arterial e Diabetes, sendo as duas primeiras variáveis de entrada e a terceira variável de saída. E na etapa de defuzificação apresenta-se a probabilidade para um dado paciente apresentar quadro clínico de diabetes. As representações gráficas dos conjuntos das variáveis linguísticas podem ser observadas nas figuras 3 (a, b e c). Também foi possível gerar a curva resultante do processo de defuzificação por meio da

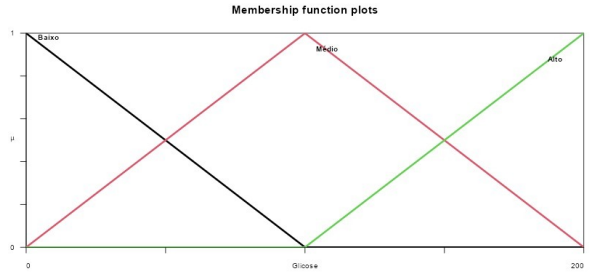

Fig. 3a – Variável linguística Glicose

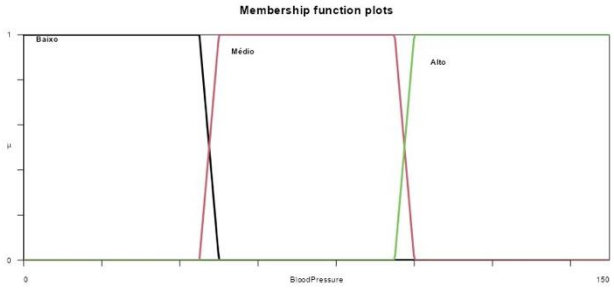

Fig. 3b – Variável linguística Pressão Arterial

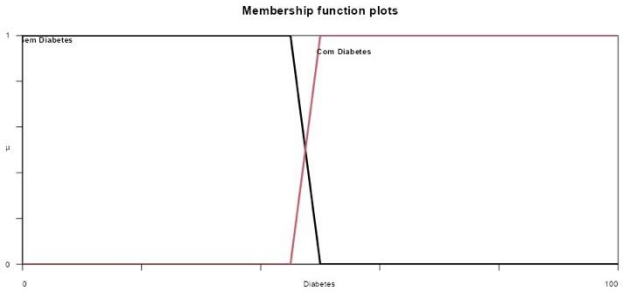

Fig. 3c – Variável linguística Diabetes

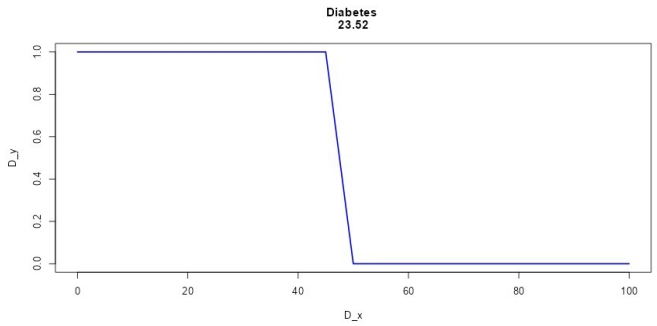

Fig. 4 – Defuzificação por implicação pelo minimo (conectivo AND)

IV. CONCLUSÕES

Este trabalho desenvolveu um modelo de MLP capaz de prever o diagnóstico do diabetes. Para o modelo de classificação proposto, o algoritmo demostrou ter uma acurácia próxima a 74%. Também se utilizou a teoria dos sistemas Fuzzy sendo possível gerar as funções de pertinência e regras de inferência de onde saiu a probabilidade de um paciente apresentar ou não o quadro clínico para diabetes a partir da análise para as variáveis de glicose e pressão arterial. Uma limitação do estudo, foi não utilizar as duas técnicas de modo unificado num modelo do tipo neuro Fuzzy. Outro fator limitante foi considerar apenas 3 variáveis linguísticas na composição do sistema Fuzzy. Assim, sugerimos para trabalhos futuros uma ampliação nas amostras para se aumentar a taxa de acurácia e robustez do modelo MLP. Além disso, recomenda-se fazer o uso de outras arquiteturas de machine learning a fim de comparar as acurácias. Espera-se que este modelo melhore a precisão diagnóstica e contribua para o avanço da prática clínica, aumentando a taxa de cura do diabetes.

REFERÊNCIAS

[1] K. P. Fonseca, C. D. A. Rached. Complicações do diabetes mellitus. International Journal of Health Management. 2019.

[2] International Diabetes Federation. IDF diabetes atlas. 6th Ed. Brussels: International Diabetes Federation; 2013.

[3] JE Shaw, RA Sicree, PZ Zimmet. Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes Res Clin Pract 2010; 87:4-14.

[4] A. F. Costa, L. S. Campos, A. F. Oliveira, et al. Carga do diabetes mellitus tipo 2 no Brasil. Cadernos de Saúde Pública. 2017. doi: 10.1590/0102-311X00197915

[5] C. Carpinteiro, J. Lopes, A. Abelha, M. F. Santos. A Comparative Study of Classification Algorithms for Early Detection of Diabetes. Procedia Computer Science. Vol. 220, 2023, Pages 868-873.

[6] M. M Hassan, S. Mollick, F. Yasmin, An unsupervised cluster-based feature grouping model for early diabetes detection. Healthcare Analytics. Volume 2, November 2022, 100112. DOI: https://doi.org/10.1016/j.health.2022.100112

[7] S. K. Kalagotla, S. V. Gangashetty, K. Giridhar. A novel stacking technique for prediction of diabetes. Computers in Biology and Medicine. Volume 135, August 2021, 104554. DOI: https://doi.org/10.1016/j.compbiomed.2021.104554

[8] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.

[9] R. E. V. Silva, Um estudo comparativo entre redes neurais convolucionais para a classificação de imagens. 2018. 51 f. TCC (Graduação em Sistemas de Informação) Universidade Federal do Ceará, Campus de Quixadá, Quixadá, 2018.

[10] L. Silva, L. Araújo, V. Souza, A. Santos, R. Neto, Redes neurais convolucionais aplicadas na detecção de pneumonia através de imagens de raio-x. v. 11 n. 1 (2020): Computer on The Beach 2020. DOI: https://doi.org/10.14210/cotb.v11n1.p419-426

[11] L. R. B. Maciel, I. S. M. Oliveira, A. H. M. Oliveira, G. M. Botelho. Previsão de Mortalidade de Câncer de Mama usando Redes Perceptron de Múltiplas Camadas e Lógica Fuzzy. XVII Encoinfo Congresso de Computação e Sistemas de Informação, 2015.

[12] S. Haykin, Neural Networks and Learning Machines, 3a edição. Prentice Hall, 2008.