



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO RIO GRANDE (FURG)
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL
CURSO DE DOUTORADO EM MODELAGEM COMPUTACIONAL

Tese de Doutorado

Uma rede social para registro e monitoramento de dados de saúde baseada em modelos de aprendizado de máquina e lógica fuzzy

Alice Tissot Garcia Pintanel

Orientadora:	Prof[a]. Dr[a] Graçaliz Pereira Dimuro
Co-orientadores:	Prof. Dr. Eduardo Nunes Borges
	Prof. Dr. Julian Moises Sejje Suarez
Colaboradora:	Dr[a] Camila Rose Guadalupe Barcelos

Rio Grande, 2026





Tese de Doutorado

Uma rede social para registro e monitoramento de dados de saúde baseada em modelos de aprendizado de máquina e lógica fuzzy

Alice Tissot Garcia Pintanel

Banca examinadora:

Profª. Drª. Graçaliz Pereira Dimuro (orientadora)

Prof. Dr. Eduardo Nunes Borges (coorientador)

Prof. Dr. Julian Moises Sejje Suarez (coorientador)

Profª. Drª. Viviane Leite Dias de Mattos

Profª. Drª. Heloísa de Arruda Camargo

Prof. Dr. Giancarlo Lucca

Dr. Vinicius Carrera Souza

P659r

Pintanel, Alice Tissot Garcia
Uma rede social para registro e monitoramento de dados de saúde baseada em modelos de aprendizado de máquina lógica fuzzy / Alice Tissot Garcia Pintanel. - 2026.
231f.
Tese (Doutorado) – Universidade Federal do Rio Grande – Programa de Pós-Graduação em Modelagem Computacional, 2026.

Orientadora: Dra. Graçaliz Pereira Dimuro
Co-orientadores: Dr. Eduardo Nunes Borges
Dr. Julian Moises Sejje Suarez
Colaboradora: Dra. Camila Rose Guadalupe Barcelos

1. Modelagem Computacional. 2. Exames de sangue. 3. Extração de informação de PDF. 4. Classificação laboratorial. 5. Apoio à decisão clínica. 6. Lógica fuzzy. 7. ANFIS. 8.TSK. I. Dimuro, Graçaliz Pereira. II. Borges, Eduardo Nunes. III. Suarez, Julian Moises Sejje. IV.Barcelos, Camila Rose Guadalupe. V. Título.

CDU 004.414.23

Catalogação na fonte: Bibliotecária Brenda Sequeira CRB10/1979

ATA DE DEFESA DE TESE DE DOUTORADO

No dia 29 do mês de abril do ano de dois mil e vinte e seis, por meio híbrido (videoconferência e presencial), realizou-se o exame de defesa de tese da doutoranda **Alice Tissot Garcia Pintanel** do Programa de Pós-Graduação em Modelagem Computacional, que apresentou e defendeu a tese **"Uma rede social para registro e monitoramento de dados de saúde baseada em modelos de aprendizado de máquina e lógica *fuzzy*"** como requisito para a conclusão do Doutorado em Modelagem Computacional. A sessão foi aberta às nove horas pela Profa. Dra. Graçaliz Pereira Dimuro, orientadora e presidente da Comissão de Avaliação. A referida Comissão esteve integrada, além da presidente e orientadora, pelo Prof. Dr. Eduardo Nunes Borges (coorientador), pelo Prof. Dr. Julian Moises Sejje Suarez (coorientador), pela Dra. Camila Rose Guadalupe Barcelos (colaboradora), pela Profa. Dra. Viviane Leite Dias de Mattos (FURG), pela Profa. Dra. Heloísa de Arruma Camargo (UFSCAR), pelo Prof. Dr. Giancarlo Lucca (FURG) e pelo Dr. Vinicius Carrera Souza. Após a apresentação do trabalho, a banca arguiu o candidato e, a seguir, deliberou pela

( ) aprovação da Tese.

(x) aprovação da Tese, sugerindo modificações no texto.

( ) reprovação da Tese.

ASSINADO DIGITALMENTE
EDUARDO NUNES BORGES
CPF
000.570.350-67
A conformidade com a assinatura pode ser verificada em:
https://serpro.gov.br/assinador-digital
Serpro

Profa. Dra. Graçaliz Pereira Dimuro
Orientadora

Prof. Dr. Eduardo Nunes Borges
Coorientador

Documento assinado digitalmente
gov.br
CAMILA ROSE GUADALUPE BARCELOS
Data: 11/05/2026 10:57:46-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Julian Moises Sejje Suarez
Coorientador

Dra. Camila Rose Guadalupe Barcelos
Colaboradora

Documento assinado digitalmente
VIVIANE LEITE DIAS DE MATTOS
Data: 04/05/2026 18:23:11-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Viviane Leite Dias de Mattos
Membro Interno

Documento assinado digitalmente
GIANCARLO LUCCA
Data: 05/05/2026 15:55:16-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Giancarlo Lucca
Membro Interno

Documento assinado digitalmente
HELOISA DE ARRUDA CAMARGO
Data: 06/05/2026 10:41:00-0300
Verifique em https://validar.iti.gov.br

Prof. Dra. Heloísa de Arruda Camargo
Membro Externo

Documento assinado digitalmente
VINICIUS CARRERA SOUZA
Data: 12/05/2026 10:10:08-0300
Verifique em https://validar.iti.gov.br

Dr. Vinicius Carrera Souza
Membro Externo

Dedico esta tese ao Lucas, meu companheiro incansável em todas as horas; ao Théo, que trouxe luz e renovou minhas forças no momento mais intenso desta jornada; e aos meus pais, cuja dedicação e amor incondicional foram meu alicerce em cada passo do caminho.

AGRADECIMENTOS

A conclusão desta tese de doutorado não teria sido possível sem o apoio, a colaboração e o incentivo de muitas pessoas que estiveram ao meu lado ao longo desta jornada de quatro anos e meio. Foram inúmeros desafios, superados com o auxílio e a presença de pessoas especiais, às quais registro minha sincera gratidão.

Em primeiro lugar, agradeço aos meus orientadores, Dra. Graçaliz Pereira Dimuro e Dr. Eduardo Nunes Borges, pela orientação competente, pela paciência e pelo constante apoio em todas as etapas deste trabalho. Estendo meus agradecimentos à Dra. Camila Barcelos, pelo valioso suporte na área da saúde, e ao Dr. Julian Moises Sejje Suarez, pela contribuição fundamental no processo de modelagem ANFIS. A experiência e os conselhos de cada um de vocês foram essenciais para o direcionamento e a consolidação desta pesquisa.

Aos colegas do Projeto Dados48, deixo meu especial agradecimento não apenas pelas contribuições técnicas e científicas, mas também pelo companheirismo e amizade construídos ao longo do percurso. O apoio de vocês tornou esta caminhada mais leve e enriquecedora.

Sou imensamente grata à minha família. Ao meu pai, João, e à minha mãe, Marilia, pelo amor incondicional e pelo incentivo constante. Ao meu marido, Lucas, pela paciência, compreensão e apoio nos momentos mais difíceis. E ao meu filho, Théo, que chegou na fase final da tese, trazendo luz e renovando minhas forças quando eu mais precisei. Vocês foram meus alicerces, sustentando-me diante das adversidades e mostrando que seria possível conciliar todas as demandas.

Aos meus amigos e amigas, agradeço por cada palavra de incentivo, pela compreensão nas ausências e, principalmente, por trazerem leveza e alegria aos dias mais desafiadores.

Finalmente, agradeço à Universidade Federal do Rio Grande (FURG), instituição que me acolhe desde 2014 e que foi fundamental para minha formação acadêmica, pessoal e profissional. Registro também minha gratidão ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e demais órgãos de fomento, pelo suporte financeiro que possibilitou a realização desta pesquisa. Por fim, deixo meu agradecimento ao Hospital Universitário por terem aceito a proposta de projeto a fim de fornecer os dados necessários para o desenvolvimento deste trabalho.

A todos vocês, meu mais profundo obrigado. Sem o apoio de cada um, esta tese não teria sido possível.

“Todas as vitórias ocultam uma abdicação”.
— SIMONE DE BEAUVOIR

RESUMO

PINTANEL, Alice Tissot Garcia. **Uma rede social para registro e monitoramento de dados de saúde baseada em modelos de aprendizado de máquina e lógica fuzzy**. 2026. 234 f. . Universidade Federal do Rio Grande (FURG), Rio Grande.

Este estudo aborda o desenvolvimento de uma abordagem inteligente para apoio à decisão clínica voltada à classificação laboratorial e à estratificação de risco associadas ao Diabetes *Mellitus* Tipo 2. Diante do crescente impacto das doenças crônicas e das limitações dos métodos tradicionais baseados em limiares fixos, propõe-se a utilização de técnicas de aprendizado de máquina aliadas à lógica *fuzzy*, com o objetivo de lidar com as incertezas inerentes aos dados clínicos e fornecer resultados interpretáveis. A metodologia proposta está estruturada em três frentes principais. A primeira consiste na extração automática de informações relevantes a partir de exames laboratoriais em formato PDF, por meio de técnicas de aprendizado de máquina, demonstrando elevado desempenho em termos de precisão, *recall* e *F1-score*. A segunda envolve o desenvolvimento de um Sistema de Inferência Neuro-*Fuzzy* Adaptativo (ANFIS) baseado em TSK, organizado em dois níveis: um modelo de classificação laboratorial, fundamentado em dados de glicose e hemoglobina glicada, e um modelo de estratificação de risco, que incorpora variáveis adicionais como índice de massa corporal (IMC), sexo, raça e classificação clínica prévia. A terceira frente contempla o desenvolvimento de uma plataforma digital integradora, capaz de armazenar, processar e apresentar dados clínicos em formato de linha do tempo, promovendo a interação entre pacientes e profissionais de saúde. Os resultados obtidos demonstram que a abordagem proposta apresenta elevado desempenho na categorização de perfis glicêmicos, bem como resultados consistentes na estratificação de risco, mesmo em cenários com dados sintéticos e maior heterogeneidade. Destaca-se a capacidade do modelo *fuzzy* de representar o risco de forma contínua, proporcionando maior aderência à prática clínica. Além disso, a utilização de regras linguísticas garante interpretabilidade e transparência, características fundamentais em aplicações na área da saúde. Conclui-se que a metodologia desenvolvida é robusta, interpretável e adequada para integração em sistemas de apoio à decisão clínica, validando as hipóteses científicas e tecnológicas propostas. Como perspectivas futuras, destacam-se a aplicação da abordagem a outras patologias, a integração de mecanismos de alerta automático e a ampliação da plataforma para diferentes contextos clínicos.

Palavras-chave: Exames de sangue, extração de informação de PDF, classificação laboratorial, apoio à decisão clínica, lógica *fuzzy*, ANFIS, TSK.

ABSTRACT

PINTANEL, Alice Tissot Garcia. **A social network for recording and monitoring health data based on machine learning models and fuzzy logic**. 2026. 234 f. . Universidade Federal do Rio Grande (FURG), Rio Grande.

This study presents the development of an intelligent clinical decision support approach for laboratory classification and risk stratification related to Type 2 Diabetes Mellitus. Given the growing impact of chronic diseases and the limitations of traditional methods based on fixed thresholds, machine learning techniques combined with fuzzy logic are employed to address the uncertainties inherent in clinical data and provide interpretable results. The proposed methodology is structured into three main components. The first consists of the automatic extraction of relevant information from laboratory reports in PDF format using machine learning techniques, demonstrating high performance in terms of precision, recall, and F1-score. The second involves the development of a Takagi–Sugeno–Kang (TSK)-based Adaptive Neuro-Fuzzy Inference System (ANFIS), organized into two levels: a laboratory classification model based on glucose and glycated hemoglobin data, and a risk stratification model that incorporates additional variables such as body mass index (BMI), sex, race, and previous clinical classification. The third component comprises the development of an integrated digital platform capable of storing, processing, and presenting clinical data in a timeline format, thereby promoting interaction between patients and healthcare professionals. The results demonstrate that the proposed approach achieves high performance in glycemic profile categorization, as well as consistent outcomes in risk stratification, even in scenarios involving synthetic data and greater heterogeneity. The fuzzy model's ability to represent risk continuously is particularly noteworthy, providing closer alignment with clinical practice. Furthermore, the use of linguistic rules ensures interpretability and transparency, which are essential characteristics for healthcare applications. It is concluded that the developed methodology is robust, interpretable, and suitable for integration into clinical decision support systems, thereby validating the proposed scientific and technological hypotheses. Future work includes the application of the approach to other pathologies, the integration of automated alert mechanisms, and the expansion of the platform to different clinical contexts.

Keywords: Blood tests, information extraction, laboratory classification, clinical decision support, fuzzy logic, ANFIS, TSK.

LISTA DE FIGURAS

1 Diagrama de Fluxo de Processo 30

2 Pseudo-algoritmo retirado das diretrizes da SBD e utilizado como base para pré-classificação dos pacientes em diabéticos, pré-diabéticos e não diabéticos 44

3 Arquitetura do modelo MC-TSK-FS. 53

4 Fluxograma da metodologia da revisão teórica da Extração de Informação 57

5 Fluxograma da metodologia da revisão teórica dos sistemas *Fuzzy* . . 64

6 Nuvem de palavras montada a partir dos seis artigos selecionados . . 71

7 Arquitetura ANFIS 72

8 Etapas gerais da abordagem proposta 73

9 Modelo *wrapper* usando algoritmo de otimização seno-cosseno e método KNN. 74

10 Fluxo de trabalho do modelo proposto por [1]. 76

11 Modelo de estrutura proposto. 77

12 Fluxograma da metodologia da revisão teórica da metodologia ANFIS 80

13 Resumo geral referente ao total de artigos após todas as etapas 85

14 Gráfico que relaciona a quantidade de artigos pelo tipo de dado . . . 86

15 Gráfico que relaciona a quantidade de artigos pela disponibilidade ou não do código 86

16 Gráfico que relaciona a quantidade de artigos pelo tipo da doença . . 87

17 Gráfico que relaciona a quantidade de artigos pelos métodos utilizados em cada um deles 87

18 Organização metodológica do trabalho 96

19 Tela de Login ou Cadastro do Usuário 97

20 Principais elementos presentes no aplicativo 98

21 Estrutura da aplicação na nuvem 101

22 REST API 102

23 Tela de Login ou Cadastro do Usuário 106

24 Tela para selecionar o perfil que deseja criar 107

25 Telas de Criação de Perfil 109

26 Tela para selecionar o perfil que deseja ingressar no APP 110

27 Telas disponíveis para o perfil paciente 111

28 Telas disponíveis para o médico visualizar ou aceitar pacientes 112

29 Telas disponíveis para o médico visualizar as informações referentes à um determinado paciente 113

30 Fluxograma referente a metodologia para extração de informação dos exames de sangue 114
31 Fluxograma do passo a passo para detecção das variáveis de interesse 118
32 Etapas do processo 120
33 Exemplo de documento processado utilizando o modelo *layout* 121
34 Exemplo de localização das *Keys and Endpoint* 122
35 Exemplo parcial de extração de dados de exames de sangue a partir da utilização da ferramenta *Azure* da *Microsoft* 123
36 Exemplo de extração de algumas variáveis de interesse de vários exames de sangue 123
37 Exemplo de arquivo gerado a partir do processo de extração das variáveis de interesse de vários exames de sangue 124
38 Exemplo de *layout* do laboratório Gram 125
39 Exemplo de *layout* do laboratório Analisa 126
40 Exemplo de *layout* do laboratório Sérgio Franco 130
41 Texto extraído para o exemplo de *layout* do laboratório Gram 131
42 Texto extraído para o exemplo de *layout* do laboratório Analisa . . . 132
43 Texto extraído para o exemplo de *layout* do laboratório Sérgio Franco 133
44 Variáveis de interesse obtidas a partir da utilização da técnica *form understanding* para os três *layouts* 133
45 Variáveis de interesse obtidas a partir da utilização da técnica *form understanding* para os três *layouts* apresentadas no .csv gerado 134
46 Comparação do resultado obtido para o exemplo de *layout* do laboratório Gram 135
47 Comparação do resultado obtido para o exemplo de *layout* do laboratório Analisa 136
48 Comparação do resultado obtido para o exemplo de *layout* do laboratório Sérgio Franco 137

49 Diagrama representando a intersecção dos dados necessários para a construção do banco destinado à modelagem de classificação diagnóstica laboratorial 146
50 Exemplo do arquivo **data**, contendo os valores médios de glicose e hemoglobina glicada dos pacientes. 146
51 Exemplo do arquivo ***target***, contendo a classificação diagnóstica dos pacientes. 147
52 Caracterização da amostra do HU. São apresentadas as distribuições dos participantes quanto ao sexo, raça, faixa etária e classificação glicêmica. 148
53 Exemplo de trecho do arquivo **data** resultante da filtragem dos dados do simulador. 154
54 Exemplo de trecho do arquivo ***target*** resultante da filtragem dos dados do simulador. 155

55 Caracterização da amostra utilizada no estudo. São apresentadas as distribuições referentes ao sexo, raça, faixa etária, índice de massa corporal (IMC) e classificação glicêmica dos 1.352 indivíduos que compõem a base de dados. 156
56 Diagrama que resume as duas modelagens *fuzzy* para DM2 158
57 Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável glicose para a classificação normal 162
58 Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável glicose para a classificação intermediário . . . 163
59 Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável glicose para a classificação alto 165
60 Conjuntos *fuzzy* trapezoidais representando os termos linguísticos da variável glicose a partir das três classificações 166
61 Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável HbA1c para a classificação normal 167
62 Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável HbA1c para a classificação intermediário . . . 169
63 Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável HbA1c para a classificação alto 170
64 Conjuntos *fuzzy* representando os termos linguísticos da variável HbA1c a partir das três classificações 171
65 Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável IMC para a classificação normal 176
66 Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável IMC para a classificação intermediário 178
67 Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável IMC para a classificação alto 179
68 Conjuntos *fuzzy* representando os termos linguísticos da variável IMC a partir das três classificações 180
69 Resumo dos dados utilizados para cada um dos níveis 183
70 Fluxograma resumo sobre a análise dos resultados 183

LISTA DE TABELAS

1 Estrutura da matriz de confusão para classificação binária. 35

2 Combinações de palavras-chave utilizadas para revisão sistemática de extração de informação. 58
3 Total de artigos retornados para cada um dos passos descritos 60
4 Informações básicas dos artigos selecionados 62
5 Informações básicas dos artigos selecionados (continuação) 63
6 Combinações de palavras-chave utilizadas para revisão sistemática de sistemas *fuzzy* 65
7 Total de artigos retornados para cada um dos passos descritos 68
8 Total de artigos selecionados a partir da aplicação de todos os critérios de exclusão 69
9 Informações básicas dos artigos selecionados 70
10 Desempenho de cada algoritmo em relação as métricas propostas . . 76
11 Tabela de glicose no sangue 77
12 Tabela comparativa entre ANFIS e ANFIS-PSO a partir da métrica RMSE 79
13 Combinações de palavras-chave utilizadas para revisão sistemática de ANFIS 81
14 Total de artigos retornados para cada um dos passos descritos 84
15 Total de artigos selecionados a partir da aplicação de todos os critérios de exclusão 85
16 Comparação de estudos que aplicam técnicas de Inteligência Artificial para apoio à classificação, estratificação de risco ou diagnóstico de diabetes. 94

17 Resultados Quantitativos da Avaliação de Desempenho da API 105

18 Resultados obtidos considerando alguns exames de sangue 138
19 Relação de exames de sangue pertencentes ao Grupo 1 139
20 Relação de exames de sangue pertencentes ao Grupo 2 139
21 Resumo das métricas considerando cada um dos três testes realizados 139
22 Resultados obtidos para cada uma das variáveis de interesse considerando todos os exames de sangue 140
23 Resultados obtidos para cada uma das variáveis de interesse considerando todos os exames de sangue (continuação) 141
24 Resumo das métricas considerando as variáveis de interesse para cada um dos três testes realizados 142

25 Características da amostra utilizada no estudo - Dados do HU. 149
26 Caracterização da amostra segundo a classificação glicêmica - Dados HU. 150
27 Características da amostra utilizada no estudo - Dados sintéticos. . . . 155
28 Caracterização da amostra segundo a classificação glicêmica - Dados sintéticos. 157
29 Critérios laboratoriais para Diabetes *Mellitus* e pré-diabetes 158
30 Critérios adotados para o paciente quanto à classificação laboratorial de Diabetes *Mellitus*, pré-diabetes e não diabético 159
31 Critérios adotados para o paciente quanto à classificação laboratorial de Diabetes *Mellitus*, pré-diabetes e não diabético a partir da consideração do limite de hipoglicemia 160
32 Classificação do IMC segundo estado da arte 174
33 Critérios adotados para o IMC quanto ao risco metabólico atual associado ao DM2 174
34 Critérios adotados para as entradas binárias quanto ao risco metabólico atual associado ao DM2 175
35 Divisão dos dados do Nível 01 para treino e teste para cada uma das classes 185
36 Distribuição dos dados em cada uma das classes do Nível 02 185
37 Métricas Globais para análise do Nível 01 186
38 Métricas por Classe para análise do Nível 01 187
39 Comparação dos resultados obtidos no Nível 02 para o teste original e teste balanceado 187
40 Resultados obtidos para a análise por classe realizada para o teste balanceado 187
41 Matriz de Confusão para análise de desempenho do teste balanceado . 188

42 Informações básicas dos artigos selecionados 213
43 Informações básicas dos artigos selecionados (continuação) 214
44 Informações básicas dos artigos selecionados (continuação) 215
45 Informações básicas dos artigos selecionados (continuação) 216
46 Informações básicas dos artigos selecionados (continuação) 217
47 Informações básicas dos artigos selecionados (continuação) 218
48 Informações básicas dos artigos selecionados (continuação) 219
49 Informações básicas dos artigos selecionados (continuação) 220
50 Informações básicas dos artigos selecionados (continuação) 221
51 Informações básicas dos artigos selecionados (continuação) 222
52 Informações básicas dos artigos selecionados (continuação) 223
53 Informações básicas dos artigos selecionados (continuação) 224

54 Resultados obtidos considerando todos os exames de sangue 226
55 Resultados obtidos considerando todos os exames de sangue (continuação) 227
56 Resultados obtidos considerando todos os exames de sangue (continuação) 228
57 Resultados obtidos considerando todos os exames de sangue (continuação) 229

58 Resultados obtidos considerando todos os exames de sangue (continuação) 230
59 Resultados obtidos considerando todos os exames de sangue (continuação) 231
60 Resultados obtidos considerando todos os exames de sangue (continuação) 232
61 Resultados obtidos considerando todos os exames de sangue (continuação) 233
62 Resultados obtidos considerando todos os exames de sangue (continuação) 234

LISTA DE ABREVIATURAS E SIGLAS

AIDS	Acquired Immunodeficiency Syndrome
ALP	Fosfatase Alcalina
ANFIS	Adaptive Neuro-Fuzzy Inference System
API	Application Programming Interface
AUC	Area Under the Curve
AUCH	Area Under the Convex Hull
AWS	Amazon Web Services
BiLSTM	Bidirectional Long Short-Term Memory
CAAE	Certificado de Apresentação para Apreciação Ética
CE	Critério de Exclusão
CI	Critério de Inclusão
CPF	Cadastro de Pessoa Física
COVID-19	Coronavirus Disease 2019
CRM	Conselho Regional de Medicina
CRUD	Create, Read, Update and Delete
CSV	Comma-Separated Values
DM2	Diabetes Mellitus Tipo 2
DPF	Função de Hereditariedade
DPP	Diabetes Prevention Program
DT	Decision Tree
DWT	Discrete Wavelet Transform
FBS	Fasting Blood Sugar
FCM	Fuzzy C-Means
FDPS	Finish Diabetes Prevention Study
FINDRISC	Finnish Diabetes Risk Score
FLS	Fuzzy Logic System

FN	Falso Negativo
FP	Falso Positivo
FS-ANFIS	Feature Selection-based ANFIS
GNC	Graph Convolutional Networks
HbA1c	Hemoglobina Glicada
HDL	High-Density Lipoprotein
HIV	Human Immunodeficiency Virus
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
HU	Hospital Universitário
IA	Inteligência Artificial
ICEIS	International Conference on Enterprise Information Systems
IDE	Integrated Development Environment
IDF	International Diabetes federation
IMC	Índice de Massa Corporal
IoT	Internet of Things
JBHI	Journal of Biomedical and Health Informatics
JSON	JavaScript Object Notation
KNN	K-Nearest Neighbors
KRM	Knowledge Representation Module
LDL	Low-Density Lipoprotein
LOF	Local Outlier Factor
MC-TSK-FS	Multi-Class Takagi-Sugeno-Kang Fuzzy System
MLP	Multilayer Perceptron
MWC	Minimum Weighted Coefficient
MySQL	My Structured Query Language
NEIWDA	Natural Exponential Inertia Weight-Based Dragonfly Algorithm
NFIS	Neuro-Fuzzy Inference System
OCR	Optical Character Recognition
OP	Ordem de Pesquisa
PDF	Portable Document Format
PIDD	Pima Indians Diabetes Dataset
PMC	Perceptron Multicamadas
PPG	Fotopletismografia (Photoplethysmography)

PCR	Proteína C Reativa
PHP	Hypertext Preprocessor
PSA	Prostate-Specific Antigen
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
REST	Representational State Transfer
RMSE	Root Mean Square Error
RNA	Redes Neurais Artificiais
SBD	Sociedade Brasileira de Diabetes
SEL	Semantic Entity Labeling
SOM	Self-Organizing Maps
SQI	Signal Quality Index
SUS	Sistema Único de Saúde
SVM	Support Vector Machine
T4	Tiroxina
TGO	Transaminase Glutâmico-Oxalacética
TGP	Transaminase Glutâmico-Pirúvica
TN	Verdadeiro Negativo
TP	Verdadeiro Positivo
TPS	Transações por Segundo
TSH	Thyroid Stimulating Hormone
TSK	Takagi-Sugeno-Kang
TTGO	Teste de Tolerância à Glicose Oral
UF	Unidade Federativa
VLDL	Very Low-Density Lipoprotein
WEB	World Wide Web
XML	Extensible Markup Language

LISTA DE SÍMBOLOS

G	Valor da glicose plasmática do paciente (mg/dL)
$HbA1c$	Valor da hemoglobina glicada (%)
IMC	Índice de Massa Corporal (kg/m^2)
S	Sexo do paciente
R	Raça do paciente
C	Classe diagnóstica atribuída ao paciente
x	Variável de entrada genérica
y	Variável de saída do sistema fuzzy
n	Número total de amostras
m	Número de variáveis de entrada
r	Número de regras fuzzy
A_i	Conjunto fuzzy associado à variável de entrada i
B_i	Conjunto fuzzy associado à variável de saída i
$\mu_A(x)$	Grau de pertinência do elemento x ao conjunto fuzzy A
α_i	Grau de ativação da regra fuzzy i
w_i	Peso da regra fuzzy i
$\bar{w}_i$	Peso normalizado da regra fuzzy i
f_i	Saída da regra fuzzy i
z_i	Consequente da regra fuzzy i
a, b, c, d	Parâmetros da função de pertinência trapezoidal
θ	Vetor de parâmetros do modelo
ε	Erro associado à predição
$RMSE$	Raiz do Erro Quadrático Médio
Acc	Acurácia
$Prec$	Precisão
Rec	Recall (Sensibilidade)

F_1	Medida F1
AP	Average Precision
HL	Hamming Loss
TP	Número de Verdadeiros Positivos
TN	Número de Verdadeiros Negativos
FP	Número de Falsos Positivos
FN	Número de Falsos Negativos
μ	Média aritmética
σ	Desvio padrão
Min	Valor mínimo observado
Max	Valor máximo observado
Q_1	Primeiro quartil
Q_2	Mediana
Q_3	Terceiro quartil
$\mathcal{R}$	Conjunto de regras fuzzy do sistema
$\mathcal{X}$	Espaço de entrada do modelo
$\mathcal{Y}$	Espaço de saída do modelo
$P(\cdot)$	Probabilidade de ocorrência de um evento

SUMÁRIO

1 Introdução **26**
1.1 Contextualização e Motivação . . . 27
1.2 Hipótese Científica e Tecnológica . . . 28
1.3 Objetivos . . . 28
1.3.1 Objetivos Específicos . . . 29
1.4 Metodologia . . . 29
1.5 Projeto Dados48 . . . 29
1.6 Estrutura do Trabalho . . . 29

2 Fundamentação Teórica **31**
2.1 Conhecimentos Técnicos Gerais . . . 31
2.1.1 *Tokenização* . . . 31
2.1.2 *Form Understanding* . . . 31
2.1.3 *Optical Character Recognition* (OCR) . . . 32
2.1.4 Extração de pares chave–valor . . . 33
2.2 Métricas de avaliação . . . 34
2.2.1 Métricas para classificação . . . 34
2.2.2 Métricas para extração de informação . . . 36
2.2.3 Métricas para regressão . . . 37
2.2.4 Rede Social . . . 37
2.3 Conhecimentos associados à área da saúde . . . 39
2.3.1 Diabetes *Mellitus* . . . 39
2.3.2 Exames de Sangue . . . 44
2.4 Teoria de conjuntos *fuzzy*, lógica *fuzzy* e sistemas *fuzzy* . . . 47
2.4.1 Conjuntos *fuzzy* . . . 48
2.4.2 Funções de pertinência . . . 48
2.4.3 Sistemas de inferência *fuzzy* . . . 49
2.4.4 Modelos Mamdani e Takagi–Sugeno–Kang . . . 50
2.4.5 Funções de agregação . . . 50
2.4.6 T-normas . . . 51
2.4.7 Modelos caixa branca baseados em regras *fuzzy* . . . 52
2.4.8 Sistemas de Inferência Neuro-*Fuzzy* Adaptativos (ANFIS) . . . 52
2.4.9 Aprendizado dos parâmetros . . . 54
2.5 Resumo do capítulo . . . 55

3 Revisões Sistemáticas **56**
3.1 Revisão 01: Extração de Informação de PDFs . . . 56
3.1.1 Questões de Pesquisa . . . 56
3.1.2 Termos de Busca . . . 57
3.1.3 Bases utilizadas . . . 57
3.1.4 Critérios de Exclusão/Inclusão . . . 58
3.1.5 Selecionando Trabalhos . . . 59
3.1.6 Trabalhos Selecionados . . . 61
3.2 Revisão 02: Sistemas baseados em regras *Fuzzy* para predição e classificação . . . 64
3.2.1 Questões de Pesquisa . . . 64
3.2.2 Termos de Busca . . . 65
3.2.3 Bases utilizadas . . . 66
3.2.4 Critérios de Exclusão/Inclusão . . . 66
3.2.5 Selecionando Trabalhos . . . 67
3.2.6 Trabalhos Selecionados . . . 71
3.3 Revisão 03: Metodologia ANFIS para predição e/ou classificação de doenças . . . 80
3.3.1 Questões de Pesquisa . . . 80
3.3.2 Termos de Busca . . . 81
3.3.3 Bases utilizadas . . . 81
3.3.4 Critérios de Exclusão/Inclusão . . . 82
3.3.5 Filtrando Trabalhos . . . 82
3.3.6 Trabalhos Selecionados . . . 83

4 Metodologia Geral **95**

5 Aplicativo desenvolvido **97**
5.1 Visão geral da arquitetura da plataforma . . . 98
5.1.1 Aplicativo . . . 99
5.1.2 Banco de Dados . . . 100
5.2 Resultados associados ao desenvolvimento do aplicativo . . . 105
5.2.1 Análise do Desempenho da API . . . 105
5.2.2 Interfaces . . . 106

6 Extração de Informação **114**
6.1 Método para extração de informação de exames de sangue . . . 114
6.1.1 PDF . . . 114
6.1.2 Texto . . . 115
6.1.3 Variáveis . . . 116
6.1.4 Resultado . . . 118
6.2 Estudo de caso . . . 119
6.2.1 Utilização da ferramenta *Azure* . . . 119
6.2.2 Utilização da técnica *Form Understanding* . . . 121
6.2.3 Geração de arquivo .csv . . . 122
6.2.4 Exemplo de aplicação para 3 laboratórios diferentes . . . 122
6.3 Análise dos Resultados . . . 123
6.3.1 Análise de desempenho considerando cada um dos exames de sangue . . . 126
6.3.2 Análise de desempenho considerando cada uma das variáveis de interesse 127

6.3.3 Análise de Complexidade 128

7 Modelagem Neuro-*Fuzzy* para apoio à decisão clínica de DM2 **143**
7.1 Banco de Dados para o apoio à decisão clínica de DM2 143
7.1.1 Dados Reais: Hospital Universitário (HU) 144
7.1.2 Dados Sintéticos: Simulador *Sint Multiverse* 149
7.2 Modelagem Neuro Fuzzy em Dois Níveis 155
7.2.1 Nível 01: Implementação *Fuzzy* dos Critérios Laboratoriais de Classificação Glicêmica 157
7.2.2 Nível 02: Categorização fuzzy do perfil de risco metabólico associado ao DM2 172
7.3 Análise dos Resultados da Metodologia Neuro-*Fuzzy* em Dois Níveis . . 182
7.3.1 Protocolo de avaliação experimental 183

8 Conclusões e Trabalhos Futuros **189**
8.1 Limitações do Estudo 192
8.2 Trabalhos Futuros 193

Referências **195**

ANEXOS

A Tabelas com as informações básicas dos artigos selecionados na Revisão ANFIS **212**

B Tabelas com os resultados na íntegra para cada um dos 210 exames **225**

1 INTRODUÇÃO

A necessidade de desenvolver soluções voltadas ao monitoramento e à prevenção de doenças ganhou destaque ainda maior após a pandemia de COVID-19 (*Coronavirus Disease 2019*), iniciada em 2019, que evidenciou fragilidades nos processos de acompanhamento e gestão de pacientes. Embora esse período tenha impulsionado a criação de diversas metodologias que contribuíram para avanços expressivos no sistema de saúde, persistem lacunas relevantes que demandam investigação e aprimoramento contínuos [2].

Nesse contexto, destaca-se também a realidade hospitalar, caracterizada por rotinas intensas e emocionalmente exaustivas para os profissionais de saúde. A sobrecarga de trabalho, frequentemente associada a condições estruturais inadequadas, favorece a ocorrência de inconsistências no atendimento, especialmente quando o paciente é acompanhado por diferentes profissionais ao longo de sua internação. Assim, metodologias que promovam maior padronização, integração e agilidade no acompanhamento clínico surgem como alternativas promissoras para qualificar o cuidado em saúde.

Diante desse cenário, o presente trabalho propõe o desenvolvimento de um repositório digital inteiramente gerenciado pelo paciente, destinado ao registro e ao monitoramento de informações de saúde. A plataforma permite a inserção de dados referentes a condições pré-existentes, resultados de exames, hábitos de vida (alimentação, prática de atividade física, consumo de álcool, tabagismo), indicadores de saúde mental e até mesmo aspectos rotineiros, como tempo dedicado à televisão ou aos videogames.

Esses dados ficam acessíveis a uma rede de profissionais de saúde previamente cadastrados, como médicos, nutricionistas, enfermeiros, psicólogos, psiquiatras, educadores físicos, entre outros, os quais se conectam à plataforma por meio de registro profissional. O paciente, posicionado como elemento central do sistema, decide quais informações serão compartilhadas e pode, quando necessário, delegar essa função a um tutor, preferencialmente um familiar.

Além disso, os profissionais podem utilizar o ambiente para organizar suas redes de pacientes e colegas, bem como para moderar grupos voltados à educação em saúde. Dessa forma, o aplicativo estimula a troca de experiências, a disseminação de orientações especializadas e o fortalecimento do suporte oferecido aos pacientes.

A plataforma também prevê funcionalidades adicionais, como o compartilhamento seguro de informações entre diferentes profissionais envolvidos em um mesmo caso, bem como a solicitação de segundas ou terceiras opiniões (incluindo a recomendação automática de especialistas com apoio de inteligência artificial). Soma-se ainda a possibilidade de realização de atendimentos por telemedicina, quando apropriado.

Trata-se, portanto, de um estudo de caráter multidisciplinar, que busca integrar abordagens computacionais avançadas, com destaque para inteligência artificial e lógica *fuzzy*, como ferramentas de apoio para o aprimoramento dos sistemas de saúde.

1.1 Contextualização e Motivação

A saúde constitui um tema de relevância universal, independentemente do campo de atuação, e a busca por mecanismos capazes de aprimorar os sistemas de atenção tem impulsionado pesquisas contínuas em múltiplas áreas do conhecimento. Estudos direcionados ao desenvolvimento de novas metodologias e soluções inovadoras vêm ganhando visibilidade, sobretudo no que diz respeito à reestruturação dos processos de registro e análise de informações clínicas. Esses avanços não apenas buscam tornar os procedimentos mais seguros, ágeis e eficientes, como também visam explorar métodos de predição de comportamentos anômalos que possam influenciar de maneira significativa os tratamentos e a evolução dos pacientes [2].

Conforme destaca [3], a extração de informações provenientes de registros eletrônicos de saúde e sua aplicação em pesquisas científicas e na gestão hospitalar têm demonstrado resultados promissores. Entretanto, quando comparados a outras fontes de dados, esses registros ainda apresentam limitações relevantes, sobretudo no que se refere à eficácia e à aplicabilidade das informações. O estudo evidencia, ainda, a importância de apoiar os profissionais de saúde na compreensão dos processos de extração e análise de dados, de modo a ampliar o potencial dessas ferramentas. Nesse cenário, pesquisas que proponham metodologias ou soluções capazes de otimizar tais etapas tornam-se cada vez mais pertinentes.

No contexto brasileiro, a inexistência de um prontuário eletrônico universal, que integre informações tanto do Sistema Único de Saúde (SUS) quanto das redes privadas, configura um desafio persistente relacionado à centralização e à disponibilidade de dados clínicos. Diante disso, emerge a motivação para desenvolver uma solução inovadora que ultrapasse as limitações do prontuário eletrônico tradicional. A proposta de uma plataforma com características semelhantes às de uma rede social fundamenta-se na necessidade de disponibilizar uma ferramenta intuitiva, acessível e de fácil compreensão, que favoreça a conectividade e o compartilhamento de informações. Nesse modelo, o paciente assume posição central, determinando quais dados deseja compartilhar e com quais profissionais pretende estabelecer interação.

A pandemia de COVID-19 reforçou ainda mais a relevância do acesso rápido, preciso e integrado às informações de saúde. No período mais crítico da crise, a constante atualização de evidências científicas evidenciou que a disponibilidade ampla de dados clínicos dos pacientes poderia desempenhar papel decisivo no monitoramento, na prevenção e no tratamento de complicações.

Além disso, a incorporação de recursos de telemedicina, combinada ao uso de inteligência artificial e de técnicas baseadas em lógica *fuzzy*, amplia substancialmente o potencial da solução proposta. Essas tecnologias podem oferecer suporte à compreensão do comportamento das doenças, ao direcionamento terapêutico, ao aperfeiçoamento de estratégias de educação em saúde e ao desenvolvimento de estudos científicos, conferindo valor adicional ao presente trabalho.

1.2 Hipótese Científica e Tecnológica

As hipóteses científicas dessa tese são:

- Métodos de extração de informações de exames médicos guiados por aprendizado de máquina podem mitigar o problema da partida fria da rede e melhorar o apoio à classificação laboratorial e à tomada de decisão clínica quando comparados a métodos clássicos;
- Modelos interpretáveis baseados em regras *fuzzy* para detectar anomalias em variáveis colhidas manualmente ou por meio de *wearables* (alimentação, sono, frequência cardíaca, glicemia, exercício físico etc.) permitem compreender os relacionamentos entre os preditores e as variáveis alvo relacionadas às patologias do paciente, em particular diabetes *mellitus* tipo 2 (DM2).

A hipótese tecnológica dessa tese consiste em:

- Uma plataforma, no formato de linha do tempo, que se constitua em um repositório de informações pregressas e atuais de saúde do usuário e que esteja conectado a uma rede social, formada por profissionais de saúde e responsáveis pelo seu cuidado, pode auxiliar na prevenção de eventos e desfechos através do monitoramento e controle da saúde dos usuários.

1.3 Objetivos

O objetivo geral deste trabalho consiste em desenvolver uma plataforma para prevenção e monitoramento da saúde que integre um repositório de informações de saúde de seus usuários com uma rede social formada por profissionais de saúde e responsáveis pelo cuidado, integrando modelos de extração de informação baseado em aprendizado de máquina com modelos de classificação baseados em regras difusas.

1.3.1 Objetivos Específicos

Para atingir o objetivo geral os seguintes objetivos específicos serão necessários:

- Avaliar o estado da arte acerca da utilização da lógica *fuzzy* para classificação, da extração de informação de arquivos PDF (*Portable Document Format*) e da utilização do modelo ANFIS (*Adaptive Neuro-Fuzzy Inference System*) para classificação de doenças;
- Desenvolver um método baseado em aprendizado de máquina para extrair um conjunto de informações de saúde automaticamente a partir de documentos com resultados de exames médicos;
- Desenvolver modelos interpretáveis baseados em regras *fuzzy* para desfechos em saúde a partir de variáveis extraídas automaticamente;
- Desenvolver a plataforma propriamente dita.

1.4 Metodologia

A presente seção tem como objetivo apresentar os passos utilizados para o desenvolvimento da tese. Para isso, optou-se pela elaboração de um diagrama de fluxo de processo, apresentado pela Figura 1. A partir da figura pode-se observar que a mesma está dividida em quatro grandes etapas: a primeira delas consiste na definição do problema a ser resolvido, ou seja, em que área e de que maneira o trabalho pretende contribuir; a segunda, as revisões sistemáticas realizadas a fim de estudar o estado da arte; a terceira apresenta as metodologias adotadas a partir do estudo dos artigos encontrados e selecionados em cada uma das revisões sistemáticas; e, a quarta e última etapa consiste na apresentação dos resultados.

1.5 Projeto Dados48

O Dados48 consiste em um projeto submetido ao edital CNPq Programa Doutorado Acadêmico para Inovação - DAI Nº 21/2020, intitulado “Repositório dinâmico de informações pregressas e atuais de saúde do usuário”, vinculado à empresa Wiplay.

Tal projeto visa, com base em técnicas de inteligência artificial, desenvolver uma plataforma (APP) no formato de linha do tempo, que se constitua em um repositório dinâmico de informações pregressas e atuais de saúde do usuário e que esteja conectado em uma rede social, formada por profissionais de saúde responsáveis pelo seu cuidado.

1.6 Estrutura do Trabalho

O presente trabalho está dividido em outros 8 capítulos, são eles:

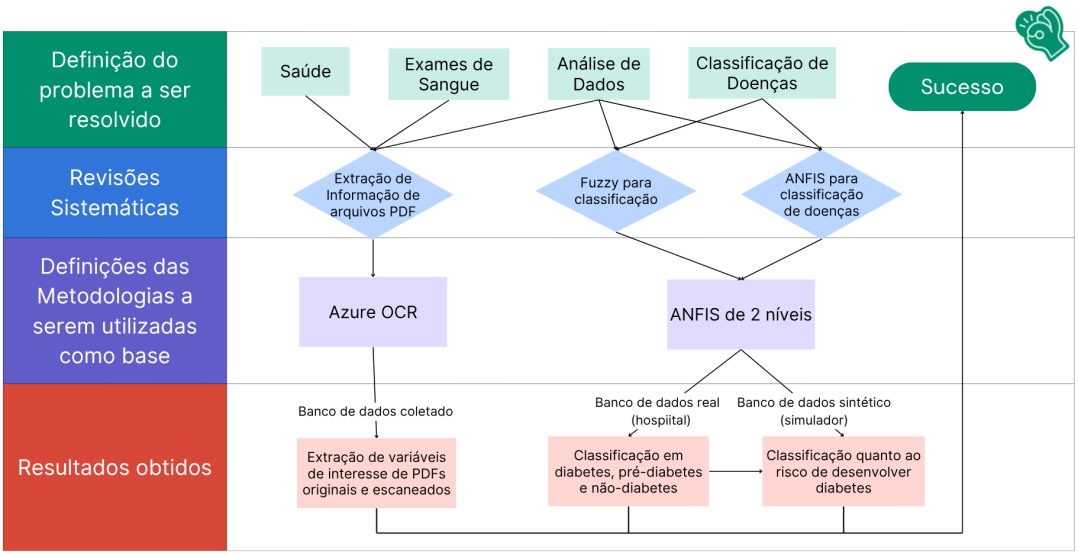


Figura 1: Diagrama de Fluxo de Processo
Fonte: Autora.

- CAPÍTULO 2: apresenta a fundamentação teórica técnica, onde são apresentados conceitos importantes para o melhor entendimento teórico e técnico acerca do trabalho desenvolvido;
- CAPÍTULO 3: apresenta as revisões sistemáticas realizadas a fim de avaliar o estado da arte sobre temas específicos e relacionados à este estudo;
- CAPÍTULO 4: apresenta a metodologia geral proposta neste estudo, destacando quais os três grandes focos de desenvolvimento;
- CAPÍTULO 5: apresenta o aplicativo desenvolvido, bem como os resultados obtidos após análise de desempenho da API e as interfaces geradas;
- CAPÍTULO 6: apresenta a etapa de extração de informação, contendo o método desenvolvido, o estudo de caso e os resultados obtidos;
- CAPÍTULO 7: apresenta a classificação laboratorial e o apoio À decisão clínica associado à DM2, contendo o banco de dados utilizado, a modelagem proposta e a análise dos resultados;
- CAPÍTULO 8: onde são apresentadas as conclusões obtidas a partir do resultados gerados, as limitações do estudo e as possibilidades de continuidade do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

O objetivo deste capítulo consiste em apresentar, de forma breve e resumida, algumas das principais definições acerca de termos e/ou tecnologias computacionais, técnicas, da área da saúde e da teoria *fuzzy* que são necessárias para um melhor entendimento no que diz respeito ao estudo multidisciplinar na qual consiste esta tese. Além disso, são apresentados conceitos associados à métricas de avaliação que serão utilizadas ao longo deste estudo para avaliar e validar as metodologias desenvolvidas e a definição de rede social.

2.1 Conhecimentos Técnicos Gerais

Esta seção visa apresentar conceitos necessários para o entendimento do processo de extração de informação de arquivos PDF, principalmente no que tange à técnicas utilizadas como base para a construção da metodologia proposta.

2.1.1 *Tokenização*

Segundo [4], a tokenização se trata do processo de conversão de uma sequência de texto em partes menores, conhecidas como *tokens*. Estes *tokens* podem ser tão pequenos quanto caracteres ou tão longos quanto palavras.

O principal objetivo da tokenização consiste em representar o texto de uma maneira que seja significativa para as máquinas sem perder o contexto. Ao converter o texto, no que se chama de *tokens*, os algoritmos conseguem identificar padrões com mais facilidade.

Os métodos de tokenização variam de acordo com a granularidade do detalhamento do texto e os requisitos específicos da tarefa em questão. Dessa forma, estes métodos podem variar desde dissecar o texto em palavras individuais até dividi-lo em caracteres ou unidades ainda menores [4].

2.1.2 *Form Understanding*

Segundo o estudo apresentado em [5], *form understanding* pode ser visto como um processo de extração automatizada de texto e estruturação de formulários baseado na

detecção e reconhecimento de palavras contidas no mesmo. Para isso, faz-se necessário que primeiro seja feita a análise das informações espaciais do *layout* do documento, logo após se faz o cruzamento destes dados com o texto e por fim, é necessário compreender e assimilar como estas entidades estão relacionadas entre si [6].

Quando *form understanding* é aplicado a documentos não estruturados, este processo pode ser descrito a partir das seguintes etapas [5]:

- ***Word Grouping***: Segundo [6, 7] está etapa é responsável por agregar entidades compostas por múltiplos *tokens* em uma única célula semântica. Nesta etapa são aplicada diversas técnicas, como por exemplo: técnicas baseadas em estatística, distribuição espacial, além de técnicas mais robustas como a GCN (*Graph Convolutional Networks*) [8];

- ***Semantic Entity Labeling* (SEL)**: Segundo [6], está atividade é responsável por classificar entidades em determinadas categorias pré-definidas, sendo elas: Q - *Question*, A - *Answers*, H - *Header* e O - *Others*. Esta etapa baseia-se na utilização de diversas *features* disponíveis no documento (por exemplo: disposição espacial, características da fonte do texto e a relação semântica entre as entidades), ou seja, segundo [9] utiliza-se do contexto em que o *token* está inserido no documento.

- ***Entity Linking***: Segundo [6], esta etapa tem como objetivo realizar a geração de modelo da relação entre as entidades semânticas classificadas na etapas SEL, ou seja, as informações classificadas são então empregadas como entrada para algoritmos de aprendizado de máquina capaz de analisar as possíveis relações entre as entidades de *Question* e *Answer* disponíveis para aquele documento [10].

Sendo assim, pode-se destacar que o presente trabalho apresenta como importante tarefa alvo a atividade de *Semantic Entity Labeling*. Tal atividade serve como base para o funcionamento da ferramenta *Azure Form Recognizer* da Microsoft, pois além dela apresentar a mesma descrição e funcionamento do estudo apresentado por [5] e da descrição apresentada neste trabalho, ela apresenta diversas funções dentre elas: extração de tabelas, formulários e pares-chave, ou seja, perguntas e respostas, disponibilizando os resultados em formato texto ou JSON (*JavaScript Object Notation*).

2.1.3 *Optical Character Recognition* (OCR)

OCR é um processo bastante utilizado, que consiste em reconhecer e extrair textos de arquivos de diferentes origens. Segundo [6], OCR tem como finalidade converter textos impressos ou manuscritos que foram digitalizados em texto selecionável e editável.

A partir da utilização do OCR é possível transformar documentos físicos em arquivos de texto editáveis ou em formatos que podem ser pesquisados e analisados digitalmente.

Isto acaba por facilitar a automação de processos, a extração de informação de documentos e o arquivamento digital.

Atualmente, o OCR está sendo utilizado em várias aplicações, como por exemplo: digitalização de documentos para criar versões digitais de arquivos físicos; automação de entrada de dados em sistemas de informação; extração de texto de imagens para análise ou processamento; acessibilidade para pessoas com deficiência visual, permitindo que leiam texto em documentos físicos por meio de sintetizadores de voz; dentre outras.

O processo todo do OCR pode ser dividido, resumidamente, em 4 etapas, são elas: processo de aquisição de imagem propriamente dita; o pré-processamento, onde os erros são removidos ou minimizados ao máximo; o reconhecimento do texto propriamente dito; e por fim, o pós-processamento, onde os dados de texto extraídos são convertidos em um arquivo informatizado.

Apesar do constante aperfeiçoamento das ferramentas de OCR e da capacidade destas ferramentas em lidar com uma variedade de idiomas, formatos e estilos de texto, a precisão do OCR depende muito da qualidade do texto de entrada, da resolução das imagens e do tipo de texto. Além disso, a complexidade do documento também pode ser vista como um fator que pode influenciar diretamente a precisão do OCR.

2.1.4 Extração de pares chave–valor

No contexto desta tese, um exame laboratorial em PDF é tratado como uma fonte de dados semiestruturada. Cada variável de interesse pode ser representada como uma chave, enquanto o resultado numérico extraído representa o valor associado a essa chave. Assim, um par chave–valor pode ser exemplificado por:

$$\text{chave} = \text{“glicose”}, \qquad \text{valor} = 102\ \text{mg/dL}.$$

Essa estrutura é adequada para integrar a informação extraída ao banco de dados da plataforma, pois permite armazenar e recuperar resultados laboratoriais de forma padronizada. Além disso, a estruturação dos pares chave–valor facilita a etapa posterior de modelagem, na qual variáveis como glicose, hemoglobina glicada e índice de massa corporal podem ser utilizadas como entradas dos sistemas *fuzzy*.

Entretanto, a extração de pares chave–valor em exames laboratoriais apresenta desafios importantes. Um mesmo exame pode aparecer com nomes diferentes, abreviações, variações de grafia ou unidades distintas. Além disso, valores de referência, comentários e observações podem estar próximos ao resultado do paciente, dificultando a identificação do valor correto. Por esse motivo, a avaliação do processo de extração deve considerar não apenas se a variável foi encontrada, mas também se o valor extraído corresponde ao resultado correto do exame.

2.2 Métricas de avaliação

A avaliação de modelos computacionais depende da tarefa realizada. Em problemas de classificação, o objetivo é atribuir amostras a classes ou categorias previamente definidas. Em problemas de regressão, o objetivo é estimar valores numéricos contínuos. Já em tarefas de extração de informação, frequentemente avalia-se se uma entidade foi corretamente localizada e se o valor associado a ela foi extraído de forma correta.

Nesta tese, as métricas são utilizadas principalmente em duas situações: (i) avaliação do processo de extração de informações de exames laboratoriais; e (ii) avaliação dos modelos de classificação ou categorização baseados em sistemas *fuzzy*. Por isso, esta seção separa explicitamente as métricas de classificação, as métricas utilizadas em extração de informação e as métricas de regressão.

2.2.1 Métricas para classificação

Problemas de classificação consistem em atribuir uma amostra a uma ou mais classes. Em classificação binária, existem apenas duas classes possíveis, geralmente denominadas positiva e negativa. Em classificação multiclasse, existem três ou mais classes mutuamente exclusivas. Em classificação multirrótulo, uma mesma amostra pode pertencer simultaneamente a mais de uma classe.

2.2.1.1 Matriz de confusão

A matriz de confusão é uma representação tabular que compara as classes previstas pelo modelo com as classes reais observadas. Em problemas de classificação binária, ela é formada por quatro quantidades principais:

- **Verdadeiros Positivos (VP):** amostras positivas corretamente classificadas como positivas;
- **Verdadeiros Negativos (VN):** amostras negativas corretamente classificadas como negativas;
- **Falsos Positivos (FP):** amostras negativas incorretamente classificadas como positivas;
- **Falsos Negativos (FN):** amostras positivas incorretamente classificadas como negativas.

A Tabela 1 apresenta a estrutura básica da matriz de confusão para classificação binária.

Tabela 1: Estrutura da matriz de confusão para classificação binária.

	Predito positivo	**Predito negativo**
Real positivo	VP	FN
Real negativo	FP	VN

A matriz de confusão permite compreender não apenas quantas predições foram corretas, mas também qual tipo de erro o modelo comete. Em aplicações na área da saúde, essa distinção é importante, pois falsos negativos e falsos positivos podem ter impactos clínicos distintos.

2.2.1.2 Acurácia

A acurácia representa a proporção de classificações corretas em relação ao total de amostras avaliadas. Para classificação binária, é dada por:

$$\text{Acurácia} = \frac{VP+VN}{VP+VN+FP+FN}. \tag{1}$$

Embora seja uma métrica intuitiva, a acurácia pode ser pouco informativa quando as classes estão desbalanceadas. Nesses casos, um modelo pode obter alta acurácia simplesmente classificando a maior parte das amostras como pertencentes à classe majoritária.

2.2.1.3 Precisão

A precisão, ou *precision*, mede a proporção de amostras classificadas como positivas que realmente pertencem à classe positiva:

$$\text{Precisão} = \frac{VP}{VP+FP}. \tag{2}$$

Essa métrica é particularmente importante quando falsos positivos representam um problema relevante. Em extração de informação, por exemplo, uma precisão elevada indica que, quando o sistema extrai uma variável, essa extração tende a estar correta.

2.2.1.4 Recall ou sensibilidade

O *recall*, também denominado sensibilidade, mede a capacidade do modelo de identificar corretamente as amostras positivas existentes:

$$Recall = \frac{VP}{VP+FN}. \tag{3}$$

Valores elevados de *recall* indicam baixa quantidade de falsos negativos. Em tarefas de extração de informação, um *recall* elevado indica que o sistema consegue encontrar a maior parte das variáveis que deveriam ser extraídas.

2.2.1.5 *F1-Score*

O *F1-Score* corresponde à média harmônica entre precisão e *recall*:

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \textit{Recall}}{\text{Precisão} + \textit{Recall}}. \tag{4}$$

Essa métrica é útil quando se deseja obter uma medida única que equilibre a proporção de extrações corretas e a capacidade de recuperar os elementos relevantes. Por isso, é frequentemente utilizada em tarefas de classificação desbalanceada e de extração de informação.

2.2.1.6 *Hamming Loss*

Em problemas de classificação multirrótulo, uma mesma amostra pode estar associada a vários rótulos simultaneamente. Nesses cenários, a métrica *Hamming Loss* quantifica a fração de rótulos classificados incorretamente em relação ao total de rótulos avaliados:

$$HL = \frac{1}{N \cdot L} \sum_{i=1}^{N} \sum_{j=1}^{L} \mathbb{I}(y_{ij} \neq \widehat{y}_{ij}), \tag{5}$$

em que N é o número de amostras, L é o número de rótulos, y_{ij} é o valor real do rótulo j para a amostra i, $\widehat{y}_{ij}$ é o valor previsto e $\mathbb{I}(\cdot)$ é a função indicadora. Quanto menor o valor da *Hamming Loss*, melhor é o desempenho do modelo.

2.2.1.7 *Average Precision*

A *Average Precision* (AP) avalia a qualidade do ranqueamento produzido por um modelo, combinando precisão e *recall* ao longo de diferentes limiares de decisão:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n, \tag{6}$$

em que P_n representa a precisão e R_n representa o *recall* no limiar n. Valores próximos de 1 indicam melhor capacidade de ordenar corretamente as amostras positivas acima das negativas.

2.2.2 Métricas para extração de informação

Na etapa de extração de informação dos exames laboratoriais, as métricas de precisão, *recall* e *F1-Score* são interpretadas em termos das variáveis extraídas. Assim, considera-se:

- **VP**: variável de interesse que deveria ser extraída e foi extraída corretamente;
- **FP**: variável extraída incorretamente, seja por não existir no exame, seja por estar associada a um valor incorreto;

- **FN**: variável que deveria ter sido extraída, mas não foi identificada pelo sistema.

Essa avaliação pode ser realizada em dois níveis. No primeiro, considera-se cada exame como unidade de análise, avaliando quantas variáveis deveriam ser extraídas e quantas foram corretamente recuperadas. No segundo, considera-se cada variável de interesse individualmente, avaliando seu desempenho ao longo do conjunto de exames. Essa distinção é importante porque uma ferramenta pode apresentar bom desempenho médio por exame, mas ainda assim falhar em variáveis específicas de maior relevância clínica.

2.2.3 Métricas para regressão

Em problemas de regressão, a saída do modelo é um valor contínuo. Nesses casos, métricas baseadas diretamente na matriz de confusão não são adequadas. Uma das métricas mais utilizadas é a raiz do erro quadrático médio, ou *Root Mean Squared Error* (RMSE), definida por:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}, \tag{7}$$

em que y_i é o valor observado, $\widehat{y}_i$ é o valor estimado pelo modelo e N é o número de amostras.

No contexto de sistemas *fuzzy* do tipo Takagi–Sugeno–Kang, as saídas internas podem ser contínuas e posteriormente interpretadas como graus de suporte, escores ou categorias. Por isso, é importante distinguir entre a etapa matemática de estimação de valores contínuos e a etapa final de interpretação classificatória.

2.2.4 Rede Social

Entende-se por rede social qualquer estrutura social composta por pessoas ou organizações, conectadas por um ou vários tipos de relações, que compartilham valores e objetivos comuns.

Quando se fala em rede social logo vem à mente sites ou aplicativos como *Facebook* (Meta), *TikTok*, *Twitter* (X) e *Instagram*. Essas plataformas criaram uma nova forma de relacionamento entre empresas e clientes, abrindo caminhos para interação e para o anúncio de produtos ou serviços.

Segundo [11], a maioria das redes sociais contam com elementos em comum, sendo os 5 principais:

- Perfil público: seja de uma pessoa ou de uma marca, traz informações básicas sobre o dono(a) do perfil, incluindo uma foto, uma pequena biografia e número de seguidores;

- Amigos e seguidores: são as pessoas que se deseja acompanhar e com quem se deseja compartilhar conteúdo;

- *Feed* de publicações: a maioria das plataformas tem uma página inicial, que a pessoa vê quando faz *login*. Na grande maioria, apresenta um *feed* com atualizações de perfis, que permite acompanhar atividades, pensamentos e notícias compartilhadas;
- Curtidas, comentários e compartilhamentos: obter e fornecer *feedbacks* são elementos fundamentais das redes sociais;
- *Hashtags*: Uma *hashtag* consiste em uma palavra ou frase precedida pelo símbolo #, usadas nos conteúdos compartilhados na maioria das redes. Possui o objetivo de reunir postagens com o mesmo tema e ajudar outras pessoas a encontrá-las.

As principais vantagens das redes sociais são, segundo [12]:

- Aproxima as pessoas que vivem em locais diferentes, pois é uma maneira fácil de manter as relações e o contato;
- Possibilita a interação em tempo real;
- Facilita a relação com quem está mais perto, permitindo manter uma relação de proximidade sem se encontrar fisicamente;
- Oferece uma forma rápida e eficaz de comunicar algo para um grande número de pessoas ao mesmo tempo;
- Permite avisar sobre um acontecimento, a preparação de uma manifestação ou a mobilização de um grupo para um protesto;
- Facilita a organização de eventos, enviando convites e solicitando a confirmação de presença.

Por outro lado, as principais desvantagens são [12]:

- Falta de privacidade;
- Exige cuidado na divulgação de certos pormenores da vida de cada um;
- Crianças e adolescentes divulgam informações sobre a escola e locais que frequentam;
- Pode causar dependência, pois em alguns casos as pessoas não conseguem se ”desligar” das redes sociais, deixando coisas importantes por fazer;
- Criação de perfil falso para postar comentários racistas, preconceituosos e racistas;
- Facilidade de divulgação de notícias, fatos e imagens sem a verificação da fonte, podendo ser ”*fake news*”.

2.3 Conhecimentos associados à área da saúde

Esta seção visa apresentar conceitos necessários para o entendimento associado à área da saúde, uma vez que se trata de um trabalho multidisciplinar e que tem um foco grande em conceitos associados à saúde.

2.3.1 Diabetes *Mellitus*

Segundo [13], o diabetes trata-se de uma doença causada pela produção insuficiente ou pela má absorção de insulina.

A insulina, por sua vez, consiste em um hormônio produzido pela células β das ilhotas de Langerhans do pâncreas. Sua principal função consiste em reduzir a concentração de glicose no sangue, facilitando sua entrada nas células para ser utilizada como fonte de energia ou armazenada. Em resumo, a insulina tem a função de promover a captação celular da glicose via GLUT4 (*Glucose Transporter Type* 4). No DM2, a insulina está presente, mas as células respondem menos ao seu sinal. Como consequência, ocorre menor translocação de GLUT4 e menor entrada de glicose nas células.

O diabetes, quando não controlado, pode acarretar no aumento da glicemia e as altas taxas podem levar a complicações no coração, nas artérias, nos olhos, nos rins e nos nervos. Em casos mais graves, o diabetes pode levar inclusive à morte.

De acordo com a Sociedade Brasileira de Diabetes (SBD) [14], existem no Brasil, atualmente, mais de 13 milhões de pessoas vivendo com a doença, o que acaba por representar cerca de 6,9% da população nacional.

Apesar de não se conhecer a causa do tipo de diabetes, a melhor forma de prevenir a doença consiste na realização de práticas de vida saudáveis, como: realização de atividades físicas, alimentação saudável, evitar ingestão de álcool, tabaco e outras drogas.

Além dos fatos genéticos e a ausência de hábitos saudáveis, existem alguns outros fatores que podem, de certa forma, contribuir para o desenvolvimento do diabetes, são eles [13]:

- Diagnóstico de pré-diabetes;
- Pressão alta;
- Colesterol alto ou alterações na taxa de triglicerídeos no sangue;
- Sobrepeso, principalmente se a gordura estiver concentrada em volta da cintura;
- Pais, irmãos ou parentes próximos com diabetes;
- Doenças renais crônicas;
- Mulher que deu à luz de uma criança com mais de 4kg;

- Diabetes gestacional;
- Síndrome de ovários policísticos;
- Diagnóstico de distúrbios psiquiátricos - esquizofrenia, depressão, transtorno bipolar;
- Apneia do sono;
- Uso de medicamentos da classe dos glicocorticoides.

Em geral, os principais sintomas associados ao diabetes são: fome e sede excessiva e vontade de urinar várias vezes ao dia.

O diabetes *mellitus* pode apresentar diversas formas e possui diversos tipos diferentes [14, 13]. A seguir, os tipos de diabetes são apresentados de forma detalhada.

Ressalta-se que as diretrizes nacionais da SBD [14] foram consideradas em conjunto com referências internacionais, como a *American Diabetes Association* (ADA) [15], a *World Health Organization* (WHO) [16] e a *International Diabetes Federation* (IDF) [17].

2.3.1.1 Diabetes Tipo 1

Pode ser definido como uma doença metabólica crônica de origem autoimune, caracterizada pela destruição progressiva das células β das ilhotas pancreáticas, responsáveis pela produção de insulina. Com isso, ocorre deficiência absoluta ou quase absoluta desse hormôniom comprometendo a capacidade do organismo de controlar os níveis de glicose no sangue.

Embora exista uma forte predisposição genética associada principalmente aos genes do sistema HLA (*Human Leukocyte Antigen*) classe II, o desenvolvimento da doença depende também da interação com fatores ambientais. Essa combinação desencadeia uma resposta imunológica inadequada, na qual o próprio sistema imune passa a reconhecer as células β como estruturas estranhas e as destrói gradativamente.

A ausência de insulina impede que tecidos insulino-dependentes, como músculos e tecido adiposo, realizem adequadamente a captação da glicose por meio dos transportadores GLUT4. Como resultado, a glicose permanece acumulada na corrente sanguínea, provocando hiperglicemia.

Os principais sintomas associados ao diabetes tipo 1 são:

- Fome frequente;
- Sede constante;
- Vontade de urinar diversas vezes ao dia;
- Perda de peso;

- Fraqueza;
- Fadiga;
- Mudanças de humor;
- Náusea e vômito.

O diabetes tipo 1 manifesta-se com maior frequência na infância e adolescência, embora possa surgir em qualquer fase da vida. O diagnóstico geralmente ocorre após o aparecimento dos sintomas clínicos ou pela identificação de autoanticorpos específicos, como os anticorpos anti-GAD, anti-IA2 e anti-insulina.

Como o organismo perde a capacidade de produzir insulina suficiente, o tratamento requer administração exógena de insulina por toda a vida, associada ao monitoramento frequente da glicemia, orientação nutricional e prática regular de atividade física.

Em resumo, o Diabetes Mellitus Tipo 1 é uma doença autoimune multifatorial, resultante da interação entre predisposição genética e fatores ambientais, que leva à destruição das células produtoras de insulina e ao consequente aumento dos níveis de glicose no sangue. Sua compreensão é fundamental para o diagnóstico precoce, prevenção de complicações e adequado manejo clínico dos pacientes.

2.3.1.2 Diabetes Tipo 2

O diabetes tipo 2 aparece quando o organismo não consegue usar adequadamente a insulina que produz, ou ainda, quando não produz insulina suficiente para controlar a taxa de glicemia.

Atualmente, cerca de 90% das pessoas com diabetes possuem o tipo 2. Este tipo tende a se manifestar com maior frequência em adultos, mas crianças também podem apresentar. Dependendo da gravidade o diabetes tipo 2 pode ser controlado somente com atividade física e planejamento alimentar, porém, em casos mais graves, faz-se necessário que se utilize insulina e/ou outros medicamentos para controlar a glicose.

Os principais sintomas associados ao diabetes tipo 2 são:

- Fome frequente;
- Sede constante;
- Formigamento nos pés e mãos;
- Vontade de urinar diversas vezes ao dia;
- Infecções frequentes na bexiga, rins, pele e infecções de pele;
- Feridas que demoram para cicatrizar;

- Visão embaçada.

Destaca-se que, segundo as diretrizes clínicas, para um paciente ser confirmado com diabetes tipo 2 quando o mesmo não apresenta sintomas clássicos de hiperglicemia, ele precisa apresentar dois testes confirmatórios. Dessa forma, segundo as diretrizes, o diagnóstico pode ser estabelecido por um dos seguintes exames ([16, 14, 15]:

- Glicemia de jejum $\geq$ 126 mg/dL;
- Hemoglobina glicada (HbA1c) $\geq$ 6,5 %;
- Glicemia 2 horas após TOTG (75 g de glicose) $\geq$ 200 mg/dL;
- Glicemia plasmática casual $\geq$ 200 mg/dL associada a sintomas clássicos de diabetes.

Portanto, quando o paciente é assintomático, um único exame alterado não é suficiente para confirmar o diagnóstico. Neste caso, faz-se necessário que o paciente repita o mesmo exame em outro dia e obtenha o resultado alterado novamente.

Diagnósticos diferenciais

Quando se fala em DM2, é importante considerar alguns diagnósticos diferenciais, especialmente em pacientes que apresentam características clínicas atípicas. Entre eles, destacam-se o *Latent Autoimmune Diabetes in Adults* (LADA) e o *Maturity-Onset Diabetes of the Young* (MODY). Ambos apresentam glicose e HbA1c elevadas e seriam classificados como DM2, mas apresentam fisiopatologia, prognóstico e tratamento distintos:

- LADA corresponde a aproximadamente 10% dos diagnosticados como DM2 em idade adulta. Apresenta evolução autoimune lenta, anticorpos positivos e pode demandar insulinoterapia precoce;
- MODY corresponde a até 5% dos casos de DM, herança autossômica dominante, demanda rastreio em familiares e tem subtipos com manejo específico, como MODY 2 (GCK) e MODY 3 (HNF1A), que apresentam implicações terapêuticas distintas.

Dessa forma, apesar de saber desses casos particulares, esta tese não pretende, pelo menos em um primeiro momento, lidar com tais limitações, deixando para que tais casos sejam melhor explorados em trabalhos posteriores.

2.3.1.3 Diabetes Gestacional

Durante o período da gravidez, a fim de permitir o correto desenvolvimento do bebê, a mulher passa por uma série de mudanças em seu equilíbrio hormonal. A placenta, por exemplo, é uma fonte importante de hormônios que reduzem a ação da insulina, responsável pela captação e utilização da glicose pelo corpo. Consequentemente, o pâncreas acaba aumentando a produção de insulina para compensar este quadro.

Porém, em algumas mulheres, este processo não ocorre e elas acabam desenvolvendo um quadro chamado de diabetes gestacional, caracterizado pelo aumento do nível de glicose no sangue. Este quadro pode de alguma forma afetar o bebê, uma vez que quando ele é exposto a grandes quantidades de glicose, mesmo que ainda no ambiente intrauterino, há maior risco de crescimento excessivo (macrossomia fetal) e, consequentemente, partos traumáticos, hipoglicemia neonatal e até mesmo risco de obesidade e diabetes na vida adulta [14].

Os principais fatores de risco associados ao desenvolvimento da diabetes gestacional são:

- Idade materna mais avançada;
- Ganho de peso excessivo durante a gestação;
- Sobrepeso ou obesidade;
- Síndrome dos ovários policísticos;
- História prévia de bebês grandes (mais de 4kg) ou de diabetes gestacional;
- História de diabetes gestacional na mãe da gestante;
- História familiar de diabetes em parentes de 1º grau (pais e irmãos);
- Hipertensão arterial na gestação;
- Gestação múltipla (gravidez de gêmeos).

2.3.1.4 Pré-Diabetes

O termo pré-diabetes é utilizado quando os níveis de glicose no sangue estão mais altos do que o normal, mas não o suficiente para um diagnóstico de DM2. Pessoas consideradas no grupo de risco são aquelas que apresentam alterações nos lipídios, obesidade e hipertensão.

Segundo [14], cerca de 50% dos pacientes no estágio pré-diabetes vão acabar por desenvolver a doença. É importante destacar que o pré-diabetes é especialmente importante uma vez que se trata da única etapa que ainda pode ser revertida ou mesmo permite retardar a evolução para diabetes e suas complicações.

A mudança de hábito alimentar e a prática de exercícios são os principais fatores de sucesso para o controle. No entanto, para cerca de 60% dos pacientes, a dieta acaba por ser o passo mais difícil a ser incorporado na rotina. Ao todo, em torno de 95% apresentam dificuldades com o controle de peso, dieta saudável e exercícios regulares.

Após a definição dos principais tipos de diabetes *mellitus*, o próximo passo consistia em saber interpretar e entender como uma determinada pessoa é classificada em diabética, pré-diabética ou não diabética, ou seja, quais as variáveis analisadas e seus respectivos limites. Para isso, utilizou-se como base um pseudo-algoritmo apresentado pela Figura 2 e retirado da [14].

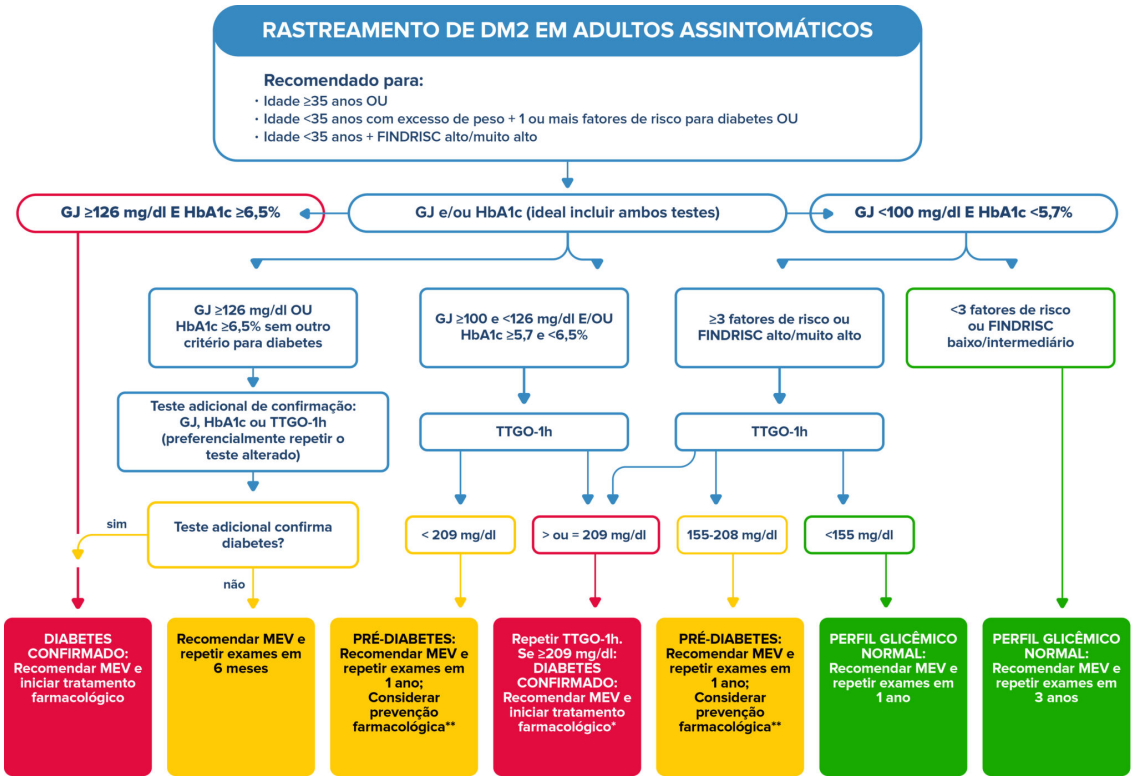


Figura 2: Pseudo-algoritmo retirado das diretrizes da SBD e utilizado como base para pré-classificação dos pacientes em diabéticos, pré-diabéticos e não diabéticos

Fonte: [14]

2.3.2 Exames de Sangue

O exame de sangue consiste em um termo popularmente utilizado para se referir a qualquer análise realizada a partir do sangue do paciente [18]. Os exames de sangue são solicitados pelos médicos com o intuito de avaliar, de forma geral, o estado de saúde do paciente.

Segundo [19], realizar exames de sangue de forma regular colabora para o diagnóstico precoce e para a prevenção de muitas doenças e, consequentemente, para um tratamento mais eficaz. Algumas doenças que podem ser diagnosticadas a partir da realização de exames de sangue são:

- Inflamações e infecções;
- Patologias cardíacas;
- Pressão alta;
- Diabetes;
- Sobrepeso e obesidade;
- Diversos tipos de câncer;
- Problemas na vista;
- Dermatites e doenças de pele.

Além disso, o exame de sangue é visto como um exame essencial antes de cirurgias, a fim de verificar se o paciente em questão se encontra apto a passar pelo procedimento [20].

A seguir, são apresentados os exames de sangue mais comuns, a fim de apresentar a finalidade e a diferença de cada um dos exames mais solicitados pelos médicos [19, 18, 21].

- Hemograma: Considerado o exame mais pedido, apresenta como estão as células sanguíneas do paciente e consegue identificar muitas doenças. O hemograma apresenta os resultados de:
  - Glóbulos brancos: é a defesa do corpo contra doenças, infecções e inflamações;
  - Glóbulos vermelhos: transportam oxigênio e, a partir da sua redução, pode ser caracterizada a anemia;
  - Plaquetas: ajudam a indicar a coagulação do sangue e possíveis problemas de circulação.
- Perfil Lipídico: consiste no exame que analisa as quantidades de determinadas lipoproteínas responsáveis pelo transporte do colesterol no corpo do paciente. No exame, constam os valores de:
  - Colesterol *High-Density Lipoprotein* (HDL);
  - Colesterol *Low-Density Lipoprotein* (LDL);
  - Colesterol *Very Low-Density Lipoprotein* (VLDL);
  - Colesterol não HDL;
  - Colesterol Total;

  - Triglicerídeos.

  O HDL (considerado colesterol bom), contribui para a proteção dos vasos sanguíneos do acúmulo de placas de gordura, sendo assim, é importante que a taxa esteja em uma alta concentração. Por outro lado, o LDL (considerado colesterol ruim) e o VLDL se apresentarem um alto acúmulo nos vasos acabam gerando infarto ou derrame, por isso, as taxas devem estar sempre abaixo do limite recomendado. Já a análise de triglicerídeos no sangue verifica o risco de entupimento das artérias, assim como, da tendência do paciente em desenvolver outras doenças cardíacas;

- Glicose e Glicemia: a análise da dosagem da glicose na corrente sanguínea auxilia no diagnóstico e até mesmo no controle do tratamento da diabetes *mellitus*. Enquanto a glicose é importante para o organismo pois se trata da principal fonte de energia, a glicemia serve para mensurar o nível de açúcar. Este exame consegue indicar se o paciente apresenta glicemia normal, pré-diabetes ou diabetes.

- Ácido Úrico: consiste no indicador na corrente sanguínea capaz de identificar a presença de gota ou cálculo renal;

- Transaminases: taxas importantes para avaliação do fígado, uma vez que se tratam de enzimas responsáveis pela metabolização das proteínas. Quando tanto o Transaminase Glutâmico-Pirúvica (TGP) quanto o Transaminase Glutâmico-Oxalacética (TGO) estão com níveis elevados, significa que o órgão está com algum problema.

- *Thyroid Stimulating Hormone* (TSH) e Tiroxina (T4) Livre: são dois hormônios que ajudam no diagnóstico de funcionamento da tireoide. Quando o TSH encontra-se alto e T4 baixo, é porque o paciente pode estar com hipotireoidismo. Por outro lado, se o TSH estiver baixo e o T4 elevado, é possível que o paciente esteja com hipertireoidismo. Já se ambos estiverem altos, o problema é na hipófise.

- Creatinina e Ureia (Função renal): a avaliação dos níveis séricos de creatinina e ureia é amplamente utilizada na análise da função renal e na identificação de possíveis quadros de insuficiência renal. A creatinina, associada a variáveis como idade e sexo do paciente, é empregada no cálculo da Taxa de Filtração Glomerular (TFG), um dos principais indicadores da capacidade de filtração dos rins. Valores elevados de creatinina e ureia, bem como a redução da TFG, sugerem comprometimento da função renal, indicando uma diminuição da capacidade dos rins de filtrar adequadamente os resíduos metabólicos presentes no sangue.

- *Prostate-Specific Antigen* (PSA): consiste em um marcador no sangue que ajuda a identificar alterações na glândula da próstata, sendo um marcador útil para rastreamento e seguimento do câncer.

- Proteína C Reativa (PCR): consiste em uma pentraxina, ou seja, proteína de fase aguda produzida pelo fígado em resposta à IL-6. Se a taxa estiver elevada significa que há um processo inflamatório no organismo.

- Eletrólitos: os eletrólitos mais conhecidos são o sódio, cálcio, potássio e o fósforo. Eles ajudam a transportar água, bem como conduzir eletricidade pelo corpo. Quando as taxas estão dentro do normal, eles asseguram o bom funcionamento do organismo. Porém, se esses níveis encontram-se muito elevados ou reduzidos na circulação sanguínea, faz-se necessário que haja uma investigação a fim de fazer um tratamento correto.

- *Human Immunodeficiency Virus* (HIV): ajuda a detectar a presença do vírus da *Acquired Immunodeficiency Syndrome* (AIDS), em suas duas variações (HIV 1 e HIV 2), no sangue.

2.4 Teoria de conjuntos *fuzzy*, lógica *fuzzy* e sistemas *fuzzy*

Esta seção apresenta os fundamentos da teoria de conjuntos *fuzzy*, da lógica *fuzzy* e dos sistemas de inferência *fuzzy*. Essa distinção é importante porque esses termos, embora relacionados, não são equivalentes.

A **teoria de conjuntos *fuzzy*** fornece a base matemática para representar pertencimento gradual. Diferentemente da teoria clássica de conjuntos, em que um elemento pertence ou não pertence a um conjunto, a teoria de conjuntos *fuzzy* permite que um elemento pertença parcialmente a um conjunto, com grau de pertinência no intervalo $[0, 1]$ [22].

A **lógica *fuzzy*** utiliza essa ideia de gradualidade para generalizar operadores lógicos, como “E”, “OU” e “NÃO”, permitindo raciocinar com proposições parcialmente verdadeiras. Assim, enquanto a teoria de conjuntos *fuzzy* define como representar conceitos vagos por funções de pertinência, a lógica *fuzzy* define como combinar e inferir informações a partir desses graus de verdade.

Por fim, um **sistema *fuzzy*** é uma aplicação computacional que utiliza conjuntos *fuzzy*, regras linguísticas, operadores de inferência e, quando necessário, mecanismos de defuzzificação para resolver problemas de classificação, controle, tomada de decisão ou apoio à decisão. No contexto desta tese, os sistemas *fuzzy* são utilizados por sua capacidade de representar transições graduais entre categorias clínicas e por sua interpretabilidade.

2.4.1 Conjuntos *fuzzy*

Seja X um universo de discurso. Um conjunto *fuzzy* A em X é definido por uma função de pertinência

$$\mu_A : X \to [0,1],$$

que associa a cada elemento $x \in X$ um grau de pertencimento ao conjunto A.

Formalmente, o conjunto *fuzzy* pode ser representado como:

$$A = \{(x, \mu_A(x)) \mid x \in X\}. \tag{8}$$

Quando $\mu_A(x) = 0$, entende-se que x não pertence ao conjunto A. Quando $\mu_A(x) = 1$, entende-se que x pertence completamente a A. Quando $0 < \mu_A(x) < 1$, o elemento pertence parcialmente ao conjunto. Essa estrutura permite modelar conceitos vagos, como "glicose normal", "risco intermediário" ou "IMC alto", sem impor fronteiras abruptas entre categorias.

2.4.2 Funções de pertinência

As funções de pertinência descrevem a forma como os graus de pertencimento são atribuídos aos elementos do universo de discurso. A escolha da função depende do conhecimento do especialista, da natureza da variável e do objetivo da modelagem [23, 24].

As principais funções utilizadas em sistemas *fuzzy* são apresentadas a seguir.

2.4.2.1 Função triangular

A função triangular é definida por três parâmetros $a < b < c$:

$$\mu_A(x) = \begin{cases} 0, & x \leq a, \\ \dfrac{x-a}{b-a}, & a < x \leq b, \\ \dfrac{c-x}{c-b}, & b < x < c, \\ 0, & x \geq c. \end{cases} \tag{9}$$

Essa função é simples, computacionalmente eficiente e útil quando se deseja representar um ponto central de máxima pertinência com transições lineares para ambos os lados.

2.4.2.2 *Função trapezoidal*

A função trapezoidal é definida por quatro parâmetros $a < b < c < d$:

$$\mu_A(x) = \begin{cases} 0, & x \leq a, \\ \dfrac{x-a}{b-a}, & a < x \leq b, \\ 1, & b < x \leq c, \\ \dfrac{d-x}{d-c}, & c < x < d, \\ 0, & x \geq d. \end{cases} \tag{10}$$

Essa função possui uma região central de pertinência máxima e é adequada para representar intervalos nos quais determinado conceito é plenamente válido. No contexto desta tese, funções trapezoidais são particularmente relevantes para representar faixas laboratoriais ou clínicas com regiões de transição em torno de limiares.

2.4.2.3 *Função gaussiana*

A função gaussiana é definida por:

$$\mu_A(x) = \exp\left(-\frac{(x-c)^2}{2\sigma^2}\right), \tag{11}$$

em que c representa o centro da curva e σ controla sua largura.

Funções gaussianas são comuns em modelos neuro-*fuzzy* e em aplicações de aprendizado, pois apresentam transições suaves e diferenciáveis.

2.4.2.4 *Função sigmoidal*

A função sigmoidal pode ser escrita como:

$$\mu_A(x) = \frac{1}{1+\exp[-a(x-c)]}, \tag{12}$$

em que a controla a inclinação da curva e c define o ponto de inflexão.

Esse tipo de função é útil para representar conceitos com transição gradual crescente ou decrescente, como risco crescente em função de determinada variável.

2.4.3 Sistemas de inferência *fuzzy*

Um sistema de inferência *fuzzy* é geralmente composto por três etapas principais:

1. **Fuzzificação**: converte valores numéricos de entrada em graus de pertinência associados a conjuntos *fuzzy*;
2. **Inferência**: aplica regras linguísticas do tipo "SE–ENTÃO" para combinar

informações e produzir uma conclusão *fuzzy* ou uma saída intermediária;

3. **Defuzzificação**: quando necessária, converte a saída *fuzzy* em um valor numérico único.

Em problemas de apoio à decisão, essa estrutura é útil porque permite representar conhecimento de especialistas de forma explícita. Por exemplo, uma regra pode ser escrita como:

> *SE a glicose é alta E a hemoglobina glicada é alta, ENTÃO a classificação laboratorial possui forte suporte para diabetes.*

Esse tipo de regra é interpretável, pois explicita quais condições contribuem para a saída do modelo.

2.4.4 Modelos Mamdani e Takagi–Sugeno–Kang

Entre os modelos de inferência *fuzzy*, destacam-se os modelos de Mamdani e Takagi–Sugeno–Kang (TSK). No modelo de Mamdani, tanto os antecedentes quanto os consequentes das regras são expressos por termos linguísticos. Esse modelo é frequentemente utilizado em aplicações nas quais a interpretabilidade linguística das regras é prioridade.

No modelo TSK, os antecedentes das regras continuam sendo expressos por conjuntos *fuzzy*, mas os consequentes são funções matemáticas das variáveis de entrada. Em um sistema TSK de primeira ordem, o consequente da regra k assume a forma:

$$f^k(x) = p_0^k + \sum_{j=1}^{A} p_j^k x_j, \tag{13}$$

em que $x = (x_1, \ldots, x_A)$ é o vetor de entrada e p_j^k são parâmetros do consequente.

Nesta tese, as regras clínicas utilizadas para construir determinadas classificações de referência possuem natureza linguística e podem ser descritas no estilo Mamdani. Já a arquitetura neuro-*fuzzy* utilizada para aprendizado segue a estrutura TSK, pois combina antecedentes *fuzzy* com consequentes paramétricos.

2.4.5 Funções de agregação

Em sistemas *fuzzy*, funções de agregação são utilizadas para combinar múltiplos graus de pertinência em um único valor representativo [25]. Formalmente, uma função de agregação pode ser definida como uma aplicação:

$$\mathcal{A} : [0,1]^A \to [0,1],$$

crescente em todos os seus argumentos e satisfazendo as condições de contorno:

$$\mathcal{A}(0,\ldots,0) = 0, \qquad \mathcal{A}(1,\ldots,1) = 1.$$

Essas funções são importantes na etapa de ativação das regras, pois permitem combinar os graus de pertinência dos antecedentes. Por exemplo, em uma regra com duas condições, como “glicose alta” e “HbA1c alta”, é necessário definir como os dois graus de pertinência serão combinados para determinar a força de ativação da regra.

2.4.6 T-normas

As t-normas, ou normas triangulares, são operadores utilizados para generalizar a conjunção lógica “E” na lógica *fuzzy* [26]. Uma t-norma é uma função $T : [0,1]^2 \to [0,1]$ que satisfaz as seguintes propriedades, para quaisquer $a, b, c \in [0,1]$:

- **Comutatividade**: a ordem dos operandos não altera o resultado:

$$T(a,b) = T(b,a).$$

- **Associatividade**: o agrupamento das operações não altera o resultado:

$$T(a,T(b,c)) = T(T(a,b),c).$$

- **Monotonicidade**: se $a_1 \leq a_2$ e $b_1 \leq b_2$, então:

$$T(a_1,b_1) \leq T(a_2,b_2).$$

- **Elemento neutro**: o valor 1 atua como elemento neutro da conjunção:

$$T(a,1) = a.$$

Entre as t-normas mais utilizadas, destacam-se:

$$T_{\min}(a,b) = \min(a,b), \tag{14}$$

$$T_{\prod}(a,b) = a \cdot b, \tag{15}$$

$$T_{\text{Łuk}}(a,b) = \max(0, a+b-1). \tag{16}$$

A escolha da t-norma influencia diretamente a força de ativação das regras. A t-norma do mínimo adota uma interpretação conservadora, pois a ativação da regra é determinada pelo antecedente menos satisfeito. A t-norma do produto, por sua vez, considera a contribuição conjunta dos antecedentes, penalizando combinações em que um dos graus de pertinência é baixo.

2.4.7 Modelos caixa branca baseados em regras *fuzzy*

Modelos de classificação caixa branca baseados em regras *fuzzy* são sistemas interpretáveis que representam conhecimento por meio de regras linguísticas e funções de pertinência. Diferentemente de modelos caixa preta, nos quais a relação entre entradas e saídas pode ser difícil de interpretar, os modelos baseados em regras permitem analisar explicitamente como cada variável contribui para a decisão.

Essa característica é particularmente relevante em aplicações na saúde, nas quais é desejável que o sistema forneça resultados compreensíveis por especialistas. No caso desta tese, a utilização de regras e funções de pertinência permite representar faixas clínicas de forma gradual, reduzir a rigidez de limiares fixos e oferecer uma interpretação mais transparente dos resultados.

2.4.8 Sistemas de Inferência Neuro-*Fuzzy* Adaptativos (ANFIS)

Os Sistemas de Inferência Neuro-*Fuzzy* Adaptativos, conhecidos como ANFIS, combinam a interpretabilidade dos sistemas *fuzzy* com a capacidade de aprendizado de modelos baseados em dados [27]. Em geral, a estrutura ANFIS utiliza regras *fuzzy* cujos parâmetros podem ser ajustados a partir de um conjunto de treinamento.

Seja $x = (x_1, \ldots, x_A) \in \mathbb{R}^A$ o vetor de entrada, em que A representa o número de atributos, e seja $y \in \mathbb{R}^C$ o vetor de saída, em que C representa o número de classes ou categorias consideradas. O conjunto de dados pode ser representado por:

$$\mathcal{D} = \left\{ (x_i, y_i) \in \mathbb{R}^A \times \mathbb{R}^C \right\}_{i=1}^{N}.$$

No contexto desta tese, a saída do modelo pode ser interpretada como um vetor de suporte às categorias consideradas. Assim, os componentes de y não devem ser lidos como diagnóstico clínico definitivo, mas como escores ou graus de confiança associados a categorias laboratoriais ou de risco.

A abordagem utilizada tem como base um sistema *fuzzy* Takagi–Sugeno–Kang multiclasse, aqui denotado por MC-TSK-FS. A sigla TSK corresponde a Takagi–Sugeno–Kang e deve ser mantida nessa ordem em todo o texto [28, 29, 30, 31].

O MC-TSK-FS é composto por um conjunto de K regras de inferência *fuzzy*. A k-ésima regra pode ser escrita como:

$$R^k: \text{ SE } x_1 \text{ é } B_1^k \text{ E } \ldots \text{ E } x_A \text{ é } B_A^k \text{ ENTÃO } y = L^k(x, p^k),$$

em que B_j^k é o conjunto *fuzzy* associado à variável x_j na regra k, e L^k é a função conse-

quente linear. Para um consequente de primeira ordem, tem-se:

$$L^k(x) = p_0^k + \sum_{j=1}^{A} p_j^k x_j, \tag{17}$$

em que $p_0^k, \ldots, p_A^k \in \mathbb{R}$ são parâmetros a serem estimados.

A Figura 3 ilustra a arquitetura geral do MC-TSK-FS, composta por cinco camadas: fuzzificação, força de ativação das regras, normalização, avaliação dos consequentes e agregação ponderada da saída.

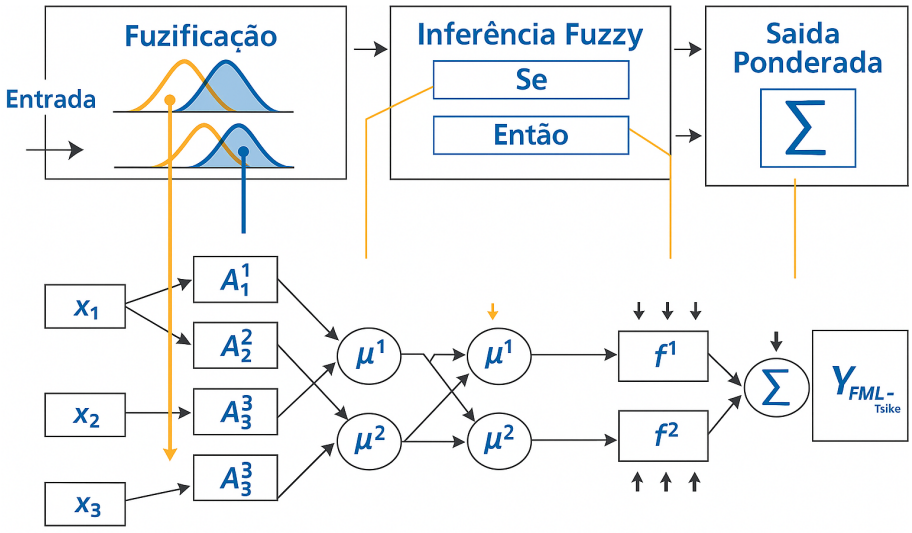


Figura 3: Arquitetura do modelo MC-TSK-FS.
Fonte: Autora.

2.4.8.1 Camada 1: Fuzzificação

A primeira camada converte valores numéricos de entrada em graus de pertinência. Para cada variável x_j e cada regra R^k, calcula-se:

$$\mu_{B_j^k}(x_j),$$

que representa o grau com que x_j satisfaz o antecedente B_j^k.

Essa etapa permite representar transições graduais entre categorias e incorporar conhecimento de especialistas por meio das funções de pertinência.

2.4.8.2 Camada 2: Força de ativação das regras

A segunda camada calcula a força de ativação de cada regra. Essa força representa o grau com que todos os antecedentes da regra são simultaneamente satisfeitos. Utilizando

uma função de agregação $\mathcal{A}$, a força de ativação da regra R^k é:

$$\mu^k_{\mathcal{A}}(x) = \mathcal{A}\left(\mu_{B^k_1}(x_1), \ldots, \mu_{B^k_A}(x_A)\right). \tag{18}$$

Quando a t-norma do produto é adotada, obtém-se:

$$\mu^k_{\mathcal{A}}(x) = \prod_{j=1}^{A} \mu_{B^k_j}(x_j). \tag{19}$$

Quanto maior a força de ativação, maior é a contribuição potencial da regra correspondente para a saída final do sistema.

2.4.8.3 Camada 3: Normalização

A terceira camada normaliza as forças de ativação, de modo que a contribuição relativa das regras possa ser comparada. O peso normalizado da regra R^k é dado por:

$$\widetilde{\mu}^k_{\mathcal{A}}(x) = \frac{\mu^k_{\mathcal{A}}(x)}{\sum_{i=1}^{K} \mu^i_{\mathcal{A}}(x)}. \tag{20}$$

Assim, $\widetilde{\mu}^k_{\mathcal{A}}(x)$ representa a contribuição relativa da regra k em relação ao conjunto de regras ativadas.

2.4.8.4 Camada 4: Avaliação dos consequentes

Na quarta camada, cada regra produz uma saída local por meio de sua função consequente linear, definida na Equação 17. Essa etapa representa a parte paramétrica do modelo TSK e permite ajustar a resposta local de cada regra aos dados de treinamento.

2.4.8.5 Camada 5: Agregação ponderada

A quinta camada combina as saídas locais das regras por meio dos pesos normalizados. A saída final do sistema é:

$$\widehat{y} = \sum_{k=1}^{K} \widetilde{\mu}^k_{\mathcal{A}}(x)\, L^k(x, p^k). \tag{21}$$

Essa agregação ponderada garante que regras mais ativadas tenham maior influência na saída final. Como consequência, o modelo produz uma resposta contínua e interpretável, adequada para representar graus de suporte ou categorias de risco.

2.4.9 Aprendizado dos parâmetros

O processo de aprendizado do MC-TSK-FS tem como objetivo estimar os parâmetros dos consequentes das regras. Após a fuzzificação, a ativação das regras e a normalização,

a saída do modelo pode ser escrita de forma compacta como:

$$\widehat{Y} = P^T G,$$

em que P representa a matriz de parâmetros dos consequentes e G contém as informações das entradas processadas pelas camadas *fuzzy*.

Uma função objetivo típica para estimar P pode ser definida como:

$$O_f(P) = \frac{1}{2} \left\| P^T G - Y \right\|_F^2 + \beta \|P\|_1 , \tag{22}$$

em que o primeiro termo mede o erro entre as saídas previstas e observadas, enquanto o segundo termo corresponde à regularização ℓ_1, ponderada pelo parâmetro β.

A regularização contribui para evitar sobreajuste e favorecer soluções mais parcimoniosas. A minimização pode ser realizada por métodos de otimização adequados a funções compostas por um termo diferenciável e um termo regularizador não diferenciável, como a descida de gradiente proximal.

2.5 Resumo do capítulo

Este capítulo apresentou os conceitos necessários para fundamentar a metodologia desenvolvida na tese. Inicialmente, foram discutidos os elementos técnicos associados à extração automática de informações de exames laboratoriais em PDF, incluindo tokenização, OCR, *form understanding* e pares chave–valor. Em seguida, foram apresentadas as métricas de avaliação, separando explicitamente métricas de classificação, extração de informação e regressão.

Na sequência, foram abordados os fundamentos clínicos relacionados ao Diabetes *Mellitus*, com ênfase no Diabetes Tipo 2, nos critérios laboratoriais, no pré-diabetes, nos diagnósticos diferenciais e nos escores de risco. Também foi destacada a diferença entre diagnóstico clínico e classificação diagnóstica laboratorial, que é essencial para interpretar corretamente o escopo da proposta.

Por fim, foram apresentados os fundamentos da teoria de conjuntos *fuzzy*, da lógica *fuzzy*, dos sistemas de inferência *fuzzy* e dos modelos neuro-*fuzzy* adaptativos. Essa base permite compreender por que a metodologia proposta utiliza regras interpretáveis e funções de pertinência para representar, de forma gradual, categorias clínicas e níveis de risco.

3 REVISÕES SISTEMÁTICAS

O presente capítulo tem como objetivo apresentar as revisões sistemáticas realizadas durante a construção desta tese. Tais revisões foram utilizadas como base para o entendimento do estado da arte acerca dos assuntos principais deste estudo, visando aprender o que se tinha de tecnologia sendo utilizada e de que forma este estudo pode contribuir e, ao mesmo tempo, se diferenciar, do que já está posto e validado no mercado.

Foram realizadas três revisões sistemática: a primeira com o objetivo de apresentar o estado da arte associado ao processo de extração de informação de PDFs; a segunda, o estado da arte associado aos sistemas baseados em regras *fuzzy* para predição e classificação; e, por fim, a terceira, que buscava entender o estado da arte acerca da metodologia ANFIS para predição e/ou classificação de doenças. A seguir, cada uma delas é apresentada de forma detalhada, seguindo a mesma ideia de processo.

3.1 Revisão 01: Extração de Informação de PDFs

A metodologia utilizada para a realização da revisão sistemática seguiu o método apresentado por [32] e está sumarizada no fluxograma apresentado pela Figura 4. A cada evolução do processo apresentada no fluxograma, os artigos encontrados nas buscas vão sendo retirados do escopo do estudo, a fim de que restem somente aqueles que melhor atendem aos critérios pré-estabelecidos. A seguir, para cada uma das etapas chaves apresentadas no fluxograma são discutidos em detalhes os critérios adotados.

3.1.1 Questões de Pesquisa

O primeiro passo para iniciar o processo de revisão sistemática consiste na definição das questões de pesquisa que se deseja responder, ou seja, quais as questões que se deseja que os artigos selecionados sejam capazes de responder. Tais questões são as responsáveis por dar origem aos termos de busca e as combinações de palavras-chave utilizadas durante as pesquisas nas bases de dados selecionadas.

Sendo assim, as questões de pesquisa que se deseja que os artigos selecionados sejam capazes de responder são:

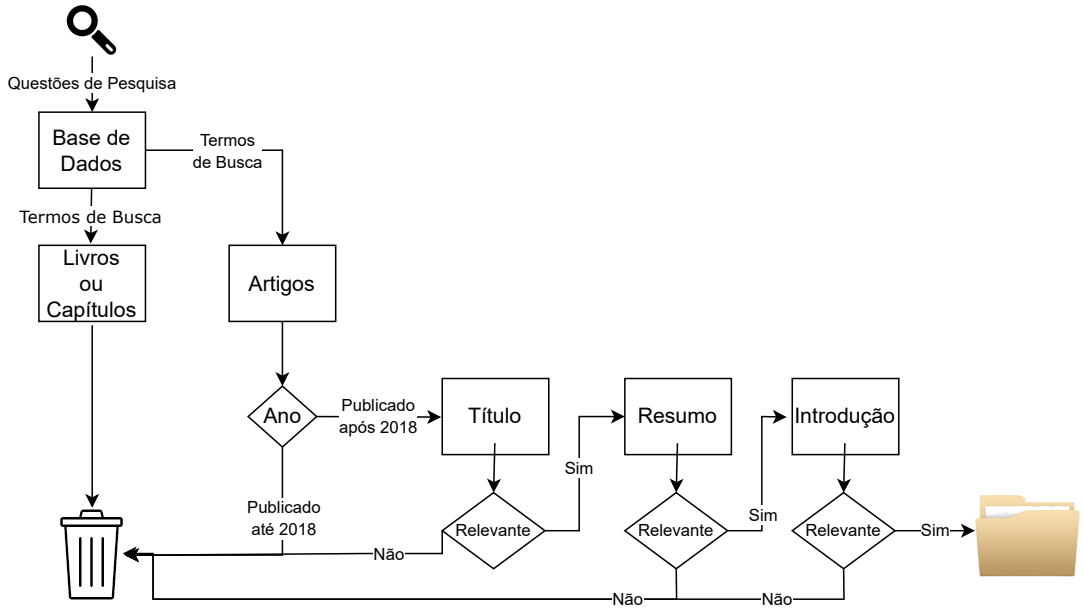


Figura 4: Fluxograma da metodologia da revisão teórica da Extração de Informação
Fonte: Autora.

- **Questão 1**: Quais os métodos usados para extração de documentos PDF?
- **Questão 2**: Quais os métodos usados para extração de informação de dados de exames de sangue?

3.1.2 Termos de Busca

Os termos de busca foram definidos a partir das duas questões de pesquisa apresentadas anteriormente. Ressalta-se que se optou por considerar somente os termos de busca em inglês. Dessa forma, os termos de busca considerados para a construção das combinações de palavras-chave foram:

- *information extraction*;
- *PDF*;
- *blood*.

A partir destes três termos de busca selecionados foram elaboradas as combinações de palavras-chave que foram utilizadas na pesquisa sistemática nas bases de dados selecionadas. A Tabela 2 apresenta para cada Ordem de Pesquisa (OP), as combinações de palavras-chave utilizadas.

3.1.3 Bases utilizadas

Para realizar uma pesquisa completa e mais robusta, diferentes bibliotecas digitais foram consideradas. Precisamente, as bases consideradas foram:

Tabela 2: Combinações de palavras-chave utilizadas para revisão sistemática de extração de informação.

OP	Combinação
1	"*information extraction*"
2	"*information extraction*" *and* "PDF"
3	"*information extraction*" *and* "*blood*"
4	"*information extraction*" *and* "PDF" *and* "*blood*"

Fonte: Autora.

- *Web of Science*[1];
- PubMed[2];
- *Scopus*[3];

3.1.4 Critérios de Exclusão/Inclusão

Os Critérios de Exclusão (CE) são definidos como aspectos do estudo que preenchem os Critérios de Inclusão (CI), mas apresentam características adicionais, que poderiam interferir no sucesso da pesquisa ou aumentar o risco de obter informações desnecessárias para o estudo. Os CEs considerados foram:

- **CE 1**: Artigos publicados antes de 2018;
- **CE 2**: Livros e capítulos de livros;
- **CE 3**: Trabalhos que não sejam no idioma português ou inglês;
- **CE 4**: Trabalhos que não estejam totalmente publicados. Ex. *drafts*;

Por outro lado, os CIs são definidos como as características chave da população ou pesquisa que está sendo realizada. Os CIs são utilizados para responder às questões de pesquisa e são apresentados a seguir:

- **CI 1 - Título**: Artigos que possuem extração de informação no título;
- **CI 2 - Resumo**: Artigos que abordem no resumo o assunto de extração de informação em arquivos PDF;
- **CI 3 - Introdução**: Artigos que destaquem na introdução que o objetivo do trabalho consiste no processo de extração de informação de arquivos PDF de exames de sangue.

[1] https://www-webofscience.ez40.periodicos.capes.gov.br/wos/woscc/basic-search
[2] https://www-ncbi-nlm-nih.ez40.periodicos.capes.gov.br/pmc/
[3] https://www-scopus.ez40.periodicos.capes.gov.br/search/form.uri?display=basic#basic

3.1.5 Selecionando Trabalhos

Nesta subseção são apresentados os dados relacionados aos CEs e CIs, bem como a quantidade de artigos que foram sendo retornados a partir da aplicação de cada um dos critérios elencados anteriormente.

Inicialmente, a pesquisa foi realizada considerando cada uma das combinações de palavra-chave destacadas anteriormente, para cada uma das bases selecionadas, sem considerar, no primeiro momento, nenhum CE. Na Tabela 3 são retornados para cada uma das OPs o total de artigos. Por linha são consideradas as bibliotecas digitais e por colunas as etapas do processo de revisão (ver Figura 4).

O próximo passo consiste na aplicação do CE relacionado ao ano de publicação dos resultados retornados. Para este caso, portanto, foram considerados apenas os resultados publicados de 2018 até o presente momento. Tal CE fez com que o total de resultados retornados fosse menor, como mostra a coluna "CE 1" Tabela 3.

Os próximos passos consistiram na aplicação dos outros CEs. A coluna "CE 2" da Tabela 3 apresenta os resultados após a exclusão dos livros e capítulos de livros, enquanto a coluna "CE 3" apresenta os resultados após a exclusão dos trabalhos que não estejam escritos em português ou inglês. Por fim, a "CE 4" da Tabela 3 apresenta os resultados dos artigos que não estavam completamente publicados.

Em resumo, a Tabela 3 apresenta para cada uma das fontes bibliográficas o total de artigos retornados para cada OP, bem como o total de artigos retornados por base de dados. Além disso, através da Tabela 3 pode-se observar que a cada CE aplicado o total de artigos retornados, não só por OP, mas também por fonte bibliográfica, diminuiu.

Analisando-se a Tabela 3 pode-ser observar que a biblioteca PubMed foi a que retornou um menor número de artigos considerando todas as OPs. Por outro lado, a *Scopus* foi a biblioteca que retornou o maior número de resultados. Além disso, a OP que retornou mais artigos foi a OP1, que considerava a relação de *information extraction*.

Realizada a etapa de aplicação dos CEs, o próximo passo consiste na aplicação dos CIs. Porém, antes de se iniciar tal processo, optou-se por desconsiderar os resultados retornados a partir da OP1, uma vez que a pesquisa realizada foi muito ampla, retornando um total muito grande de resultados.

O primeiro passo para a aplicação dos CIs consiste na avaliação inicial do título de todos os trabalhos pré-selecionados. Caso o título não atenda aos critérios estabelecidos, o artigo é rejeitado, caso atenda, avança para a próxima análise. Após a verificação do título dos 203 artigos que haviam restado, restaram um total de 72 artigos. Vale ressaltar que nesse momento ainda não estavam sendo analisados se o mesmo artigo estava aparecendo em mais de uma combinação ou em mais de uma base de dados.

O próximo passo consiste na análise do resumo e a ideia era semelhante a do título, caso o resumo não apresentasse as informações desejadas o trabalho era descartado. Neste caso, dos 72 artigos restaram 56 após a aplicação do CI2. Por fim, avaliava-se a introdução

Tabela 3: Total de artigos retornados para cada um dos passos descritos

Fontes Bibliográficas	**Inicial**	**CE 1**	**CE 2**	**CE 3**	**CE 4**
Scopus	OP1: 21765 OP2: 165 OP3: 93 OP4: 0 **Total: 22023**	OP1: 8749 OP2: 83 OP3: 42 OP4: 0 **Total: 8874**	OP1: 8521 OP2: 80 OP3: 42 OP4: 0 **Total: 8643**	OP1: 7845 OP2: 77 OP3: 35 OP4: 0 **Total: 7957**	OP1: 7707 OP2: 76 OP3: 35 OP4: 0 **Total: 7818**
PubMed	OP1: 1929 OP2: 9 OP3: 39 OP4: 0 **Total: 1977**	OP1: 980 OP2: 5 OP3: 19 OP4: 0 **Total: 1004**	OP1: 980 OP2: 5 OP3: 19 OP4: 0 **Total: 1004**	OP1: 965 OP2: 4 OP3: 17 OP4: 0 **Total: 986**	OP1: 958 OP2: 4 OP3: 17 OP4: 0 **Total: 979**
Web Of Science	OP1: 13112 OP2: 91 OP3: 66 OP4: 0 **Total: 13269**	OP1: 5567 OP2: 42 OP3: 30 OP4: 0 **Total: 5639**	OP1: 5539 OP2: 41 OP3: 30 OP4: 0 **Total: 5610**	OP1: 5442 OP2: 41 OP3: 30 OP4: 0 **Total: 5513**	OP1: 5442 OP2: 41 OP3: 30 OP4: 0 **Total: 5513**
Total por OP	OP1: 36806 OP2: 265 OP3: 198 OP4: 0	OP1: 15296 OP2: 130 OP3: 91 OP4: 0	OP1: 15040 OP2: 126 OP3: 91 OP4: 0	OP1: 14252 OP2: 122 OP3: 82 OP4: 0	OP1: 14107 OP2: 121 OP3: 82 OP4: 0

Fonte: Autora.

do trabalho, a fim de filtrar ainda mais os artigos e, após a aplicação do CI3, restaram 26 artigos. Após a aplicação de todos os critérios de inclusão foram analisados os títulos do trabalho a fim de verificar se dos 26 artigos que haviam restado algum estava aparecendo em mais de uma OP ou base de dados. Após tal análise, o total de artigos selecionados foi igual a 21, cujas informações de título, autores e ano de publicação são apresentadas nas Tabelas 4 e 5. Além disso, a tabela apresenta o número de citações considerando o *Google Scholar* e em qual OP cada um dos trabalhos foi encontrado.

3.1.6 Trabalhos Selecionados

Os 21 artigos selecionados foram estudados com o intuito de entender o estado da arte atual relacionado ao processo de extração de informação de arquivos PDFs, mas também com o intuito de encontrar quais os *softwares* que estão sendo mais utilizados no que tange ao processo de gerenciamento de documentos.

A partir do estudo mais específico destes artigos selecionados um novo artigo acabou sendo encontrado. O trabalho apresentado por [54], além de complementar o estado da arte do processo de extração, ainda destacada que os *softwares* mais populares no que tange ao gerenciamento de documentos são *Amazon Textract*, *Microsoft Azure Form Recognizer* e *Google Document AI*. Por fim, o estudo ainda apresenta de maneira mais específica o *software* da *Microsoft*, o que acabou sendo bastante proveitoso do ponto de vista deste estudo.

O estudo apresentado por [54] aponta que a melhor opção para o gerenciamento de documentos em texto está na utilização do OCR. Neste estudo são realizados diferentes testes utilizando o *Azure Form Recognizer* da *Microsoft* e para todos eles os resultados foram extremamente satisfatórios. Por isso, o estudo aponta que o *Azure Form Recognizer* é o *software* mais confiável no que tange ao processo de gerenciamento de texto em documentos, chegando a um percentual superior a 94,21% de resposta.

As indústrias armazenam diversos tipos de dados em formatos estruturados e não estruturados. Tais dados são coletados ao longo de muito anos, fazendo com que se tenha um grande volume de dados acumulados. Além do grande volume de dados, estes dados são oriundos de diversas fontes e por isso são apresentados de diferentes maneiras. Dessa forma, lidar com tais dados acaba por se tornar uma tarefa trabalhosa e custosa para a empresa, mas extremamente necessária, uma vez que tais dados são importantíssimos para o acompanhamento e monitoramento dos serviços das empresas [55].

Sendo assim, o estudo apresentado por [55], visa propor uma abordagem econômica nativa da nuvem a partir da utilização de banco de dados para resolver uma série de problemas, alguns deles destacados anteriormente. Para isso, diversos itens foram tratados no estudo, sendo um deles o processo de extração de dados de forma automática ao invés da utilização de uma equipe de funcionários para realizar tal tarefa. Para realizar tal tarefa de extração o estudo optou pela utilização da tecnologia disponível pela *Microsoft*,

Tabela 4: Informações básicas dos artigos selecionados

Referência	Título	Autores	Ano	Total de Citações	OP
[33]	A *Semi-automatic Data Extraction System for Heterogeneous Data Sources: A Case Study from Cotton Industry*	Arnab G. Chowdhury, Nils Schut, Martin Atzmuelle	2021	2	2
[34]	*A Hybrid Information Extraction Approach using Transfer Learning on Richly-Structured Documents*	Arnab Ghosh Chowdhury, Nils Schut, Martin Atzmueller	2021	0	2
[35]	*A New Annotation Method and Dataset for Layout Analysis of Long Documents*	Aman Ahuja, Kevin Dinh, Brian Dinh, William A. Ingram, Edward A. Fox	2023	1	2
[36]	*ARTIFACT: Architecture for Automated Generation of Distributed Information Extraction Pipelines*	Michael Sildatke, Hendrik Karwanni, Bodo Kraft, Albert Zündorf	2022	0	2
[37]	*A rule-based information extraction approach for extracting metadata from PDF books*	Abrar Alamoudi, Amal Alomari, Sarah Alwarthan, Atta-ur-Rahman	2021	14	2
[38]	*CyberPDF: Smart and Secure Coordinate-based Automated Health PDF Data Batch Extraction*	Reza M. Parizi, Liang Guo, Yao Bian, Ali Dehghantanha, Amin Azmoodeh, Kim-Kwang R. Choo	2018	6	2
[39]	*Extracting Semantic Relations for Scholarly Knowledge Base Construction*	Rabah A. Al-Zaidy, C. Lee Giles	2018	22	2
[40]	*Extracting ToC and Metadata from PDF Books: A Rule-Based Approach*	Huda Alghamdi, Waad Dawwas, Taghreed H. Almutairi, Atta-Ur-Rahman	2022	1	2
[41]	*Information extraction from scanned invoice images using text analysis and layout features*	H.T. Ha, A. Horák	2022	3	2
[42]	*Information Extraction from Visually Rich Documents with Font Style Embeddings*	Ismail Oussaid, William Vanhuffel, Pirashanth Ratnamogan, Mhamed Hajaiejm, Alexis Mathey, Thomas Gilles	2022	0	2

Fonte: Autora.

Tabela 5: Informações básicas dos artigos selecionados (continuação)

Referência	Título	Autores	Ano	Total de Citações	OP
[43]	*Intelligent Information Retrieval: Handling Variability in Document Structure*	Aarushi Gupta, Akhil Chawla, K. S. Shushrutha, Mohana	2022	0	2
[44]	*Method of Automatically Extracting Squeeze Casting Process Data from the Text of Periodical Literature*	Xiangming Zeng, Jianxin Deng, Rui Tang, Bin Xie	2021	1	2
[45]	*A Benchmark of PDF Information Extraction Tools using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents*	Norman Meuschke1, Apurva Jagdale, Timo Spinde, Jelena Mitrovic, Bela Gipp	2023	3	2
[46]	*Using OCR Framework and Information Extraction for Thai Documents Digitization*	Todsanai Chumwatana, Waramporn Rattana-umnuaychai	2021	6	2
[47]	*Toward A Robust Method for Understanding the Replicability of Research*	Ben Gelman, Chae Clark, Scott Friedman, Ugur Kuter, James Gentile	2021	3	2
[48]	*Text Mining from Free Unstructured Text: An Experiment of Time Series Retrieval for Volcano Monitoring*	Margherita Berardi, Luigi S. Amato, Francesca Cigna, Deodato Tapete, Mario S. de Cumis	2022	4	2
[49]	*Resume extraction with conditional random field method*	Jason Anggakusuma, Viny Christanti Mawardi, Manatap Dolok Lauro	2020	3	2
[50]	*OntoHuman: Ontology-based Information Extraction tools with Human-in-the-loop interaction*	Kobkaew Opasjumruskit, Sarah Böning, Sirko Schindler, Diana Peters	2022	4	2
[51]	*Insights to the State-of-the-Art PDF Extraction Techniques*	Hashmi, Ahmer Maqsood; Qayyum, Faiza; and Afzal, Muhammad Tanvi	2020	5	2
[52]	*Optimal dynamic treatment regime estimation using information extraction from unstructured clinical text*	Nina Zhou, Robert D. Brook, Ivo D. Dinov, Lu Wang	2022	2	3
[53]	*Knowledge Extraction of Long-Term Complications from Clinical Narratives of Blood Cancer Patients with HCT Treatments*	Weizhong Zhu, Jennifer B. Teh, Haiqing Li, Saro H. Armenian	2018	0	3

Fonte: Autora.

conhecida como *Azure Form Recognizer* e concluí que tal ferramenta apresentou uma qualidade satisfatória no processo de extração de dados.

3.2 Revisão 02: Sistemas baseados em regras *Fuzzy* para predição e classificação

A metodologia utilizada para a realização da revisão sistemática, assim como apresentado na revisão anterior, seguiu o método apresentado por [32] e está sumarizada no fluxograma apresentado pela Figura 5. A seguir, para cada uma das etapas chaves apresentadas no fluxograma, serão discutidos em detalhes os critérios adotados.

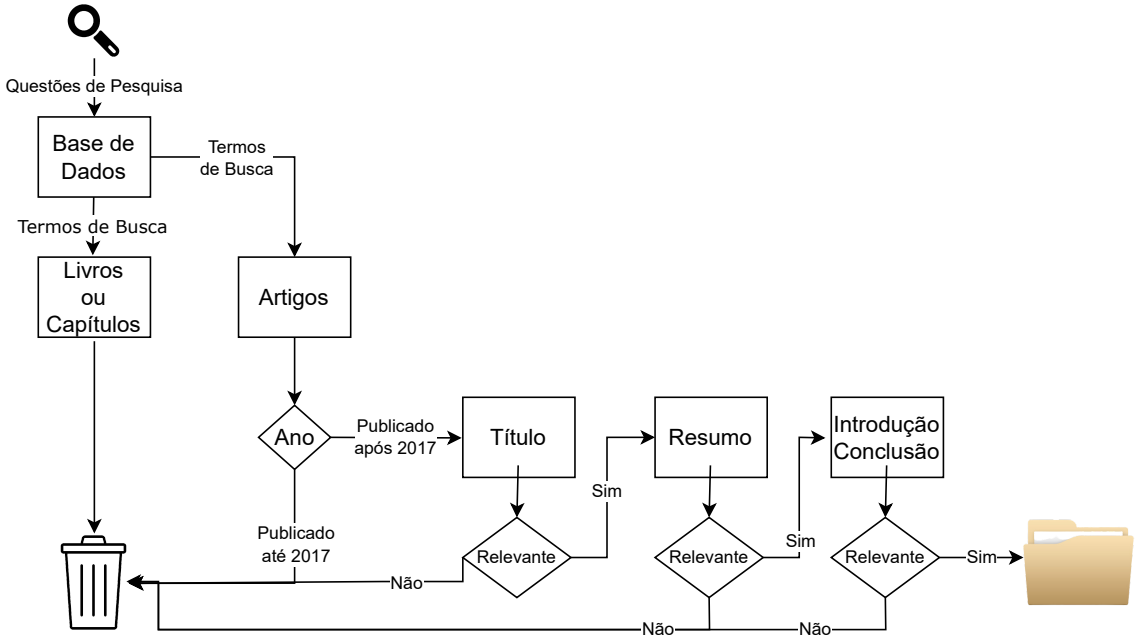


Figura 5: Fluxograma da metodologia da revisão teórica dos sistemas *Fuzzy*
Fonte: Autora.

3.2.1 Questões de Pesquisa

As questões de pesquisa que se deseja que os artigos selecionados sejam capazes de responder são:

- **Questão 1**: Quais os métodos preditivos utilizados para ajustar modelos sobre diferentes variáveis de exames laboratoriais sanguíneos?
- **Questão 2**: Quais os métodos de classificação utilizados para ajustar modelos sobre diferentes variáveis de exames laboratoriais sanguíneos?
- **Questão 3**: Quais os métodos preditivos utilizados para ajustar modelos sobre diferentes bases de dados de saúde?

- **Questão 4**: Quais os métodos de classificação utilizados para ajustar modelos sobre diferentes bases de dados de saúde?

Ressalta-se que, para cada uma das pesquisas realizadas, também foram analisados sistemas baseados em regras *fuzzy* relacionados aos métodos preditivos, pois a ideia do trabalho consiste em utilizar sistemas *fuzzy* para prever anomalias, principalmente relacionado a diabetes, com base em informações extraídas de exames de sangue.

3.2.2 Termos de Busca

Os termos de busca foram definidos a partir das quatro questões de pesquisa apresentadas anteriormente. Destaca-se o fato de que, assim como na revisão apresentada anteriormente, se optou por considerar os termos de busca somente em inglês, uma vez que se entende que se trata da língua mais utilizada na escrita de artigos para os eventos e congressos mais importantes.

Os termos de busca considerados para a construção das combinações de palavras-chave foram:

- *blood test*;
- *health database*;
- *classification*;
- *prediction*;
- *Fuzzy*.

A partir destes termos de busca selecionados foram elaboradas algumas combinações de palavras-chave que foram utilizadas na pesquisa sistemática nas bases de dados selecionadas. A Tabela 6 apresenta a relação da OP com as combinações de palavras-chave utilizadas.

Tabela 6: Combinações de palavras-chave utilizadas para revisão sistemática de sistemas *fuzzy*

OP	Combinação
1	"*blood test*" *and* "*prediction*"
2	"*blood test*" *and* "*classification*"
3	"*blood test*" *and* "*prediction*" *and* "*fuzzy*"
4	"*blood test*" *and* "*classification*" *and* "*fuzzy*"
5	"*health database*" *and* "*prediction*"
6	"*health database*" *and* "*classification*"
7	"*health database*" *and* "*prediction*" *and* "*fuzzy*"
8	"*health database*" *and* "*classification*" *and* "*fuzzy*"

Fonte: Autora.

3.2.3 Bases utilizadas

Para essa revisão sistemática as bases consideradas foram:

- *IEEE Xplore*[4];
- *ACM Digital Library*[5];
- *Scopus*[6];
- *SpringerLink*[7].

3.2.4 Critérios de Exclusão/Inclusão

Os CEs considerados para esta revisão sistemática foram:

- **CE 1**: Artigos publicados antes de 2017;
- **CE 2**: Livros e capítulos de livros;
- **CE 3**: Trabalhos que não sejam no idioma português ou inglês;
- **CE 4**: Trabalhos que não estejam totalmente publicados. Ex. *drafts*;
- **CE 5**: Artigos retornados nas combinações que não envolviam sistemas *fuzzy*.

Por outro lado, os CIs utilizados para responder às questões de pesquisa são:

- **CI 1 - Título**: Artigos que tratam sobre *machine learning* com diabetes ou ainda predição ou classificação utilizando *fuzzy* para diabetes;
- **CI 2 - Resumo**: Artigos que explicam sobre a utilização de sistemas *fuzzy* na predição de dados de exames de sangue? ou que explicam sobre a utilização de sistemas *fuzzy* na classificação de dados de exames de sangue? ou que descrevem métodos de predição ou classificação de dados de exames de sangue? ou que apresentam predição ou classificação em base de dados de exames de sangue?
- **CI 3 - Introdução e Conclusão**: Os objetivos estão claramente definidos? ou os métodos de predição ou classificação estão claramente definidos? o estudo apresenta a ideia de utilização de *fuzzy* na predição ou classificação de exames de sangue? o estudo apresenta a base de dados de exame de sangue utilizada?

[4]https://ieeexplore-ieee-org.ez40.periodicos.capes.gov.br/Xplore/guesthome.jsp
[5]https://dl-acm-org.ez40.periodicos.capes.gov.br/
[6]https://www-scopus.ez40.periodicos.capes.gov.br/search/form.uri?display=basic#basic
[7]https://link-springer-com.ez40.periodicos.capes.gov.br/

3.2.5 Selecionando Trabalhos

Nesta subseção são apresentados os dados relacionados aos CEs e CIs, bem como a quantidade de artigos que foram sendo retornados a partir da aplicação de cada um dos critérios elencados anteriormente.

O primeiro passo consiste na realização da pesquisa de cada uma das combinações de palavra-chave destacadas anteriormente em cada uma das bases selecionadas, sem considerar, no primeiro momento, nenhum CE. Na Tabela 7 são retornados, para cada uma das OPs, o total de artigos. Por linha são consideradas as bibliotecas digitais e por colunas as etapas do processo de revisão (ver Figura 5).

Em seguida, optou-se pela aplicação do CE relacionado ao ano de publicação dos resultados retornados. Para este caso, portanto, foram considerados apenas os resultados publicados de 2017 até o presente momento. Tal CE fez com que o total de resultados retornados fosse menor, como mostra a coluna “CE 1” Tabela 7.

Os próximos passos consistiram na aplicação dos outros CEs. A coluna “CE 2 e 4” da Tabela 7 apresenta os resultados após a exclusão dos livros e capítulos de livros e também dos artigos que não estavam completamente publicados, enquanto a coluna “CE 3” da Tabela 7 apresenta os resultados após a aplicação do CE referente ao idioma de publicação do artigo.

Em resumo, a Tabela 7 apresenta para cada uma das fontes bibliográficas o total de artigos retornados para cada OP, bem como o total de artigos retornados por base de dados. Além disso, através da Tabela 7 pode-se observar que a cada CE aplicado o total de artigos retornados, não só por OP, mas também por fonte bibliográfica, diminuía.

Após avaliar os resultados retornados, verificou-se que um número muito grande de resultados havia sido retornado em combinações de palavras que não englobavam a palavra ”*fuzzy*”. Dessa forma, optou-se por retirar tais combinações da análise, uma vez que o maior foco deste trabalho consiste em encontrar trabalhos que utilizassem sistemas *fuzzy* no processo de predição ou classificação de problemas de saúde. Com isso, ao invés de 8 OPs, se passou a considerar somente 4, como mostra a Tabela 8.

Analisando-se a Tabela 8, pode-ser observar que a biblioteca *IEEE Xplore* foi a que retornou um menor número de artigos considerando todas as OPs. Por outro lado, a *SpringerLink* foi a biblioteca que retornou o maior número de resultados. Além disso, a OP que retornou mais artigos foi a 4, que considerava a relação de *“blood test” and “classification” and “fuzzy”* .

Após a etapa de aplicação de todos os CEs elencados anteriormente, o próximo passo consiste na aplicação dos CIs. Neste caso, são avaliados, inicialmente, o título de todos os trabalhos pré-selecionados. Caso o título não atendesse aos critérios estabelecidos, o artigo era rejeitado, caso atendesse, avançava para a próxima análise. Após a verificação do título dos 314 artigos que haviam restado, restaram 124 artigos. Vale ressaltar que nesse momento ainda não estavam sendo analisados se o mesmo artigo estava aparecendo

Tabela 7: Total de artigos retornados para cada um dos passos descritos

Fontes Bibliográficas	Inicial		CE 1		CE 2 e 4		CE 3	
IEEE Xplore	OP1: 30	OP5: 4	OP1: 20	OP5: 4	OP1: 18	OP5: 4	OP1: 18	OP5: 4
	OP2: 32	OP6: 1	OP2: 20	OP6: 1	OP2: 19	OP6: 1	OP2: 19	OP6: 1
	OP3: 2	OP7: 0	OP3: 2	OP7: 0	OP3: 2	OP7: 0	OP3: 2	OP7: 0
	OP4: 2	OP8: 0	OP4: 1	OP8: 0	OP4: 1	OP8: 0	OP4: 1	OP8: 0
	Total: 71		**Total: 48**		**Total: 45**		**Total: 45**	
SpringerLink	OP1: 9251	OP5: 596	OP1: 5052	OP5: 374	OP1: 2902	OP5: 190	OP1: 2892	OP5: 190
	OP2: 17494	OP6: 1355	OP2: 8812	OP6: 766	OP2: 5549	OP6: 527	OP2: 5345	OP6: 526
	OP3: 861	OP7: 106	OP3: 624	OP7: 81	OP3: 123	OP7: 9	OP3: 123	OP7: 9
	OP4: 991	OP8: 116	OP4: 679	OP8: 87	OP4: 143	OP8: 13	OP4: 143	OP8: 13
	Total: 30770		**Total: 16475**		**Total: 9456**		**Total: 9241**	
Scopus	OP1: 1321	OP5: 67	OP1: 698	OP5: 44	OP1: 672	OP5: 41	OP1: 636	OP5: 41
	OP2: 974	OP6: 148	OP2: 488	OP6: 75	OP2: 475	OP6: 65	OP2: 441	OP6: 65
	OP3: 7	OP7: 0	OP3: 6	OP7: 0	OP3: 6	OP7: 0	OP3: 6	OP7: 0
	OP4: 7	OP8: 1	OP4: 2	OP8: 0	OP4: 2	OP8: 0	OP4: 2	OP8: 0
	Total: 2525		**Total: 1313**		**Total: 1261**		**Total: 1191**	
ACM	OP1: 92	OP5: 13	OP1: 66	OP5: 6	OP1: 45	OP5: 4	OP1: 45	OP5: 4
	OP2: 152	OP6: 18	OP2: 69	OP6: 6	OP2: 49	OP6: 5	OP2: 49	OP6: 5
	OP3: 16	OP7: 2	OP3: 12	OP7: 0	OP3: 7	OP7: 0	OP3: 7	OP7: 0
	OP4: 23	OP8: 3	OP4: 12	OP8: 0	OP4: 8	OP8: 0	OP4: 8	OP8: 0
	Total: 319		**Total: 171**		**Total: 118**		**Total: 118**	
Total por OP	OP1: 10694	OP5: 680	OP1: 5836	OP5: 428	OP1: 3637	OP5: 239	OP1: 3591	OP5: 239
	OP2: 18652	OP6: 1522	OP2: 9389	OP6: 848	OP2: 6092	OP6: 598	OP2: 5854	OP6: 597
	OP3: 886	OP7: 108	OP3: 644	OP7: 81	OP3: 138	OP7: 9	OP3: 138	OP7: 9
	OP4: 1023	OP8: 120	OP4: 694	OP8: 87	OP4: 154	OP8: 13	OP4: 154	OP8: 13

Fonte: Autora.

Tabela 8: Total de artigos selecionados a partir da aplicação de todos os critérios de exclusão

OP	***IEEE Xplore***	***SpringerLink***	***Scopus***	***ACM Dig. Lib.***	**Total**
3	2	123	6	7	138
4	1	143	2	8	154
7	0	9	0	0	9
8	0	13	0	0	13
Total	3	288	8	15	314

Fonte: Autora.

em mais de uma combinação ou em mais de uma base de dados.

O próximo passo consistia na análise do resumo e a ideia era semelhante a do título, caso o resumo não apresentasse as informações desejadas o trabalho era descartado, caso contrário, se avaliava a introdução e a conclusão do trabalho. Após a aplicação de todos os CIs o total de artigos selecionados foi igual a 6, cujas informações de título, autores e ano de publicação são apresentadas na Tabela 9.

A seguir, os 6 artigos selecionados são discutidos, resumidamente, usando como norteador cada uma das questões de pesquisa consideradas. Cabe destacar ainda que um trabalho pode responder uma ou mais questões.

Porém, antes de iniciar a análise individual de cada um dos seis artigos selecionados optou-se por criar uma nuvem de palavras a partir destes artigos. A nuvem foi gerada considerando o texto completo de todos os artigos, ou seja, foram consideradas todas as seções de cada um e não apenas alguma seção específica como é feito em alguns casos. O objetivo da criação de uma nuvem de palavras consiste em observar quais palavras aparecem com maior frequência. Sendo assim, a ideia é observar quais os termos que mais apareceram em todos os artigos para que se possa observar se tais termos fazem ou não parte do escopo do estudo.

Para a construção da nuvem apresentada pela Figura 6 foram consideradas as palavras/termos que apareceram mais de cinquenta vezes, a fim de não deixar a nuvem tão poluída e permanecer somente com os termos que realmente se destacaram.

Tabela 9: Informações básicas dos artigos selecionados

Referência	**Título**	**Autores**	**Ano**	**Total de Citações**	**Diabetes**	***Fuzzy***
[56]	*Predicting lab values for gastrointestinal bleeding patients in the intensive care unit: A comparative study on the impact of comorbidities and medications*	Mahani, Golnar K and Pajoohan, Mohammad-Reza	2019	2	Não	Sim
[57]	*Medicine Recommendation System For Diabetes Using Prior Medical Knowledge*	Wedagu, Mulubrhan Ayalew and Chen, Dehua and Hussain, Muhammad Ather Iqbal and Gebremeskel, Tsegay and Orlando, Mayugi Tanguy and Manzoor, Arslan	2020	0	Sim	Não
[58]	*Adaptive neuro-fuzzy inference system (ANFIS) for rapid diagnosis of COVID-19 cases based on routine blood tests*	Deif, Mohanad and Hammam, Rania and Solyman, Ahmed	2021	9	Não	Sim
[36]	*Diabetes Prediction Recommender System based on Artificial Neural Networks and Sine-Cosine Optimization Algorithm*	Faraji-Biregani, Maryam and Nematbakhsh, Nasser	2019	1	Sim	Não
[1]	*Liver Disease Prediction Using Machine Learning Algorithms*	Kalaiselvi, R and Meena, K and Vanitha, V	2021	1	Não	Sim
[59]	*Prediction of Diabetes and Clustering Based on its Levels using Fuzzy C Means Algorithm*	S. Jamuna, Dr. K. Mohan Kumar	2020	1	Sim	Sim

Fonte: Autora.



Figura 6: Nuvem de palavras montada a partir dos seis artigos selecionados
Fonte: Autora.

3.2.6 Trabalhos Selecionados

Diferente do que aconteceu no capítulo anterior, onde foi realizada apenas uma abordagem geral sobre os trabalhos, nesta revisão sistemática optou-se por apresentar de forma individual cada um dos trabalhos selecionados. Isto se deve ao fato que a descrição das metodologias/ferramentas utilizadas no processo de predição e classificação de doenças é extremamente importante, não só para a correta compreensão do estado da arte, mas também para auxiliar no processo de encontrar qual metodologia melhor se encaixa no estudo que está sendo desenvolvido neste trabalho.

3.2.6.1 Adaptive Neuro-Fuzzy Inference System (ANFIS) for rapid diagnosis of COVID-19 cases based on routine blood tests

O estudo apresentado por [58] tem como objetivo utilizar um ANFIS para detectar rapidamente casos de COVID-19 a partir de exames laboratoriais comumente disponíveis. Além disso, o estudo também apresenta a identificação das variáveis clínicas que mais influenciaram na decisão. Dessa forma, este estudo responde a questão de pesquisa 1, visto que utiliza da metodologia *fuzzy* para predizer, a partir de exames de sangue, se o paciente esta ou não com COVID-19.

O *Neuro-Fuzzy Inference System* (NFIS) surge da combinação do *fuzzy logic system* (FLS) e das Redes Neurais Artificiais (RNA), tal combinação propicia a produção de

resultados significativos no que tange a detecção de casos de COVID-19. Este estudo apresenta uma metodologia chamada de sistemas de inferência *fuzzy* baseados em redes adaptáveis (ANFIS) que é utilizada para desvendar problemas de reconhecimento de parâmetros.

ANFIS consiste em uma representação para a rede gráfica do sistema de inferência *fuzzy* do tipo *Takagi-Sugeno* que possui as habilidades de aprendizado neural. A arquitetura do ANFIS é apresentado pela Figura 7, onde os nós fixos são representados pelos nós circulares, enquanto os nós que possuem parâmetros a serem aprendidos são representados pelos nós quadrados.

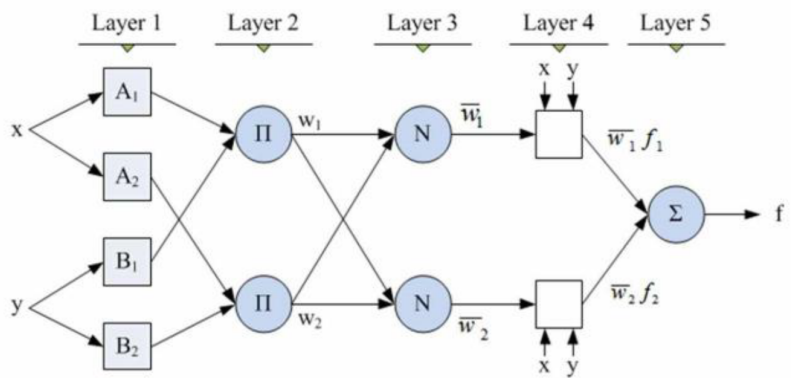


Figura 7: Arquitetura ANFIS
Fonte: [58]

Pode-se observar que a arquitetura possui 5 camadas (*layers)* como mostra a figura, são elas:

- **Camada 1:** Conhecida como camada de fuzzificação, a saída desta camada são os graus de pertinência das entradas, baseado na premissa de cada regra;
- **Camada 2:** A camada das regras, como é chamada, consiste na camada onde calcula-se o grau de pertinência ao qual é submetido o consequente de cada regra;
- **Camada 3:** Esta camada realiza a normalização dos graus de ativação das regras. Cada nó desta camada executa a função normalização, a qual é utilizada como um pré-processamento para a defuzzificação;
- **Camada 4:** Nesta camada, a saída de cada neurônio é calculada pelo produto da saída normalizada da camada anterior e o grau de ativação do consequente;
- **Camada 5:** O nó nesta camada calcula a saída total como o número de todos os sinais de entrada.

Dessa forma, seguindo a arquitetura apresentada, no estudo é definido um método de aprendizado de máquina como uma abordagem computacional para a detecção rápida de

pacientes com COVID-19 que aprende com um conjunto predefinido de recursos e passa por várias etapas diferentes. Tais etapas, apresentadas pela Figura 8, são: (1) Preparação do conjunto de dados; (2) Seleção de recursos; (3) Classificação e (4) Interpretação das previsões do modelo.

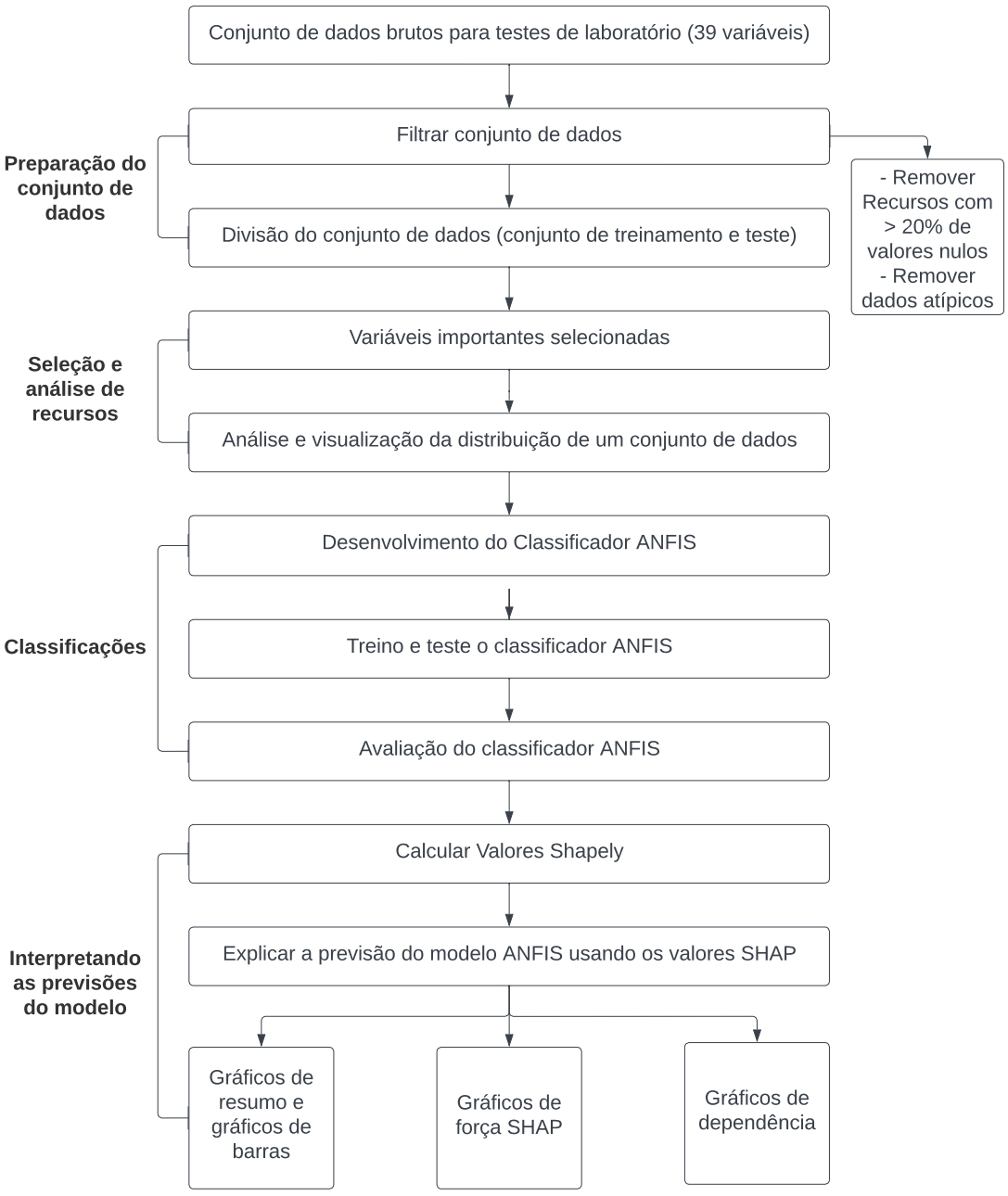


Figura 8: Etapas gerais da abordagem proposta
Fonte: Traduzido de [58]

O classificador do modelo ANFIS apresentava capacidade de discriminar pacientes positivos e negativos para COVID-19 com taxas de precisão de 95%, sensibilidade de 75% e especificidade de 97,25%.

Os resultados revelaram que pacientes com baixas contagens de leucócitos e plaquetas

são mais propensos a serem infectados pela COVID-19 do que pacientes que apresentavam valores mais altos. Sendo assim, o estudo concluiu que as contagens de leucócitos e plaquetas são capazes de atuar como indicadores relevantes e significativos para o diagnóstico de pacientes com COVID-19.

3.2.6.2 *Diabetes Prediction Recommender System based on Artificial Neural Networks and Sine-Cosine Optimization Algorithm*

Segundo o estudo apresentado por [36], a diabetes além de se tratar de uma das doenças mais comuns no mundo, ainda não possui cura. O *fasting blood sugar* (FBS) é um dos indicadores mais importantes de diabetes, porém a realização do seu teste não é viável para o público e requer preparativos antes da realização. Dessa forma, encontrar estratégias que possam auxiliar no monitoramento desse indicador acabam por ser extremamente importantes para o controle e previsão da diabetes.

Pensando na importância de se encontrar soluções para controlar e até mesmo prever a diabetes foi que este estudo extraio um conjunto de dados de testes de pacientes diabéticos e fez uso de um algoritmo de otimização seno-cosseno para seleção de recursos e de redes neurais para prever o açúcar no sangue em jejum a fim de prevenir efeitos colaterais do diabetes.

No modelo proposto pelo estudo, o conjunto de dados é selecionado com todos os seus recursos extraídos para selecionar os melhores recursos usando o algoritmo de otimização seno-cosseno e o algoritmo KNN (*K-nearest neighbors*) como seleção de recursos do *wrapper*, apresentado pela Figura 9.

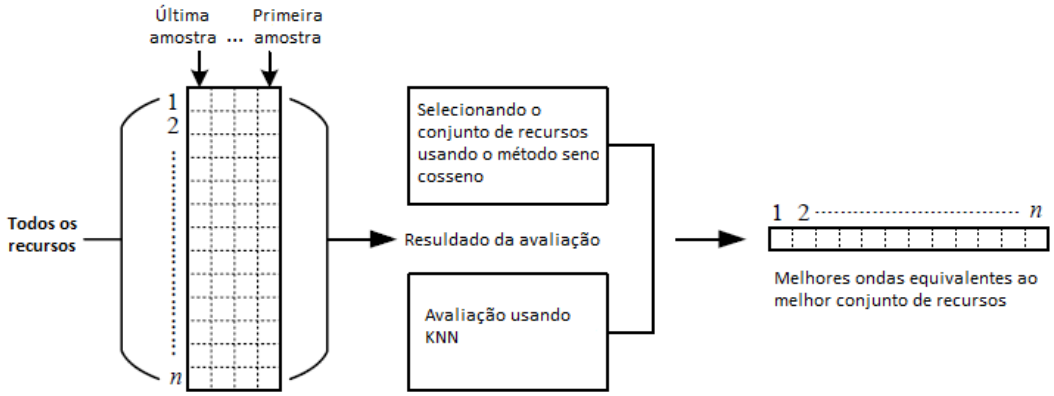


Figura 9: Modelo *wrapper* usando algoritmo de otimização seno-cosseno e método KNN.
Fonte: Adaptada de [36]

Neste estudo, três modelos NN com diferentes treinamentos foram utilizados. No treinamento com supervisão foi utilizada a rede neural *perceptron* multicamadas (PMC) [60], enquanto que no treinamento sem supervisão foram utilizados o *self-organized maps* (SOM) [61] e a *radial-basis function* (RBF) [62].

Em resumo, a partir do estudo desenvolvido, tenta-se dar um pequeno passo em

direção a uma melhor condição aos pacientes, para que se possa de alguma forma diagnosticar a doença a tempo, a fim de evitar seus efeitos colaterais, incluindo doenças oculares, doenças vasculares e etc. Os resultados do estudo comparados com outras pesquisas na previsão de diabetes sugerem que a rede neural proposta foi capaz de prever com precisão o nível de glicose no sangue em jejum a fim de predizer a diabetes.

O presente estudo foi um dos que melhor respondeu as questões de pesquisa, sendo capaz de responder as quatro questões de forma simultânea, uma vez que ele trata de um modelo de predição de diabetes, doença foco do trabalho que se pretende realizar, apesar de não utilizar lógica *fuzzy*, mas apresentou outras metodologias de predição.

3.2.6.3 Liver Disease Prediction Using Machine Learning Algorithms

O figado trata-se de uma das partes mais importantes do corpo humano, por isso encontrar metodologias que possam vir a monitorar e até controlar variáveis associadas ao seu funcionamento são extremamente importantes no combate de doenças [1]. O fígado pode ser facilmente afetado, seja pela ingestão constante de álcool, seja pelos hábitos alimentares. Porém, a identificação das doenças relacionadas ao fígado não acontece de maneira rápida e prática, uma vez que essa identificação se da através da análise de relatório de exames de sangue de função hepática e/ou pela análise de varredura, o que acaba por tornar o processo mais longo e também mais caro.

Sendo assim, buscar metodologias que possam vir a reduzir o tempo de identificação e também o custo são importantes para o combate de doenças perigosas relacionadas ao fígado. Pensando nisso, o presente estudo visa utilizar diferentes algoritmos de mineração de dados para facilitar tais processos, diminuindo assim o tempo de diagnóstico de doenças hepáticas.

Dessa forma, o principal objetivo deste trabalho consiste em encontrar qual o algoritmo, entre: o KNN, DT (decision tree) e ANFIS; mais adequado para predição de doenças hepáticas com base nos fatores de exatidão, sensibilidade, especificidade e precisão. O diagrama apresentado pela Figura 10 apresenta o fluxo de trabalho de todo o modelo proposto neste estudo.

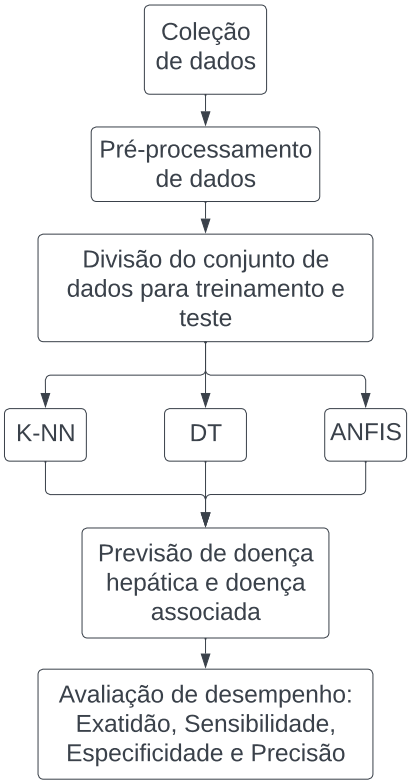


Figura 10: Fluxo de trabalho do modelo proposto por [1].
Fonte: Adaptada de [1]

Após a avaliação do desempenho dos três classificadores considerados pode-se concluir que o classificador ANFIS foi quem alcançou a maior exatidão (98%), a maior precisão (100%), a maior sensibilidade (97,4%) e maior especificidade (100%), como mostra a Tabela 10.

Tabela 10: Desempenho de cada algoritmo em relação as métricas propostas

Algoritmo	**Exatidão**	**Sensibilidade**	**Especificidade**	**Precisão**
KNN	66,6	76,61	42	76,62
DT	89,09	-	-	-
ANFIS	98	97,4	100	100

Fonte: [1]

Portanto, pode-se concluir após as análises, que o algoritmo ANFIS foi o que apresentou maior eficácia na predição de doenças hepáticas. Juntamente com o estudo apresentado anteriormente, este trabalho foi classificado como um dos que melhor respondeu as questões de pesquisa elencadas.

3.2.6.4 *Prediction os Diabetes and Clustering based on its levels using Fuzzy C Means Algorithm*

O estudo apresentado por [59] tem como principal objetivo analisar os níveis de diabetes e agrupá-los usando o algoritmo FCM (*Fuzzy C Means*). Dessa forma, a ideia consiste em detectar os níveis de diabetes e classificar os pacientes de acordo com estes níveis (não diabéticos, pré-diabéticos ou diabéticos).

A Figura 11 apresenta o modelo de estrutura utilizada para a predição de diabetes e agrupamento dos pacientes nos três níveis. Para realizar tal agrupamento o estudo utiliza o nível de glicose oriundo dos exames de sangue dos pacientes para realizar tal divisão. A comparação do nível de glicose no sangue foi realizada seguindo os dados apresentados pela Tabela 11.

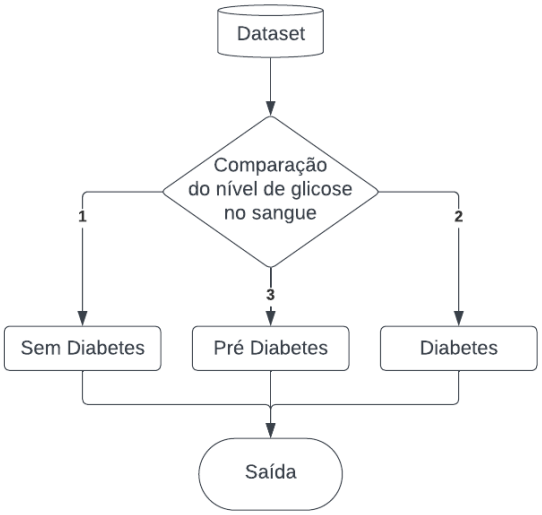


Figura 11: Modelo de estrutura proposto.
Fonte: Adaptada de [59]

Tabela 11: Tabela de glicose no sangue

Mg/DL	**Jejum**	**Depois de comer**	**2-3 horas depois de comer**
Normal	80-100	170-200	120-140
Glicose prejudicada	101-125	190-230	140-160
Diabético	126+	220-300	200+

Fonte: [63]

Sendo assim, concluí-se que este estudo apresenta de forma clara como prever os níveis de diabetes. Além disso, o estudo mostra como agrupar o conjunto de dados através da utilização do algoritmo FCM a fim de prever os pacientes que apresentam diabetes.

Assim como os dois últimos trabalhos, este estudo também foi considerado um dos três que melhor responderam as questões de pesquisa, uma vez que ele responde todos

os itens da questão de pesquisa 2, pois a partir de dados de exames de sangue, utilizando fuzzy, realiza a classificação dos pacientes de acordo com o nível de diabetes.

3.2.6.5 An ANFIS-PSO Algorithm for Predicting four grades of Non-Alcoholic fatty liver disease

O estudo apresentado por [64], assim como o apresentado anteriormente em [1], aborda a importância de se estudar metodologias que possam vir a auxiliar no diagnóstico e tratamento de doenças crônicas, principalmente aquelas relacionadas ao fígado.

O estudo foi implementado em duas etapas: na primeira delas, o conjunto de dados foi utilizado para extrair um conjunto de regras a fim de prever o resultado em *clusters* de paciente ou não paciente; já na segunda etapa, o modelo extraiu as regras e classificou os dados em quatro grupos com base nos graus de doença hepática e sua gravidade. Para realizar tais análises e classificações foram utilizadas as informações somente dos exames de sangue ao invés de utilizar imagens de ultrassom.

O sistema de inferência *neurofuzzy* adaptativo, apresentado anteriormente no trabalho de [58], também foi utilizado nesse estudo. Porém, este modelo foi integrado ao método de *particle swarm optimization* (PSO) a fim de melhorar a precisão e diminuir os erros. O algoritmo PSO consiste em um método de otimização que visa melhorar o resultado final da estrutura baseada em fatores definidos. Dessa forma, com a integração dos dois modelos, o estudo fez uso do que chamou de Modelo ANFIS-PSO, tendo como grande desafio encontrar os valores exatos para os parâmetros a fim de que se conseguisse obter um melhor desempenho quando comparado com o modelo ANFIS.

O conjunto de dados coletado contou com a ajuda de um conjunto de especialistas em fígado, que selecionaram 7 características principais que afetam o diagnóstico da doença hepática e sua gravidade, são elas:

- Peso;
- TGO;
- TGP;
- Fosfotase Alcalina (ALP);
- FBS;
- Plaquetas;
- Colesterol.

Com base em tais características, primeiramente, o conjunto de dados foi dividido em 2 grupos: saudáveis (100) e pacientes (300). Em seguida, o estudo caracterizou os dados com base nos graus de distúrbios hepáticos, incluindo: grau 1, grau 2 e grau 3.

A fim de verificar se o desempenho do modelo ANFIS-PSO foi superior ao desempenho do modelo ANFIS foi utilizada a medida *Root Mean Squared Error* (RMSE). A Tabela 12 apresenta os resultados obtidos para os dois tipos de classificação: paciente ou saudável e graus de doença hepática.

Tabela 12: Tabela comparativa entre ANFIS e ANFIS-PSO a partir da métrica RMSE

Algoritmo	**ANFIS**	**ANFIS-PSO**
Paciente ou Saudável	0,3054	0,15441
Graus de Doença Hepática	0,5048	0,3712

Fonte: [64]

Por fim, pode-se concluir que o modelo ANFIS-PSO apresentou resultado melhores no que tange as classificações realizadas. Destaca-se ainda o fato do estudo ter respondido de forma satisfatória as questões de pesquisa 1 e 2, uma vez que ele trabalha com a um sistema de predição de exames laboratoriais.

3.2.6.6 Medicine recommendation system for diabetes using prior medical knowledge

Aplicativos de recomendação estão sendo desenvolvidos e/ou aperfeiçoados com o objetivo de ajudar pacientes e médicos. Através destes aplicativos se tem a ideia de que se possa melhorar os serviços ofertados, reduzindo erros e economizando tempo [57].

Problemas como prescrição errada de medicamentos acabam por acarretar em um número significativo de mortes, o que acaba por fazer com que se busque alternativas que possam minimizar tais erros. Além disso, quando o assunto está relacionado a recomendação de medicamentos para diabetes, os problemas também são sérios e podem acarretar em problemas graves. Por isso, considerando as dificuldades elencadas, o estudo apresentado por [57] visa propor um novo método de recomendação de medicamentos para diabetes, denominado DIMERS, que integra conhecimentos médicos prévios ou totais com o modelo BiLSTM (*Bidirectional short-term memory*)).

O DIMERS consiste em dois módulos: Módulo de Recomendação e *Knowledge representation module* (KRM). Sendo assim, o sistema proposto funciona da seguinte forma: primeiramente, o módulo de representação de conhecimento extrai características do conhecimento médico prévio (por exemplo, banco de medicamentos e CID 10), em seguida, os prontuários dos pacientes e a matriz de rotulagem são conectadas como entrada do bloco ponderado. Depois disso, um atalho entre o peso que é saída do bloco e o registro do paciente são gerados como recursos. Por fim, o módulo de recomendação incorpora os recursos mencionados nas redes BiLSTM para recomendar os medicamentos apropriados para pacientes diabéticos.

O estudo apresenta um comparativo entre o DIMERS e outros modelos de recomendação de medicamentos que não consideram o conhecimento médico prévio, destacando o fato de que o modelo apresentado no estudo apresenta resultados ainda mais

satisfatórios do que se esperava no que tange a recomendação de medicamentos para diabetes.

3.2.6.7 Considerações sobre os trabalhos estudados

Após a análise de todos os trabalhos selecionados, destaca-se o fato de que as questões que melhor foram respondidas estão relacionadas as metodologias de predição e classificação utilizadas na área da saúde. Em resumo, pode-se concluir que dos 6 artigos selecionados: 4 deles responderam a questão de pesquisa 1, 5 a questão 2, 1 a questão 3 e 2 a questão 4.

3.3 Revisão 03: Metodologia ANFIS para predição e/ou classificação de doenças

A metodologia utilizada, assim como nas revisões sistemáticas anteriores, foi baseada no método apresentado por [32] e está sumarizada no fluxograma apresentado pela Figura 12. Neste fluxograma, os artigos vão sendo excluídos a medida que o processo vai evoluindo, a fim de que ao final restem somente aqueles artigos que melhor se encaixam nos critérios pré-estabelecidos.

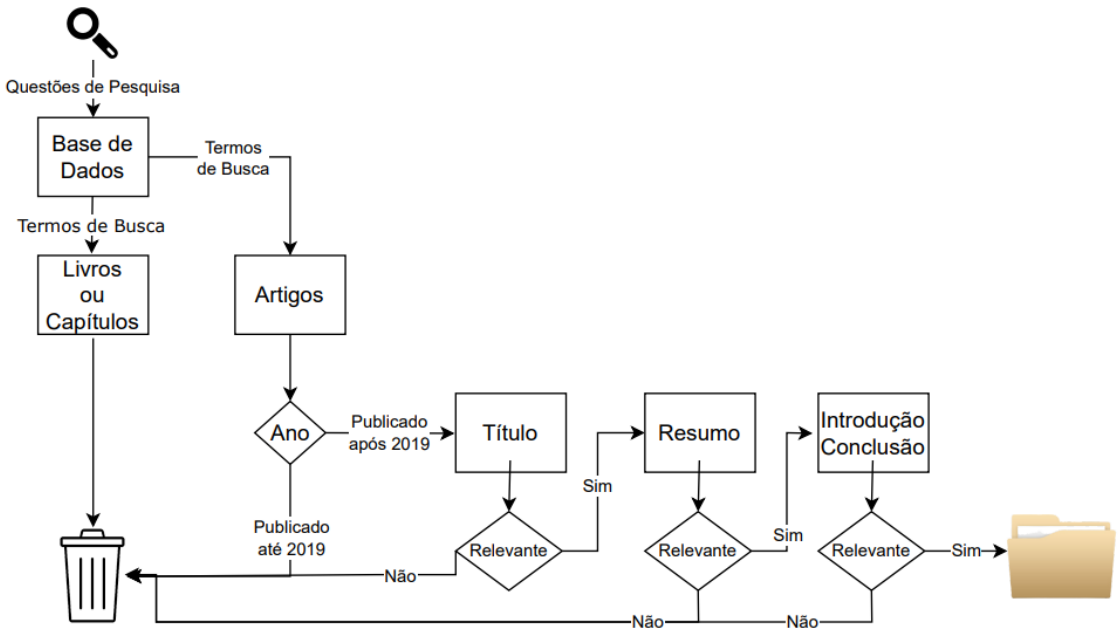


Figura 12: Fluxograma da metodologia da revisão teórica da metodologia ANFIS
Fonte: Autora.

3.3.1 Questões de Pesquisa

As questões de pesquisa levantadas para essa revisão sistemática são:

- **Questão 1**: Quais são as metodologias mais utilizadas do modelo ANFIS?

- **Questão 2**: Quais são as metodologias mais utilizadas do modelo ANFIS na classificação de doenças?
- **Questão 3**: Quais as metodologias *fuzzy* mais utilizadas para a classificação de doenças?
- **Questão 4**: Quais as metodologias mais utilizadas para a classificação de doenças?

A partir destas questões de pesquisa, o próximo passo consiste no levantamento dos termos de busca.

3.3.2 Termos de Busca

Os termos de busca, assim como nos casos anteriores, foram utilizados somente em inglês, por se tratar da língua mais utilizada em artigos.

Inicialmente, foram definidas as palavras-chave e, em seguida, estas palavras foram combinadas e separadas em 4 combinações.

As palavras-chave selecionadas são:

- *Fuzzy*;
- ANFIS;
- *disease*;
- *classification*.

A partir delas, foram realizadas as combinações apresentadas pela Tabela 13.

Tabela 13: Combinações de palavras-chave utilizadas para revisão sistemática de ANFIS

OP	Combinação
1	"*Fuzzy*" *and* "*ANFIS*" *and* "*classification*"
2	"*Fuzzy*" *and* "*ANFIS*" *and* "*classification*" *and* "*disease*"
3	"*Fuzzy*" *and* "*classification*" *and* "*disease*"
4	"*classification*" *and* "*disease*"

Fonte: Autora.

3.3.3 Bases utilizadas

As bases de dados consideradas para esta revisão sistemática foram:

- *PubMed/MedLine*[8];
- *IEEE Xplore*[9];

[8]https://pubmed-ncbi-nlm-nih-gov.ez40.periodicos.capes.gov.br/
[9]https://ieeexplore-ieee-org.ez40.periodicos.capes.gov.br/Xplore/guesthome.jsp

- *Scopus*[10];
- *ACM Digital Library*[11];

3.3.4 Critérios de Exclusão/Inclusão

Os CEs definidos para essa revisão sistemática foram:

- **CE 1**: Artigos publicados antes de 2019;
- **CE 2**: Trabalhos que não sejam no idioma português ou inglês;
- **CE 3**: Livros e capítulos de livros;
- **CE 4**: Trabalhos que não estejam totalmente publicados. Ex. *drafts*.

Por outro lado, os CIs definidos foram:

- **CI 1 - Título**: Artigos que tratam sobre metodologias que utilizam o modelo ANFIS ou metodologias *fuzzy* para classificação de doenças;
- **CI 2 - Resumo**: Artigos que explicam sobre a metodologia que utilizam o modelo ANFIS ou que explicam a utilização de metodologias *fuzzy* para classificação de doenças;
- **CI 3 - Introdução e Conclusão**: Os objetivos estão claramente definidos? ou os métodos *fuzzy* ou ANFIS estão claramente definidos? o estudo apresenta a ideia de utilização de *fuzzy* ou ANFIS na predição ou classificação de doenças?

3.3.5 Filtrando Trabalhos

Definidos as combinações de palavras-chave, bem como as bases de dados onde as pesquisas vão ser realizadas, a próxima etapa consiste em procurar cada uma das combinações em cada base de dados e, em seguida, aplicar os critérios de exclusão e inclusão para restringir as buscas.

A primeira coluna da Tabela 14 apresenta, para cada uma das OPs, o total de artigos encontrados em cada base de dados. Sendo que nas linhas são apresentadas as bases de dados e nas colunas as etapas do processo de revisão.

Em seguida, optou-se pela aplicação do CE relacionado ao ano de publicação dos resultados retornados. Para este caso, portanto, foram considerados apenas os resultados publicados de 2019 até o presente momento. Tal CE fez com que o total de resultados retornados fosse menor, como mostra a coluna "CE 1" Tabela 14.

[10] https://www-scopus.ez40.periodicos.capes.gov.br/search/form.uri?display=basic#basic
[11] https://dl-acm-org.ez40.periodicos.capes.gov.br/

Os próximos passos consistiram na aplicação dos outros CEs. A coluna “CE 2” da Tabela 14 apresenta os resultados após a exclusão dos artigos que não estavam nem em português nem em inglês; a coluna “CE 3”apresenta o total de artigos após a exclusão dos livros e capítulos de livros; por fim, a coluna “CE 4”retorna o total após a exclusão dos artigos que não estavam completamente publicados.

Em resumo, a Tabela 14 apresenta para cada uma das fontes bibliográficas o total de artigos retornados para cada OP, bem como o total de artigos retornados por base de dados. Além disso, através da Tabela 14 pode-se observar que a cada CE aplicado o total de artigos retornados, não só por OP, mas também por fonte bibliográfica, diminuía.

Após avaliar os resultados retornados, verificou-se que um número muito grande de resultados havia sido retornado nas combinações de palavras que não englobavam a palavra ”*fuzzy*”ou a palavra ANFIS. Dessa forma, optou-se por retirar tais combinações da análise por se entender que sem estes termos a pesquisa ficava muito parecida com a revisão realizada e apresentada anteriormente. Com isso, ao invés de 4 OPs, se passou a considerar somente 2, como mostra a Tabela 15.

Após a etapa de aplicação de todos os CEs elencados anteriormente, o próximo passo consiste na verificação de artigos duplicados. Neste caso, o número que antes era de 279, passou a ser de 215.

Em seguida, foram aplicados os CIs nestes 215 artigos que restaram. Neste caso, são avaliados, inicialmente, o título de todos os trabalhos pré-selecionados. Caso o título não atendesse aos critérios estabelecidos, o artigo era rejeitado, caso atendesse, avançava para a próxima análise. Após a verificação do título dos 215 artigos que haviam restado, restaram 207 artigos.

O próximo passo consistia na análise do resumo e a ideia era semelhante a do título, caso o resumo não apresentasse as informações desejadas o trabalho era descartado, caso contrário, se avaliava a introdução e a conclusão do trabalho. Após a aplicação de todos os CIs o total de artigos selecionados foi igual a 129. A Figura 13 apresenta o resumo através de um funil onde são apresentados os totais de artigos que restaram após cada uma das etapas.

Dessa forma, a próxima etapa consiste em ler e avaliar os 129 artigos selecionados, porém, ao realizar a busca pelos artigos na íntegra, não foi possível encontrar todos os artigos disponíveis de forma legítima, fazendo com que o número total de artigos caísse de 129 para 99.

3.3.6 Trabalhos Selecionados

O número de artigos, apesar de ter sido reduzido consideravelmente, ainda era considerado um número elevado para que fosse realizada a leitura detalhada de um por um. Por isso, optou-se por realizar primeiramente uma análise mais ampla de todos eles, focando em alguns itens específicos como:

Tabela 14: Total de artigos retornados para cada um dos passos descritos

FBs	Inicial	CE 1	CE 2	CE 3	CE 4
IEEE Xplore	OP1: 851 OP3: 1366 OP2: 129 OP4: 28818 **Total: 31164**	OP1: 362 OP3: 633 OP2: 67 OP4: 20098 **Total: 21160**	OP1: 362 OP3: 633 OP2: 67 OP4: 20098 **Total: 21160**	OP1: 362 OP3: 631 OP2: 67 OP4: 20043 **Total: 21103**	OP1: 362 OP3: 624 OP2: 67 OP4: 19918 **Total: 20971**
Scopus	OP1: 1375 OP3: 3874 OP2: 294 OP4: 542671 **Total: 548214**	OP1: 609 OP3: 2143 OP2: 161 OP4: 186992 **Total: 189905**	OP1: 603 OP3: 2120 OP2: 160 OP4: 178370 **Total: 181253**	OP1: 533 OP3: 1772 OP2: 144 OP4: 146889 **Total: 149338**	OP1: 524 OP3: 1727 OP2: 142 OP4: 14515 **Total: 147550**
PubMed/ Medline	OP1: 117 OP3: 410 OP2: 26 OP4: 225500 **Total: 226053**	OP1: 52 OP3: 186 OP2: 14 OP4: 60490 **Total: 60742**	OP1: 51 OP3: 184 OP2: 13 OP4: 58892 **Total: 59140**	OP1: 51 OP3: 184 OP2: 13 OP4: 58892 **Total: 59140**	OP1: 51 OP3: 184 OP2: 13 OP4: 58672 **Total: 58920**
ACM Dig. Lib.	OP1: 163 OP3: 1418 OP2: 66 OP4: 12291 **Total: 13938**	OP1: 106 OP3: 964 OP2: 57 OP4: 6899 **Total: 8026**	OP1: 106 OP3: 964 OP2: 57 OP4: 6899 **Total: 8026**	OP1: 106 OP3: 510 OP2: 57 OP4: 5950 **Total: 6623**	OP1: 106 OP3: 509 OP2: 57 OP4: 5928 **Total: 6600**
Total por OP	OP1: 2506 OP3: 7068 OP2: 515 OP4: 8092801	OP1: 1129 OP3: 3926 OP2: 299 OP4: 274479	OP1: 1122 OP3: 3901 OP2: 197 OP4: 264259	OP1: 1052 OP3: 3097 OP2: 281 OP4: 231774	OP1: 1043 OP3: 3044 OP2: 279 OP4: 22967

Fonte: Autora.

Tabela 15: Total de artigos selecionados a partir da aplicação de todos os critérios de exclusão

OP	*IEEE Xplore*	*Scopus*	*PubMed/Medline*	*ACM Dig. Lib.*	**Total**
2	67	142	57	13	279

Fonte: Autora.

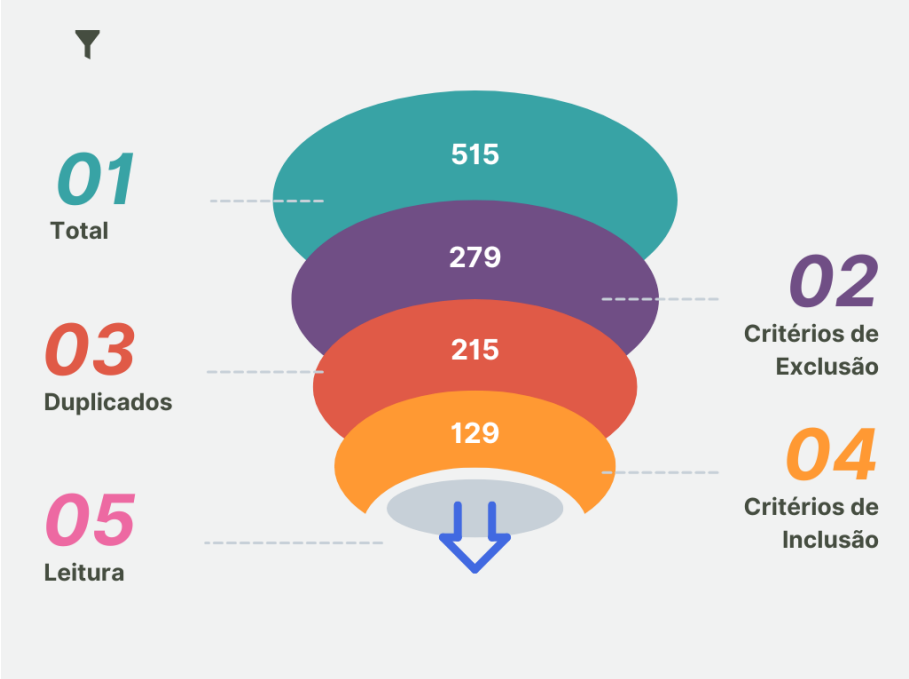


Figura 13: Resumo geral referente ao total de artigos após todas as etapas
Fonte: Autora.

- Tipos de Dados: se o artigo trabalhava com imagem, texto ou som;
- Disponibilidade do Código: se o artigo disponibilizada o código ou pelo menos o pseudocódigo para que pudesse ser utilizado como base;
- Tipo de Doença: qual a doença que o artigo tratava;
- Métodos Utilizados: se o artigo utilizava ANFIS, alguma variação dele ou até mesmo a combinação dele com algum outro método.

Dessa forma, todos os 99 artigos foram analisados de forma superficial a fim de responder tais itens. Para facilitar a visualização das informações obtidas optou-se por apresentar as mesmas em forma de gráficos. A Figura 14 apresenta o gráfico que relaciona a quantidade de artigo para cada um dos tipos de dados; a Figura 15 apresenta a quantidade de artigos que disponibiliza ou não o código ou ainda o pseudocódigo; a Figura 16, por sua vez, relacionada a quantidade de artigo para cada uma das doenças em questão; e, por fim, a Figura 17 relaciona os métodos utilizados (ou a combinação deles) com a quantidade de artigos.

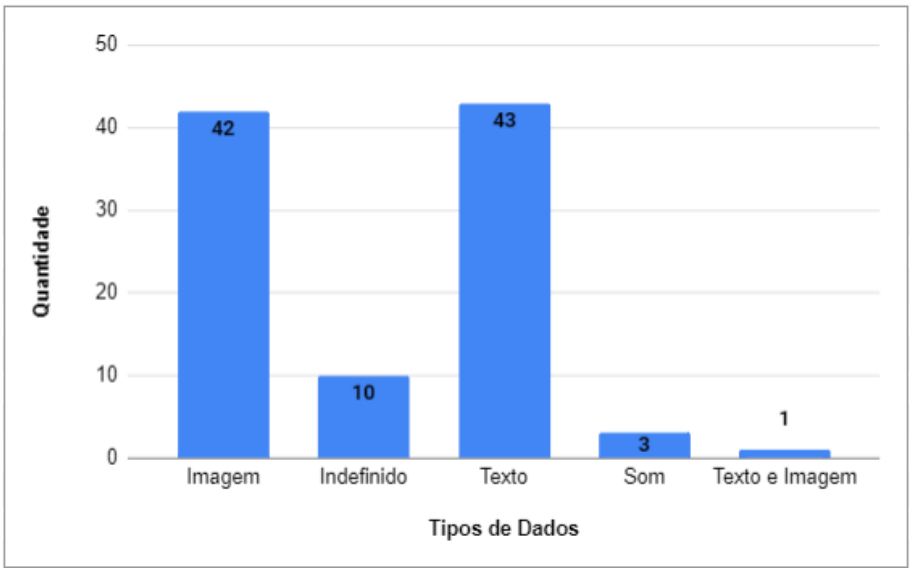


Figura 14: Gráfico que relaciona a quantidade de artigos pelo tipo de dado
Fonte: Autora.

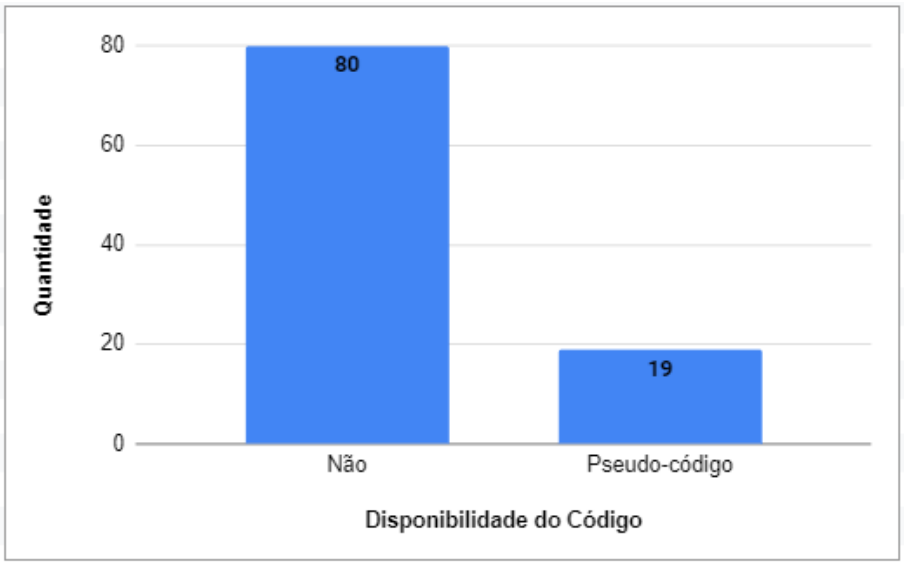


Figura 15: Gráfico que relaciona a quantidade de artigos pela disponibilidade ou não do código

Fonte: Autora.

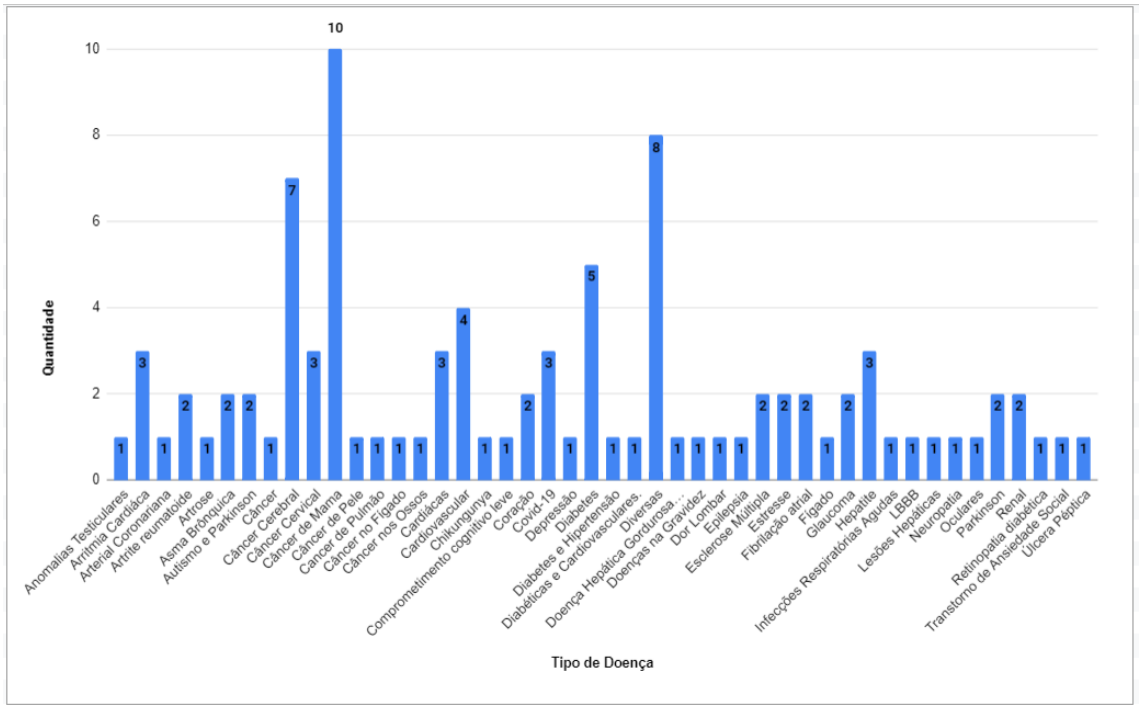


Figura 16: Gráfico que relaciona a quantidade de artigos pelo tipo da doença
Fonte: Autora.

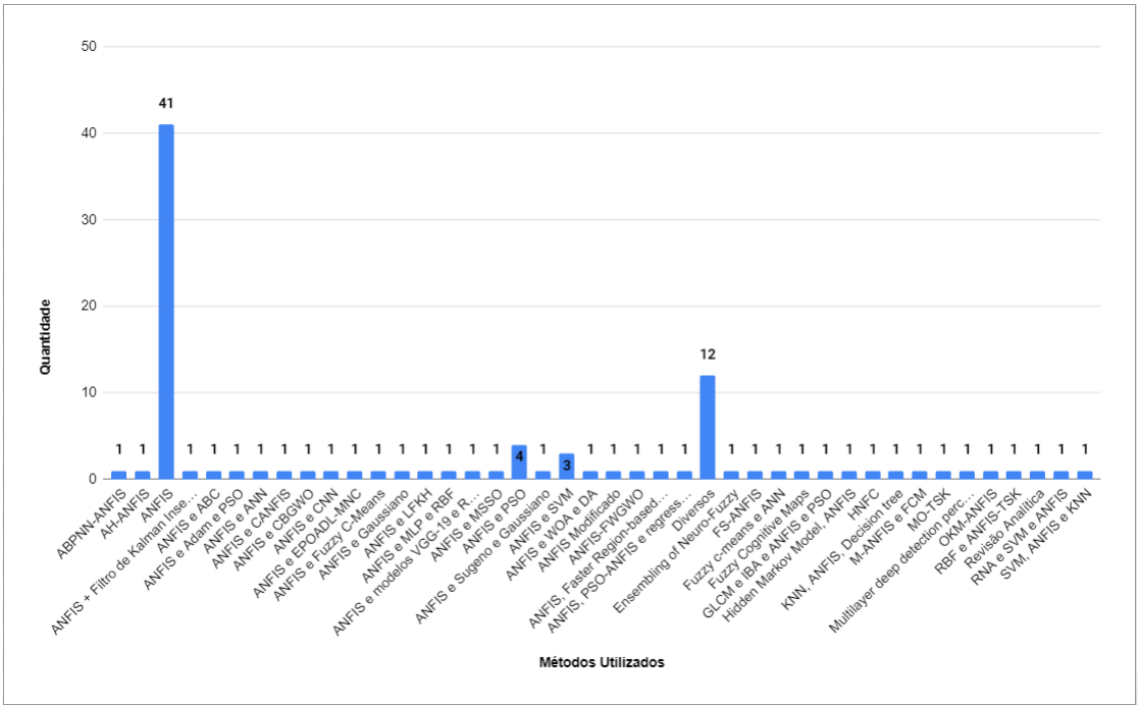


Figura 17: Gráfico que relaciona a quantidade de artigos pelos métodos utilizados em cada um deles

Fonte: Autora.

No que tange à disponibilidade ou não do código pode-se perceber que de modo geral são poucos os artigos que disponibilizam informações do código, sejam elas o código propriamente dito ou até mesmo um contato para obtenção do código caso seja de interesse do leitor. Porém, alguns deles disponibilizam o pseudocódigo de partes do estudo, o que acaba ajudando mas não torna um diferencial daquele estudo. Por isso, na hora de restringir os artigos a serem melhor explorados não foi considerado o quesito do código, uma vez que não há nenhum que se destaque em relação aos outros neste quesito.

Como era esperado, devido as palavras-chave utilizadas na revisão sistemática, a grande maioria dos artigos utiliza o método ANFIS, alguma variação do mesmo ou ainda a combinação dele com alguns outros métodos. Por isso, qualquer artigo que seja selecionado para ser melhor explorado será capa, de alguma forma, de apresentar a maneira como o método foi utilizado para classificar ou predizer alguma doença. Dessa forma, assim como para o caso do código, este quesito não será determinante na escolha dos artigos que serão selecionados para serem estudados de forma mais aprofundada.

No que tange ao tipo de dado, a grande maioria dos artigos trabalha com imagem ou texto, considerando que o interesse deste estudo consiste em trabalhar com dados textuais, optou-se por restringir a análise dos artigos aos que trabalham com dados textuais somente. Sendo assim, o valor total de artigos se restringe a 43 que trabalham somente com texto e 1 que trabalha com texto e imagem. A partir da restrição de 44 artigos impostas pelo tipo de dado que o artigo trabalha, o próximo passo consistiu em analisar o tipo de doença destes artigos.

Pensando primeiramente nos 99 artigos pré-selecionados, pode-se observar que a pesquisa realizada foi ampla no que tange ao tipo de doença, uma vez que se optou por pesquisar, inicialmente, qualquer doença. Porém, com o afunilamento do trabalho e com a definição de se abordar estratégias para a classificação de diabetes, optou-se por focar nos trabalhos que abordavam essa doença. Neste caso, analisando o gráfico apresentado pela Figura 16, observa-se que somente 5 trabalhos abordam somente diabetes, 1 trabalho fala sobre diabetes e hipertensão e 1 trabalho sobre diabetes e doenças cardiovasculares. Portanto, um total de 7 trabalhos aborda a classificação de diabetes. Além disso, estes 7 trabalhos selecionados trabalham com dados textuais, ou seja, estão dentro da restrição dos 44 artigos.

Sendo assim, após análise dos 4 itens elencados como importantes a serem avaliados nos artigos pré-selecionados pode-se chegar a um valor de 7 artigos para serem melhor explorados. A seguir, um resumo sobre estes trabalhos é mostrado, a fim de apresentar de forma resumida o que se tirou de mais importante de cada um deles focando em especial no objetivo, na metodologia, nos resultados e limitações de cada um destes estudos.

3.3.6.1 A proposed approach for diabetes diagnosis using neuro-fuzzy technique [65]

Este estudo propõe uma abordagem *neuro-fuzzy*, baseada em ANFIS, para melhorar a acurácia no diagnóstico de DM2 usando o *Pima Indians Diabetes Dataset (PIDD)*. Utiliza o PIDD (768 instâncias) e segue três etapas:

- (i) pré-processamento, com normalização, imputação de valores ausentes e detecção de *outliers* via *Local Outlier Factor* (LOF);
- (ii) classificação com ANFIS treinado por algoritmo híbrido (*backpropagation* + *least squares*) para ajustar funções de pertinência e regras *fuzzy*;
- (iii) avaliação usando acurácia, sensibilidade e especificidade, comparando o modelo com redes neurais, *Naïve Bayes*, KNN, árvore de decisão e *random forest*.

O modelo ANFIS obteve sensibilidade de 97,65%, especificidade de 84,30% e acurácia de 92,77%, superando todos os classificadores comparados, cujas acurácias ficaram entre 70% e 75%.

Porém, este estudo limita-se ao fato de utilizar um *dataset* público (PIDD), com amostra exclusivamente de mulheres Pima, o que restringe a generalização. Além disso, utiliza métricas de desempenho sem análise de interpretabilidade clínica ou impacto em fluxos reais de atendimento. Por fim, os próprios autores acabam por sugerir a inclusão futura de métodos de otimização, *deep learning* e aplicação a outras doenças, reconhecendo que o modelo ainda é exploratório.

3.3.6.2 Design of an ANFIS based Decision Support System for Diabetes Diagnosis [66]

O estudo propõe o desenvolvimento de um sistema de apoio à decisão baseado em lógica *fuzzy* e ANFIS para diagnóstico precoce de diabetes *mellitus*, capaz de lidar com dados clínicos vagos e incertos. A base de dados utilizada é o PIDD.

Primeiro, é construído um sistema especialista *fuzzy* com 240 regras definidas com apoio de especialistas, considerando como entradas glicose, insulina, Índice de Massa Corporal (IMC), função de hereditariedade (DPF) e idade. A saída *fuzzy* representa a probabilidade de diabetes em cinco níveis (muito baixa a muito alta). Em seguida, o sistema é integrado a um modelo ANFIS de cinco camadas com funções de pertinência trapezoidais; 80% dos dados são usados para treino e 20% para teste. O treinamento usa algoritmo híbrido (*least squares* + *backpropagation*), e o desempenho é comparado a uma rede *Multilayer Perceptron* (MLP).

O ANFIS alcançou acurácia de 86,48%, sensibilidade de 92,78% e especificidade de 73,08%, superando ligeiramente a MLP que apresentou acurácia de 85%, sensibilidade de 99% e especificidade de 72,7%). O RMSE estabilizou em 0,21964 tanto no treino quanto no teste para diferentes números de épocas, sugerindo convergência do modelo.

O estudo é limitado a um único *dataset* pequeno (768 registros), sem validação cruzada ou externa; não há análise de *overfitting* além do RMSE; o conjunto de regras (240 regras) é complexo para implementação em ambiente clínico real e não é discutida a interpretabilidade das regras pelos profissionais de saúde. Além disso, a comparação é feita apenas com uma MLP, sem contraste com outros algoritmos modernos de *machine learning*.

3.3.6.3 Expert diagnostic system for detection of hypertension and diabetes mellitus using discrete wavelet decomposition of photoplethysmogram signal and machine learning technique [67]

Este estudo propõe desenvolver um sistema especialista automatizado, não invasivo, para estratificação de risco de hipertensão e detecção de DM2 a partir de sinais de fotopletismografia (PPG) de ponta de dedo.

O sistema opera em modo *online/offline*. O sinal PPG é pré-processado (filtragem, segmentação, correção de *baseline*) e avaliado por um índice de qualidade de sinal (SQI), sendo que apenas segmentos acima de um limiar são usados. Em seguida, aplica-se decomposição em 11 níveis por transformada *wavelet* discreta (DWT), da qual são extraídas 130 características estatísticas. Algoritmos de seleção de atributos (correlação de *Spearman, ReliefF* e mRMR) escolhem 20 melhores características para estratificação de hipertensão e detecção de DM2. Diferentes classificadores (incluindo MLP e ANFIS) são avaliados, com enfoque em ANFIS como sistema híbrido.

O ANFIS alcançou F1-score de 92%, 98,5% e 98,3% para as categorias NT vs pré-hipertensão, NT vs hipertensão estágio 1 e NT vs hipertensão estágio 2 (usando pressão sistólica), e 83,1%, 95,6% e 86,8% com base na pressão diastólica. Para a classificação de não diabético vs diabético, o ANFIS obteve acurácia de 99,1%. O uso de DWT mostrou-se robusto a ruídos, superando abordagens baseadas em características morfológicas do PPG.

O desempenho depende fortemente da qualidade do sinal PPG. Além disso, embora o SQI filtre sinais ruins, isso pode limitar o uso em dispositivos muito portáteis/vestíveis com grande variabilidade de ruído. A base de dados utilizada restringe a avaliação de generalização para populações diversas. Não há comparação com métodos de *deep learning* mais recentes, mencionados apenas na revisão. Além disso, o estudo utiliza apenas sinais PPG, sem integrar outros dados clínicos que poderiam enriquecer a interpretação clínica.

3.3.6.4 Improving Diagnosis Accuracy of Diabetic Disease Using Radial Basis Function Network and Fuzzy Clustering [68]

Este estudo tem como objetivo aumentar a acurácia do diagnóstico de diabetes tipo 2 usando técnicas de mineração de dados baseadas em computação suave, combinando

clustering fuzzy e RBF, além de um sistema ANFIS.

A base de dados utilizada é o PIDD (768 mulheres $\geq$ 21 anos, com 8 atributos clínicos). Primeiro, aplica-se um algoritmo de *reduced fuzzy clustering* que não exige definir a priori o número de *clusters*: cada ponto é tratado como centro potencial e seu "potencial" é calculado com base na densidade local, determinando centros de *clusters* automaticamente. Em seguida, uma rede neural RBF é treinada em MATLAB, usando parâmetros otimizados (número de neurônios, *spread*, etc.). Paralelamente, constrói-se um sistema ANFIS com funções de pertinência gaussianas, oito variáveis de entrada e três funções de pertinência por variável; o treinamento usa método híbrido e múltiplas iterações (600) para reduzir o erro.

Como resultado, observou-se que o modelo híbrido (*fuzzy clustering* + RBF) e o ANFIS alcançaram acurácia de 96,10% e 97,14%, respectivamente, com sensibilidade de 95,24% e 96,54% e especificidade de 96,86% e 97,56%. Esses resultados superam diversos modelos da literatura (redes neurais *feedforward*, ANFIS isolado, MLPNN, *mixture of experts*), indicando que a combinação de técnicas *fuzzy* aumenta significativamente o desempenho na detecção de diabetes.

Porém, assim como destacado em estudos anteriores, toda a avaliação é feita apenas no PIDD, sem validação externa ou análise de *overfitting* (por exemplo, via validação cruzada). Não são explorados aspectos de interpretabilidade das regras *fuzzy* no contexto clínico. Além disso, o foco é puramente em métricas de classificação, sem análise de tempo de processamento ou viabilidade de implantação em sistemas de informação em saúde.

3.3.6.5 Blockchain-Based Secure Healthcare Application for Diabetic-Cardio Disease Prediction in Fog Computing [69]

Este estudo propõe uma arquitetura de serviços de saúde segura baseada em *fog computing* e *blockchain* para predição de doenças cardiovasculares e diabetes, garantindo baixa latência, segurança e privacidade dos dados.

A arquitetura envolve cinco componentes: sensores médicos IoT (Internet das Coisas) que coletam sinais dos pacientes; nós *fog* que agregam, processam e encaminham dados; uma rede *blockchain* que armazena de forma distribuída e imutável os registros de saúde; *cloud* para armazenamento cifrado de longo prazo; e um medical *analyzer* autorizado para análise e predição de doenças. Os dados de saúde agregados nos nós *fog* são primeiramente clusterizados por um algoritmo de *rule-based clustering* proposto pelos autores; depois, aplica-se um sistema de classificação FS-ANFIS (*feature-selection-based* ANFIS) para prever diabetes e doenças cardíacas. O desempenho do clusterizador é avaliado com *Purity* e NMI; o preditor é avaliado por acurácia, comparando-o com outras redes neurais.

O sistema alcança mais de 81% de acurácia de predição, superando outros algoritmos de redes neurais testados pelos autores. Os indicadores de *Purity* e NMI demonstram que

o algoritmo de *rule-based clustering* produz grupos coerentes de pacientes, o que reforça o desempenho subsequente do FS-ANFIS.

Porém, o trabalho limita-se a uma prova de conceito, com poucos detalhes sobre o tamanho e a diversidade dos dados reais utilizados. Não há análise aprofundada de escalabilidade da *blockchain* em cenários com grande volume de transações médicas, nem de custos computacionais nos nós *fog*. A predição é avaliada apenas por acurácia, sem discussão de sensibilidade/especificidade para os diferentes desfechos clínicos. Também não é discutida a aceitação ou integração com fluxos de trabalho hospitalares reais.

3.3.6.6 NEIWDA with Hybrid Ensemble Model for Diagnosis of Diabetes Type-II [70]

Este estudo visa melhorar a predição de DM2 por meio de um modelo híbrido de *ensemble learning* com seleção de atributos baseada no algoritmo meta-heurístico *Natural Exponential Inertia Weight-based Dragonfly Algorithm* (NEIWDA).

O estudo também utiliza o PIDDD (768 pacientes, 268 positivos e 500 negativos). Primeiro, aplica-se o NEIWDA para seleção ótima de atributos, modelando o problema como uma busca por subconjuntos de características que maximizem a acurácia de classificação. O DA é adaptado com peso de inércia exponencial natural para equilibrar exploração/exploração do espaço de busca. Em seguida, é construído um *ensemble* que combina KNN, ANFIS, ANN e *Support Vector Machine* (SVM). As saídas desses classificadores são agregadas via *majority voting*. O desempenho do *ensemble* proposto é comparado com um modelo de *ensemble* anterior e com SVM isolado, avaliando acurácia, previsão, *recall* e *F-measure*.

Como resultado, pode-se observar que o modelo NEIWDA + *ensemble* alcançou acurácia de 98,80%, superando tanto o *ensemble* anterior quanto o SVM em todas as métricas (acurácia, precisão, *recall* e *F-measure*). Gráficos comparativos são apresentados e indicam ganho consistente de desempenho, sugerindo que a combinação de seleção de atributos com *ensemble* reduz substancialmente a taxa de classificação.

Apesar da alta acurácia, o estudo é inteiramente baseado no PIDD, sem validação externa e sem análise de robustez em outros contextos clínicos ou bases desbalanceadas. O uso intensivo de meta-heurística (NEIWDA) aumenta a complexidade computacional, mas os autores não discutem tempo de execução ou custo para uso em tempo real. Além disso, o *ensemble* combina vários modelos de “caixa-preta”, o que reduz a interpretabilidade das decisões para profissionais de saúde.

3.3.6.7 eDiaPredict: An Ensemble-based Framework for Diabetes Prediction [71]

O estudo propõe um *framework* de predição de diabetes baseado em *ensemble learning* (eDiaPredict), combinando diferentes algoritmos para melhorar a acurácia do diagnóstico e a robustez do modelo. Assim como estudos anteriores, também utiliza o *dataset* PIDD, com atributos clínico como glicose, IMC, insulina e idade. Este estudo

encontra-se dividido em 6 etapas, são elas:

- Pré-processamento: detecção de valores ausentes e imputação pela mediana;
- Análise de correlação para evitar atributos redundantes;
- Seleção de atributos via RFE (*Recursive Feature Elimination*);
- Treino individual de cinco classificadores: XGBoost, *Random Forest*, SVM, *Neural Network* e *Decision Tree*;
- *Ensemble por Voting* usando os modelos com melhor desempenho (XGBoost + RF);
- 10-*fold cross-validation* para avaliar robustez do modelo.

O desempenho é avaliado com várias métricas: acurácia, sensibilidade, especificidade, Área Sob a Curva (AUC), área sob o *convex hull* (AUCH), *Gini Index*, erro mínimo e *Minimum Weighted Coefficient* (MWC).

Entre os modelos individuais, o XGBoost obteve melhor desempenho, com acurácia igual a 92,21% e AUC de 0,873. O RF foi o segundo melhor, com uma acurácia igual a 80,95%. Por outro lado, com o *ensemble* e ajuste de parâmetros, tem-se que o XGBoost alcançou uma acurácia de 95%, sensibilidade de 90,32% e especificidade de 96,8%. Dessa forma, os autores destacam que houve uma melhora de até 15% em relação aos trabalhos anteriores citados na literatura.

Porém, o próprio estudo reconhece que existem alguns pontos limitantes no que tange ao impacto clínico do estudo apresentado, sendo eles:

- Uso de um único *dataset* limitado a mulheres, sem validação externa ou multi-institucional, o que acaba reduzindo a generalização populacional;
- Dados estáticos e pouco variados, sem integração de sensores/tempo real (apesar da motivação com IoT);
- Ausência de análise de interpretabilidade e impacto da decisão para profissionais da saúde;
- *Ensemble* aumenta complexidade computacional, mas os autores não discutem tempo de execução ou possíveis gargalos em produção.

Após abordar de forma individual cada um dos 7 artigos selecionados, a Tabela 16 apresenta de forma resumida as principais abordagens baseadas em técnicas de inteligência artificial, apresentadas nos artigos destacados acima, para predição de diabetes *mellitus* a fim de facilitar a análise e comparação dos trabalhos estudados.

Tabela 16: Comparação de estudos que aplicam técnicas de Inteligência Artificial para apoio à classificação, estratificação de risco ou diagnóstico de diabetes.

Referência	**Objetivo**	**Metodologia / Técnicas**	**Base de Dados**	**Resultados Principais**	**Limitações**
[65]	Diagnóstico de diabetes com abordagem neuro-fuzzy	PIDD; normalização; detecção de outliers (LOF); ANFIS e comparação com ML	PIDD (768 instâncias)	Acc. 92,77%; Sens. 97,65%	Apenas PIDD; sem validação externa; população restrita às mulheres Pima
[66]	Sistema de apoio à decisão baseado em ANFIS	Sistema fuzzy com 240 regras + ANFIS; treino/teste 80/20	PIDD	Acc. 86,48%; Sens. 92,78%	Modelo complexo e pouco interpretável; ausência de validação cruzada
[67]	Predição não invasiva via sinais PPG	PPG + DWT + extração de 130 atributos; seleção ReliefF e mRMR; ANFIS	Sinais PPG (não detalhado)	Acc. 99,1% para DM	Dependência da qualidade do sinal; dados populacionais pouco descritos
[68]	Aumentar acurácia combinando RBF e fuzzy clustering	PIDD; clustering fuzzy baseado em densidade + RBF + ANFIS	PIDD	Acc. 97,14%	Sem validação externa; risco de overfitting; baixa interpretabilidade
[69]	Arquitetura segura para predição em fog computing e blockchain	IoT + blockchain + *rule-based clustering* + FS-ANFIS	Sensores (dados não detalhados)	Acc. $>$81%	Prova de conceito; escalabilidade e implantação clínica não discutidas
[70]	Modelo híbrido otimizado para predição de DM-II	Seleção de atributos com NEIWDA + ensemble (KNN, ANFIS, ANN, SVM)	PIDD	Acc. 98,80%	Alto custo computacional; menor interpretabilidade; sem validação externa
[71]	Melhorar desempenho via ensemble	Pré-processamento; RFE; XGBoost + Random Forest; 10-fold CV	PIDD	Acc. 95%; Sens. 90,32%; Esp. 96,8%	Dados apenas de mulheres; sem integração clínica

4 METODOLOGIA GERAL

Este capítulo tem como objetivo apresentar de forma geral as metodologias utilizadas para cada um dos três grandes focos deste trabalho: o desenvolvimento do aplicativo, o processo de extração de informação e a classificação de doenças, que em um primeiro momento será a classificação laboratorial de DM2.

A Figura 18 apresenta, de forma resumida, a organização metodológica do trabalho a fim de facilitar o entendimento acerca do que foi desenvolvido. Através dessa figura pode-se observar que existem três grandes áreas que compõem o Projeto Dados 48, são eles: o aplicativo, que refere-se ao desenvolvimento propriamente dito da solução, bem como da arquitetura e do banco de dados que rodam por trás da aplicação; o processo de extração de informação, que consiste basicamente no sistema de receber um arquivo PDF de exame de sangue e retornar um arquivo .csv com as variáveis de interesse e seus respectivos valores; e por fim, a classificação de doenças, que consiste na aplicação de uma metodologia ANFIS para a classificação laboratorial de DM2.

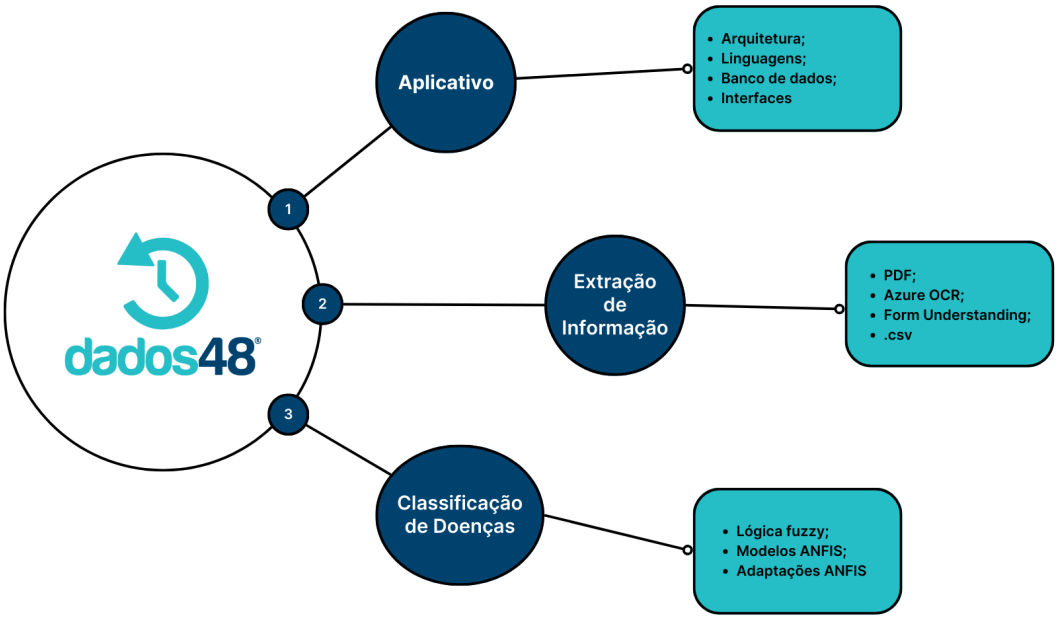


Figura 18: Organização metodológica do trabalho
Fonte: Autora.

Para facilitar a organização, cada um destes grandes grupos é apresentado em um capítulo, assim como seus resultados e respectivas análises.

Primeiramente, são apresentados os itens associados ao desenvolvimento e resultados do aplicativo em si. Para isso, o Capítulo 5 apresenta o desenvolvimento do aplicativo, com foco no *front* e no *backend*, com destaque para as linguagens utilizadas, a arquitetura proposta e o funcionamento do banco de dados. Em seguida, o capítulo apresenta os resultados associados ao desenvolvimento do aplicativo e o seu respectivo desempenho.

Seguindo, o próximo passo consiste em apresentar o grupo que compõe o processo de extração de informação, sendo que para isso, o capítulo 6 é dividido em três frentes: o item 6.1 apresenta a metodologia adotada para o processo de extração de informação; o item 6.2 apresenta um estudo de caso a fim de demonstrar os resultados obtidos a partir da metodologia adotada; e, por fim, o item 6.3 apresenta a análise dos resultados obtidos a partir da utilização de métricas de avaliação.

Na sequência, o Capítulo 7 apresenta o processo de classificação diagnóstica de DM2 e, para isso, também se encontra dividido em algumas frentes, são elas: o item 7.1 apresenta o banco de dados, a fim de mostrar de onde foram tirados os dados e quais os dados foram considerados para cada uma das modelagens; em seguida, o item 7.2 apresenta, em detalhes, a metodologia desenvolvida para o processo de classificação diagnóstica e de risco de DM2; e, por fim, o item 7.3 apresenta os resultados obtidos a partir de uma análise utilizando métricas de avaliação.

5 APLICATIVO DESENVOLVIDO

Este capítulo visa apresentar, de forma mais detalhada, a metodologia utilizada para a construção do aplicativo propriamente dito, destacando a arquitetura e as linguagens utilizadas. Além disso, são apresentados os resultados acerca do desenvolvimento do aplicativo, a partir de uma análise de desempenho da API e da apresentação das interfaces desenvolvidas.

A Figura 19 apresenta a tela inicial do aplicativo desenvolvido, sendo suas funcionalidades melhor apresentadas no decorrer deste capítulo.

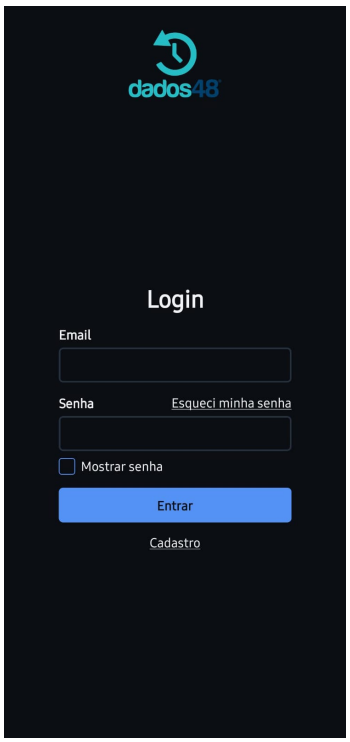


Figura 19: Tela de Login ou Cadastro do Usuário
Fonte: Autora.

5.1 Visão geral da arquitetura da plataforma

Esta seção tem como objetivo apresentar a visão geral da arquitetura que compõe o aplicativo desenvolvido. A Figura 20 apresenta um resumo contendo os principais componentes da plataforma.



Figura 20: Principais elementos presentes no aplicativo
Fonte: Autora.

De forma resumida, os quatro elementos que compõem a figura podem ser descritos como:

- **Aplicativo**: apresenta a visão geral do aplicativo desenvolvido, destacando o relacionamento existente entre os profissionais de saúde e os pacientes, bem como a maneira como os dados enviados são armazenados e apresentados no aplicativo;
- **Banco de Dados**: apresenta a estrutura que está por trás do funcionamento do aplicativo, onde todas as relações são estabelecidas e também onde todos os dados são armazenados;
- **Extração de Informação**: apresenta a proposta de metodologia para a extração automática dos dados oriundos dos exames de sangue enviados via aplicativo no formato PDF;
- **Classificação**: apresenta a metodologia desenvolvida para predizer possíveis anomalias associadas à saúde do paciente a partir dos dados fornecidos pelo próprio paciente no aplicativo. A partir desta classificação, o paciente pode receber alertas de forma automática destacando os possíveis problemas encontrados.

Destes quatro elementos, dois deles: o processo de extração de informação e de classificação dos dados, são abordados com maior ênfase neste trabalho, uma vez que possuem metodologias/ferramentas que foram desenvolvidas no decorrer deste trabalho. Por outro lado, o desenvolvimento do aplicativo e do banco de dados são apresentados de forma mais breve, uma vez que fazem parte deste trabalho mas não tiveram suas metodologias desenvolvidas efetivamente neste estudo, uma vez que foram utilizadas ferramentas/tecnologias já existentes.

5.1.1 Aplicativo

O desenvolvimento do aplicativo propriamente dito e o banco de dados são abordados a seguir, com o intuito de apresentar as técnicas/modelos que foram utilizados para a construção da arquitetura do aplicativo e do banco de dados.

5.1.1.1 Arquitetura

Para o desenvolvimento das interfaces da ferramenta, ou seja, o conjunto de elementos sensoriais (visuais, táteis e sonoros) que possibilitam que os usuários acessem os recursos oferecidos pela aplicação visando sua máxima usabilidade, denominados como o *front-end*, optou-se por utilizar o *React Native* como arquitetura da aplicação. A escolha deste *framework* levou em consideração sua capacidade de projeção de um único aplicativo que se traduz nas plataformas *Web* (*World Wide Web*), *Android* e sistema operacional da *Apple* (*iPhone Operating System* - IOS), proporcionando uma experiência nativa e integrada a cada um destes ecossistemas.

O *React Native* utiliza do *React*, uma biblioteca *JavaScript* que permite a construção de páginas HTML (*Hypertext Markup Language*) ágeis e dinâmicas, com os recursos nativos presentes em cada plataforma. O *framework* é caracterizado pelo lema: “Aprenda uma vez, use em todo lugar”, isso porque ele utiliza de elementos genéricos que possuem um correspondente para cada ecossistema, dessa forma, o *React Native* descreve os componentes visuais e funcionais com as *applications programming interfaces* (APIs) próprias, o que torna o aplicativo mais integrado e estável em cada um deles.

Como suporte à arquitetura do projeto optou-se ainda por utilizar o *Expo*, que oferece um conjunto de ferramentas que auxiliam no desenvolvimento, configuração, teste e compilação de aplicativos *React Native*. O *Expo Router* utiliza dos mesmos conceitos do *React Navigation*, outra ferramenta que agiliza a navegação entre páginas do aplicativo, para estabelecer uma relação semântica entre os diretórios e arquivos do projeto. Em suma, o *Expo* adiciona serviços que otimizam a construção da aplicação em *React Native*, além de reunir bibliotecas essenciais para seu desenvolvimento.

5.1.1.2 Linguagem TypeScript

O *TypeScript* é uma linguagem de programação criada pela *Microsoft* e construída em cima do *JavaScript*, adicionando diversos recursos não presentes na versão nativa da linguagem, esses elementos auxiliam na programação de códigos menos suscetíveis a erros, principalmente quando sua escala aumenta. O *TypeScript* é compilado em *JavaScript* nativo, por conta disso, ele é compatível com a arquitetura proposta no projeto e será utilizado no desenvolvimento do aplicativo.

5.1.1.3 Publicação da ferramenta

A compilação da aplicação é feita através do *Expo*, que resulta em um arquivo compatível com o requerido na biblioteca oficial de aplicativos do *Android*. Então, a publicação do aplicativo para a *Google Play Store* é feita através do *Google Play Console*, um serviço para desenvolvedores *mobile*.

5.1.1.4 Versões do aplicativo

O desenvolvimento do aplicativo pode ser dividido em duas grandes etapas: a primeiras delas, chamada de V0.1 onde foram implementadas somente as funcionalidades principais relacionados aos médicos e pacientes, envio e armazenamento de exames de sangue em PDF, extração automática dos dados dos exames enviados e apresentação destes dados na tela; a V0.2, como está sendo chamada a segunda versão da ferramenta, corrige todos os erros detectados na versão anterior, além de inserir novas funcionalidades, como os perfis de enfermeiro e cuidador, o envio de informações básicas dos pacientes (peso, oxigenação, pressão arterial, altura e glicose capilar) e os gráficos das variáveis extraídas dos exames de sangue. As telas geradas para o aplicativo em si são apresentadas no capítulo 9.

5.1.2 Banco de Dados

A estrutura básica do banco de dados do projeto é apresentada pela Figura 21. A partir dela, alguns pontos acerca de linguagens e arquiteturas consideradas para a construção do banco de dados são abordados e apresentados a seguir.

Uma máquina virtual criada na ferramenta *Azure* hospeda todos os aplicativos. De forma resumida, cada um dos itens que compõem a figura, apresentada anteriormente, são apresentados a seguir:

- *Backend*: responsável por realizar todas as operações associadas aos dois bancos de dados do sistema, sendo um deles um banco de dados relacional implementado pelo MySQL (*My Structured Query Language*) e o outro um banco de dados de chave valor específico para arquivos binários (MinIO). Além disso, também é o responsável por realizar a comunicação com a API de extração de informação;

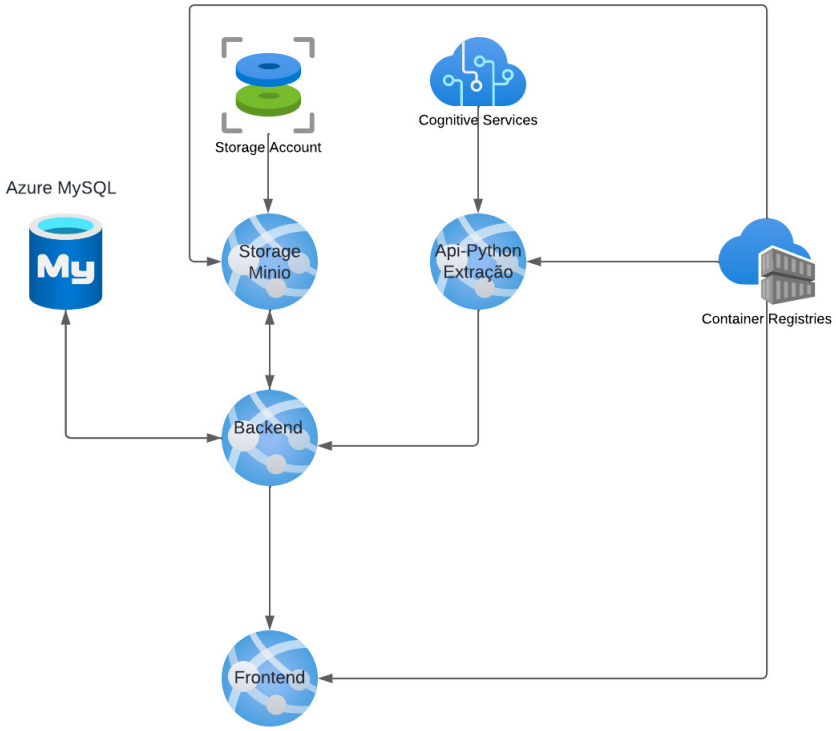


Figura 21: Estrutura da aplicação na nuvem
Fonte: Autora.

- *Frontend*: corresponde a interface da aplicação propriamente dita;
- API *Python* Extração: consiste em um microsserviço de extração que usa o *Document Intelligence* da *Azure* para aplicação de OCR e processamento de documentos;
- *Storage MinIO*: consiste em um *endpoint* que permite acessar o *object storage* S3 da AWS (*Amazon Web Services*) mesmo utilizando o *object storage* da *Microsoft*;
- *Container Registries*: local onde se sobem as três imagens do *docker*, são elas: *Storage MinIO*, a API *python* e o *frontend*.

Em resumo, pode-se dizer que o sistema desenvolvido conta com três aplicações: a primeira consiste no *backend* da aplicação, publicado como uma API (como mostra a Figura 22); a segunda, o *frontend*, que se trata de um aplicativo para *Android*; e a terceira, uma API, que compreende um componente do *backend*, específica para realizar a extração de dados de PDFs. Todos estes itens são empacotados em contêineres o que torna esta aplicação com baixo acoplamento, ou seja, os módulos são interligados mas conseguem ser autônomos nas suas funcionalidades.

5.1.2.1 Arquitetura

Para a implementação da aplicação, optou-se por utilizar uma arquitetura baseada em uma API REST (*Representational State Transfer*), apresentada pela Figura 22. A

escolha desta arquitetura se deu devido aos seus princípios e restrições bem definidos, os quais promovem a escalabilidade, flexibilidade, facilidade de compreensão do sistema desenvolvido e separação bem definida do *frontend* e *backend*.

Essa arquitetura é caracterizada por seu foco na manipulação de recursos (ou entidades) e na utilização dos métodos HTTP (*Hypertext Transfer Protocol*) padrão (*GET, POST, PUT, DELETE*) para realizar operações CRUD (*Create, Read, Update, Delete*) no banco de dados sobre esses recursos. Além disso, a API REST permite a utilização de diferentes formatos de representação de dados, como JSON ou XML (*Extensible Markup Language*), tornando-a altamente flexível e adaptável a diferentes necessidades.

A escolha da arquitetura REST para esta aplicação visa garantir às demandas do sistema, além de facilitar a integração com outros sistemas e serviços, como o sistema de extração de texto da Azure.

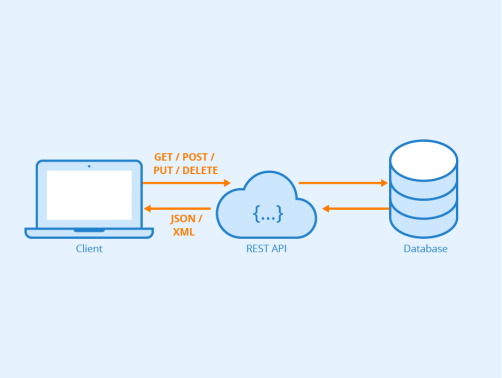


Figura 22: REST API
Fonte: Autora.

5.1.2.2 Linguagens

***Hypertext Preprocessor* (PHP)**

O PHP está sendo utilizado para desenvolver a API que interage com o banco de dados e a API de extração feita em *Python/Flask*. O motivo pelo qual a linguagem foi escolhida é pelo fácil desenvolvimento de aplicações *web*, visto que o PHP é uma das linguagens mais utilizadas para isto, com extensa documentação, ambiente de desenvolvimento simples e adaptável para diversos requerimentos.

A forma como o código está organizado usa controladores para lidar com as requisições feitas para o servidor em uma rota específica, fazendo verificações básicas, como verificar o objeto JSON e modelos de interação com o banco de dados para realizar as operações de lógica e operações de *login*, criação de usuários, *upload* de documentos e afins.

Python

O *Python* está sendo utilizado para desenvolver outra API que interage com um sistema *web* para extração de variáveis de saúde com OCR. O motivo pelo qual esta foi a

tecnologia escolhida para a tarefa foi porquê o sistema de extração já estava desenvolvido na linguagem, tornando assim mais vantajoso continuar a desenvolver a API em *Flask*, que é um *framework* para desenvolvimento de aplicações *web* da linguagem.

5.1.2.3 Armazenamento de dados

MySQL

O MySQL é um sistema de gerenciamento de banco de dados relacional de código aberto amplamente utilizado em aplicativos da *web*, e muitas vezes ele é utilizado ao lado da linguagem PHP, por isso optou-se pela utilização desse sistema para o banco de dados. Além disso, ele oferece uma estrutura confiável e escalável para armazenar e manipular dados de forma eficiente, além de ser uma tecnologia estabelecida, contribuindo para a facilidade de desenvolvimento. O MySQL é utilizado para criar a base da aplicação, com as tabelas com as principais informações, lidando com informações sensitivas de usuário, como: cadastro, variáveis de saúde, relação médico-paciente e afins.

MinIO

O MinIO é um servidor de armazenamento de objetos, altamente escalável e distribuído, projetado para ambientes de nuvem e *data centers* modernos. Ele oferece uma solução eficiente para armazenar e gerenciar grandes volumes de dados não estruturados como no caso da aplicação, os documentos de exames enviados pelo usuário. O MinIO foi utilizado para fazer o armazenamento de objetos e conseguir utilizar a mesma estrutura da API da S3 da *amazon* de armazenamento de objetos mesmo sem precisar utilizar o serviço da mesma, ou seja, ele é independente do sistema de *cloud* utilizado.

5.1.2.4 Funcionalidades Implementadas

Durante o desenvolvimento da aplicação, diversas funcionalidades foram implementadas para atender às necessidades dos usuários e garantir uma experiência eficaz. Algumas das principais funcionalidades incluem:

- Sistema de autenticação e controle de acesso: A autenticação dos usuários é realizada por meio de *tokens*, garantindo uma comunicação segura, a mesma foi projetada para garantir máxima segurança das informações e conformidade com as regulamentações de proteção de dados, como a LGPD. Para a proteção das credenciais de acesso, as senhas são criptografadas no banco de dados utilizando algoritmos hash. O controle de acesso adota a arquitetura RBAC (*Role-Based Access Control*), na qual as rotas e permissões da API são restritas diretamente no *backend*. Isso permite que o sistema possua diferenciação de perfis, como Paciente, Médico, Enfermeiro e Cuidador, assegurando que cada usuário acesse exclusivamente as informações para as quais possui autorização.

- Cadastro e gerenciamento de usuários: Esta funcionalidade foi viabilizada por

meio de uma interface de usuário intuitiva e responsiva, focada na experiência do usuário para facilitar os processos de integração e manutenção de conta. Quando um profissional da saúde, como médico ou enfermeiro, realiza o seu registro, o sistema exige e valida campos específicos e obrigatórios, por exemplo do registro profissional correspondente ao CRM ou COREN.

- *Upload* e armazenamento de documentos: Esta funcionalidade é responsável por receber os históricos médicos do paciente de forma totalmente digital e protegida. Os arquivos enviados, como exames de imagem e PDFs de laudos laboratoriais, são direcionados para um repositório de armazenamento em nuvem seguro do tipo *Object Storage*. Para mitigar qualquer risco de vazamento ou acesso indevido, o sistema utiliza criptografia de ponta a ponta e gera URLs temporárias pré-assinadas, garantindo que os documentos confidenciais fiquem acessíveis apenas durante o tempo necessário para a sua visualização ou processamento interno.

- Integração com API de extração de texto: Este módulo automatiza a transformação de documentos não estruturados em dados clínicos acionáveis através de uma conexão direta com um microsserviço independente desenvolvido em *Python* com o *framework Flask*. Logo após o *upload* do documento, esta API executa uma esteira de OCR para converter a imagem ou PDF em texto bruto. Na sequência, o sistema aplica algoritmos de Processamento de Linguagem Natural (PLN) e expressões regulares para isolar, tratar e estruturar as variáveis de saúde relevantes, como taxas de glicose, colesterol e creatinina, salvando-as diretamente no histórico do paciente para alimentar o banco de dados.

- Rede Social: Um ambiente digital controlado para conectar o paciente ao seu círculo de atendimento técnico. A associação entre os usuários ocorre por meio de um mecanismo de convites, disparados via e-mail ou por busca baseada em documentos oficiais. O fluxo digital exige que o paciente aceite o convite de um profissional para que este passe a ter visibilidade sobre seu prontuário e histórico de exames extraídos pelo OCR. Visando a autonomia do indivíduo sobre suas próprias informações de saúde, esse vínculo e a respectiva permissão de visualização dos dados podem ser revogados pelo paciente a qualquer momento.

- Módulo Neural *Fuzzy*: Este módulo visa integrar a rede social ao módulo de alertas baseado em um Modelo Neural *Fuzzy*. O mesmo combina técnicas de redes neurais com lógica *fuzzy* para analisar de forma contínua as séries temporais dos exames do paciente e caso identifique uma tendência de piora clínica ou padrão de risco iminente, é disparado alertas e notificações em tempo real na interface dos médicos e cuidadores vinculados para permitir uma análise rigorosa da equipe. Este módulo,

Tabela 17: Resultados Quantitativos da Avaliação de Desempenho da API

Endpoints	**Total de Requisições**	**Tempo de Resposta (ms)**	**Requisições/s**	**Taxa de Erro**
/sessao/criarsessao”	472	25,74	1.30	0
/paciente/criarpaciente”	399	21,52	1.10	0
/enfermeiro/criarenfermeiro”	392	4,00	1.08	0
/medico/criarmedico”	394	2,40	1.09	0
“/cuidador/criarcuidador”	396	2,42	1.09	0

apesar de não estar 100% operacional, já está previsto e pode ser melhor explorado em trabalhos futuros.

5.2 Resultados associados ao desenvolvimento do aplicativo

Nesta seção, primeiramente é apresentada uma análise de desempenho da API, a fim de avaliar o desempenho da arquitetura proposta. Em seguida, são apresentadas algumas interfaces do aplicativo, a fim de exemplificar as funcionalidades que compõem o mesmo.

5.2.1 Análise do Desempenho da API

Para avaliar o desempenho da API, foram realizados testes de carga utilizando a ferramenta Postman. O cenário de teste simulou usuários virtuais sob uma carga fixa (sem período de *ramp-up*) para avaliar a capacidade de resposta da aplicação à demanda. Os endpoints avaliados nesta análise foram /sessao/criarsessao”, /paciente/criarpaciente”, /enfermeiro/criarenfermeiro”, /medico/criarmedico” e “/cuidador/criarcuidador”. Esses endpoints representam um fluxo típico de utilização da aplicação, iniciando pela autenticação do usuário e geração do token de acesso, seguido pela criação dos diferentes perfis de usuários disponíveis no sistema (paciente, cuidador, enfermeiro e médico).

Os parâmetros do teste foram definidos da seguinte forma:

- Número de usuários virtuais: 100;
- Duração do teste: 5 minutos;
- Número total de requisições executadas: 2053.

Para uma melhor visualização, os resultados quantitativos obtidos são apresentados na Tabela 17.

Com base na análise de desempenho, pode-se concluir que a API apresentou um comportamento satisfatório sob o cenário de carga proposto. Durante o período de teste de 5 minutos, com 100 usuários virtuais concorrentes, foi executado um total de 2053 requisições distribuídas entre os cinco endpoints avaliados. Os tempos médios de resposta

permaneceram baixos para todas as rotas, variando entre 2,40 ms e 25,74 ms. Destaca-se que o endpoint de autenticação, "/sessao/criarsessao", apresentou a maior latência média, o que era esperado devido ao processamento adicional necessário para validação das credenciais e geração do token de acesso.

Além disso, a taxa de processamento permaneceu estável ao longo dos testes, demonstrando a capacidade da aplicação de lidar com múltiplas requisições simultâneas sem degradação significativa de desempenho. Nenhum erro foi registrado durante a execução dos testes, indicando estabilidade operacional mesmo sob condições de carga concorrente. Dessa forma, os resultados apresentados na Tabela 17 demonstram a robustez da API para o cenário avaliado.

5.2.2 Interfaces

A ferramenta conta com uma série de telas que representam as mais diversas funcionalidades do aplicativo, porém, nem todas elas são apresentadas neste trabalho. Optou-se por apresentar somente algumas mais gerais a fim de ilustração da ferramenta desenvolvida.

A aplicação resultante fornece uma experiência simples e direta aos seus usuários. Através de suas interfaces simples e descomplicadas, é possível cadastrar uma nova conta ou fazer login em uma conta já existente, como mostra a Figura 23.

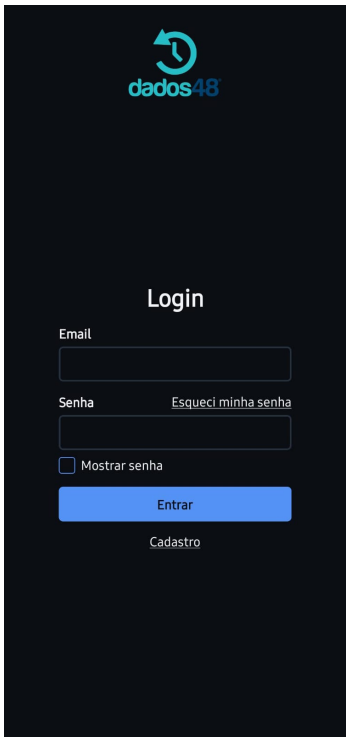


Figura 23: Tela de Login ou Cadastro do Usuário
Fonte: Autora.

Após a criação da conta no aplicativo e da autenticação do usuário, o mesmo pode de forma simples criar perfis direcionados às suas necessidades, entre eles: "Paciente",

“Médico”, “Enfermeiro” e “Cuidador”, como pode ser observado pela Figura 24.

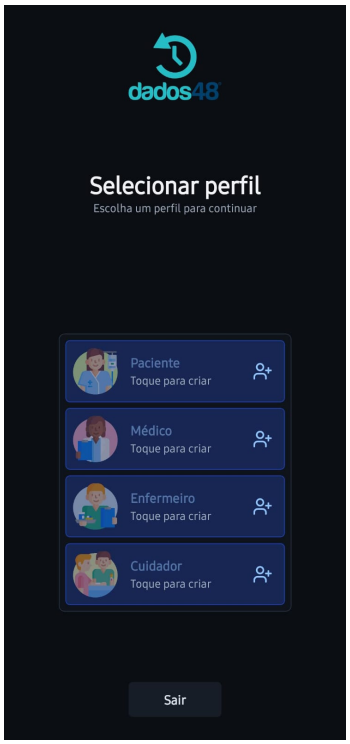


Figura 24: Tela para selecionar o perfil que deseja criar
Fonte: Autora.

A Figura 25 apresenta as telas referentes às criações das quatro opções de perfil’s. Para o caso da criação do perfil de médico, apresentado pela Figura 25a, o usuário precisa informar a sua Unidade Federativa (UF) o seu número perante o Conselho Regional de Medicina (CRM); por outro lado, para a criação do perfil de enfermeiro, apresentado pela Figura 25b, o usuário precisa informar a UF, o número do seu Conselho Regional de Enfermagem (Coren) e sua categoria; já para os perfil’s de paciente (Figura 25c) e cuidador (Figura 25d), o usuário não necessita acrescentar nenhuma outra informação, somente o número do seu Cadastro de Pessoa Física (CPF) já informada na criação da própria conta.

Ressalta-se que um mesmo usuário pode possuir mais de um perfil, portanto ao fazer o *login* na sua conta deve selecionar o perfil que deseja ingressar, como mostra a Figura 26. Cada perfil determina o nível de acesso do usuário e as funcionalidades disponíveis, permitindo a conexão entre pacientes e profissionais de saúde, a partir do claro consentimento e autorização de ambos.

Ao acessar o seu perfil de paciente, o usuário tem acesso às suas informações pessoais (Figura 27a), a lista de profissionais que cuidam da sua saúde (Figura 27b), consegue adicionar informações complementares à sua saúde (Figura 27c) e consegue adicionar/visualizar os seus exames (Figura 27d).

Por outro lado, ao acessar o perfil como médico, enfermeiro ou cuidador, os profissi-

onais podem então visualizar as informações médicas alimentadas pelo paciente através de exames enviados dentro da plataforma e de informações de saúde e hábitos de vida informados no aplicativo.

Considerando, portanto, que o usuário acessou o seu perfil de médico. Nele, o médico pode visualizar os seus pacientes (Figura 28a) e/ou visualizar se há solicitação para aceitar um novo paciente (Figura 28b).

Após vinculado à um paciente, o médico passa a ter acesso as informações do mesmo, como por exemplo: relação de exames do paciente (Figura 29a), variáveis extraídas para cada um dos exames listados (Figura 29b) e gráfico em formato de linha do tempo para cada uma das variáveis que foram extraídas dos exames enviados pelo paciente (Figura 29c).

A Figura 29b apresenta o recurso de extração proposto na plataforma. Através do envio de um exame médico no formato PDF pelo paciente ou cuidador autorizado, a aplicação é capaz de extrair as variáveis médicas e apresentar os dados listados com seus respectivos valores, facilitando o acesso e conferências destas informações.

Conforme o usuário envia mais exames para a ferramenta, os valores de cada variável extraída vão se acumulando em um histórico, demonstrado na Figura 29c, permitindo visualizar um gráfico em função do tempo que evidencia a evolução dos valores para cada variável médica, permitindo observar a variação destas informações em função da data de realização de cada exame. O profissional de saúde, uma vez autorizado a acessar estas informações, é capaz de ter um panorama geral do quadro do paciente, facilitando o acompanhamento de sua saúde, fazendo as devidas intervenções conforme a evolução apresentada, uma vez que as informações estão concentradas em um só lugar.

Cada indivíduo possui estilos de vida que impactam sua saúde de maneiras diferentes, por conta disso, ao analisar um cenário, é importante conhecer o contexto em que aquela pessoa vive, uma vez que algumas informações são fatores de risco ou ainda causadores de muitas condições. Por conta disso, visando complementar a análise e monitoramento pelos profissionais de saúde, a aplicação também conta com o preenchimento de informações complementares, apresentado pela Figura 27c, que permite aos pacientes informar peso, altura e informações recorrentes como pressão sanguínea e glicose capilar (em especial àqueles que fazem acompanhamento diário destes dados), assim como os hábitos de saúde, que auxilia a entender o perfil e estilo de vida deste paciente, o que pode ajudar a identificar possíveis fatores que podem contribuir em algum quadro que possa apresentar.



(a) Tela de Criação de Perfil de Médico



(b) Tela de Criação de Perfil de Enfermeiro



(c) Tela de Criação de Perfil de Paciente



(d) Tela de Criação de Perfil de Cuidador)

Figura 25: Telas de Criação de Perfil
Fonte: Autora.

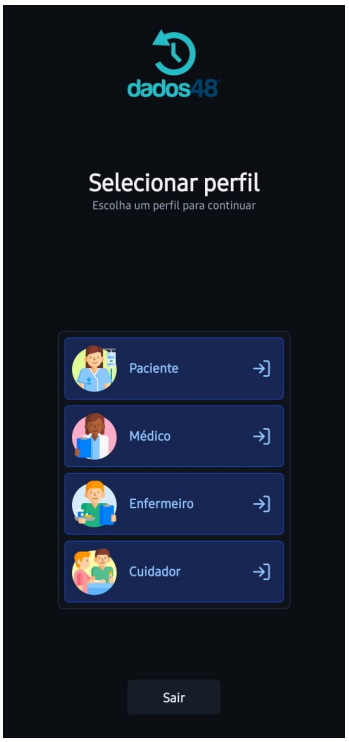
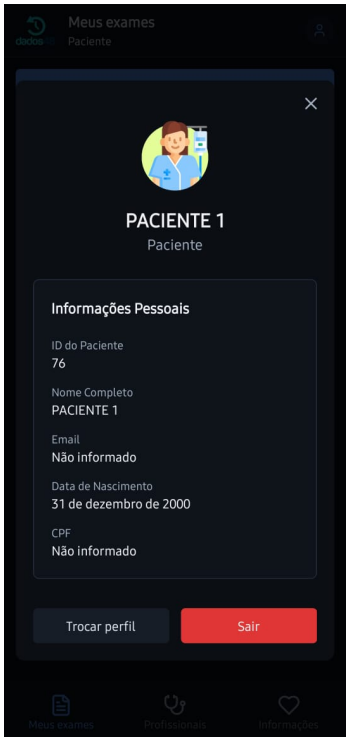


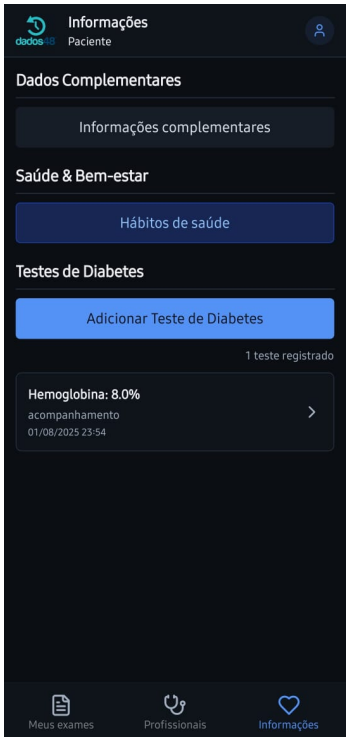
Figura 26: Tela para selecionar o perfil que deseja ingressar no APP
Fonte: Autora.



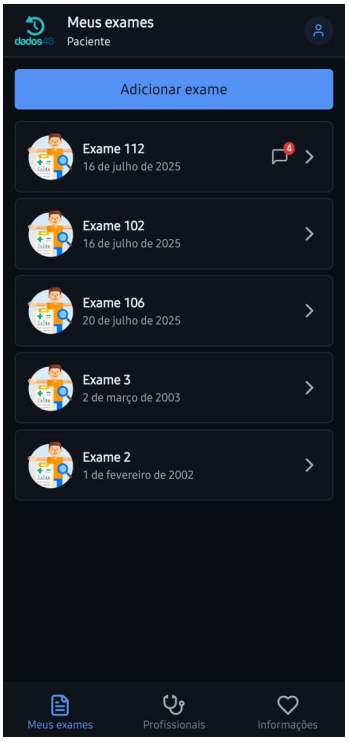
(a) Tela de Acesso as Informações Pessoais



(b) Tela de Acesso à Lista de Profissionais de Saúde



(c) Tela de Acesso e Cadastro das Informações Complementares à Saúde

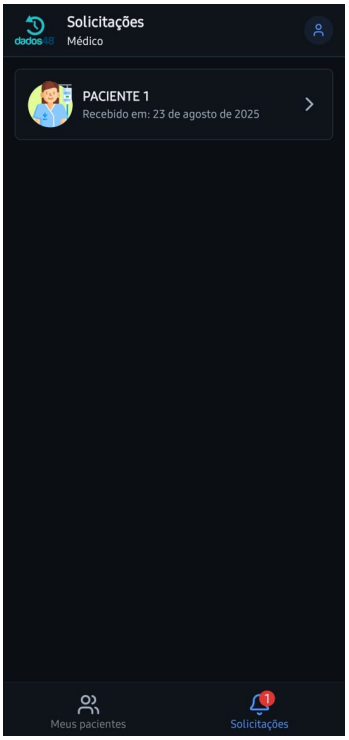


(d) Tela de Acesso e Inserção de Exames de Sangue

Figura 27: Telas disponíveis para o perfil paciente
Fonte: Autora.



(a) Tela de Acesso à sua Lista de Pacientes

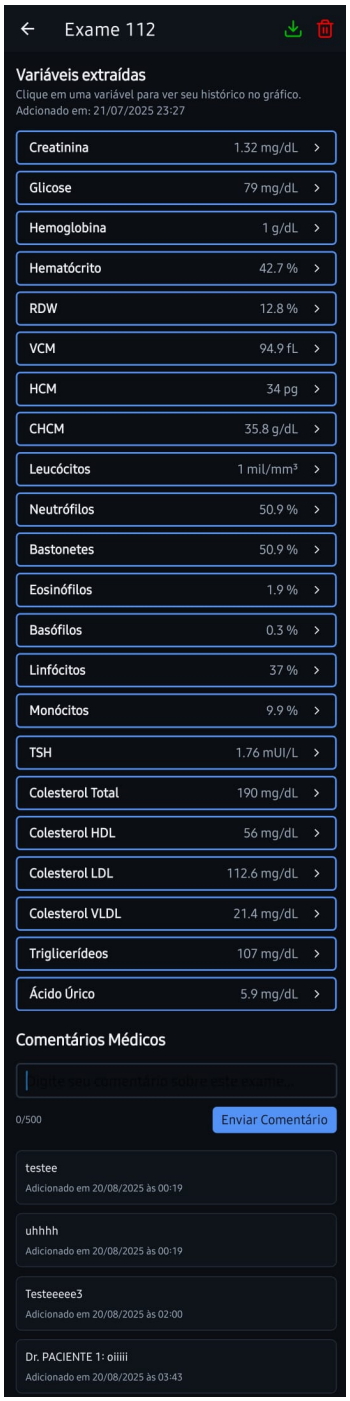


(b) Tela de Acesso às Solicitações para Aceitar novos Pacientes

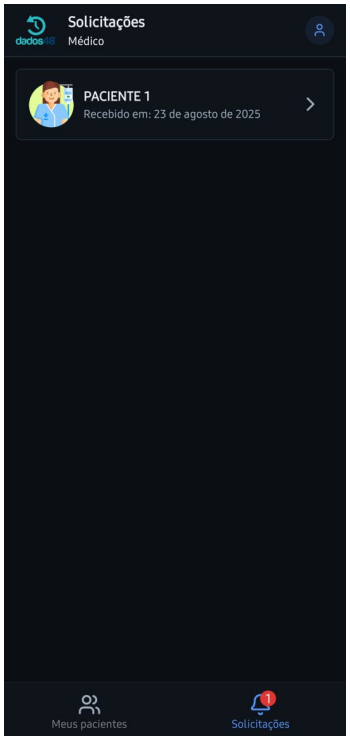
Figura 28: Telas disponíveis para o médico visualizar ou aceitar pacientes
Fonte: Autora.



(a) Tela com a relação de Exames do Paciente



(b) Tela de Acesso às Variáveis Extraídas de um determinado Exame do Paciente



(c) Tela de Acesso ao Gráfico gerado de determinada variável extraída dos exames

Figura 29: Telas disponíveis para o médico visualizar as informações referentes à um determinado paciente

Fonte: Autora.

6 EXTRAÇÃO DE INFORMAÇÃO

6.1 Método para extração de informação de exames de sangue

A metodologia desta etapa pode ser resumida em 4 etapas, como pode ser observado pela Figura 30. Cada uma das etapas será melhor descrita a seguir.

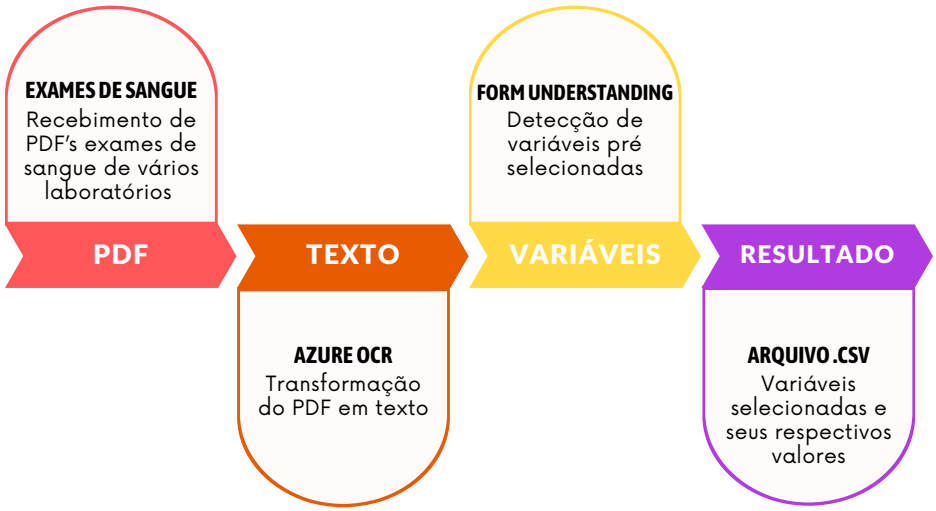


Figura 30: Fluxograma referente a metodologia para extração de informação dos exames de sangue

Fonte: Autora.

6.1.1 PDF

Os laboratórios, em sua grande maioria, disponibilizam para seus pacientes os arquivos PDF contendo os resultados de seus exames de sangue através da internet. Tais arquivos PDF apresentam *layouts* diferentes dependendo do laboratório. Neste estudo foram considerados 22 laboratórios diferentes a fim de verificar se a metodologia que está

sendo utilizada é capaz de funcionar para diferentes *layouts*. Destaca-se que por vezes o mesmo laboratório apresenta *layouts* diferentes ao longo do tempo.

Considerando que este estudo envolve dados reais de pessoas fez-se necessário que o mesmo passasse pela aprovação do comitê de ética, para que os dados fornecidos pelas pessoas pudessem realmente ser utilizados nesta pesquisa. Sendo assim, um projeto de pesquisa intitulado **Repositório dinâmico de informações pregressas e atuais de saúde do usuário** foi submetido ao comitê de ética, obtendo aprovação para trabalhar com os dados fornecidos e autorizados pelas pessoas. Para consultar tal projeto e sua respectiva aprovação utiliza-se o número de identificação **CAAE** (Certificado de Apresentação de Apreciação Ética): **60939822.5.0000.5324**.

6.1.2 Texto

O processo de extração de dados de exame de sangue no formato PDF acontece a partir de um algoritmo implementado na linguagem de programação *python*. A implementação desse algoritmo foi estruturada em uma classe, onde cada método da classe representa uma parte do procedimento.

O primeiro passo consiste em obter o conteúdo dos diferentes arquivos PDF em formato de texto. Diversos fatos tornam a extração de texto de um arquivo PDF uma tarefa desafiadora. As bibliotecas disponíveis muitas vezes retornam resultados imprecisos e, por esse motivo, os métodos que realizam esse passo utilizam ferramentas de IA (Inteligência Artificial).

Considerando os trabalhos estudados e apresentados anteriormente, bem como um estudo referente as novas possibilidades disponibilizadas pela ferramenta de IA da *Microsoft*, optou-se pela utilização da *Azure AI Document Intelligence* para a realização do processo de extração de texto dos arquivos PDF.

A *Azure AI Document Intelligence* consiste em um serviço de IA da *Azure* que permite criar soluções inteligentes de processamento de documentos, ou seja, este serviço permite gerenciar com eficácia a velocidade com que os dados são coletados e processados. O serviço conta basicamente com três opções de modelos para serem explorados:

- **Modelos de análise de documentos**: Permitem a extração de texto de formulários e documentos e retornam conteúdo estruturado pronto para a ação, uso ou o progresso das organizações;
- **Modelos predefinidos**: Permitem adicionar o processamento de documentos inteligentes aos seus aplicativos e fluxos sem precisar treinar e criar seus próprios modelos;
- **Modelos personalizados**: São treinados usando um conjunto de dados rotulados para extrair dados distintos de formulários e documentos, específicos para cada um dos usos de caso.

Considerando a necessidade deste estudo, optou-se pela utilização dos modelos de análise de documentos, uma vez que a ideia consiste em entregar um documento e obter na saída o texto pronto para ser utilizado.

A *Azure AI Document Intelligence* segue a mesma linha da técnica OCR, apresentada anteriormente, no que tange ao processo de transformar textos-imagens, uma vez que a ferramenta encontra-se disponível através de APIs *web*, que recebem uma requisição com o arquivo PDF e retornam uma resposta com o conteúdo de texto obtido.

Porém, o modelo de análise de documentos da *Azure* conta com três opções diferentes de APIs para realizar o processo de extração, são elas:

- ***Read***: Extrai texto impresso e manuscrito junto com códigos de barras, fórmulas e estilos de fontes de imagens e documentos;
- ***Layout***: Extrai tabelas, caixas de seleção e texto de formulários e documentos;
- ***General documents***: Extrai pares de valores-chave e estruturas como tabelas e marcas de seleção de qualquer formulário ou documento.

Considerando que o objetivo nesta etapa consiste em receber o arquivo PDF e extrair o texto contido nele, optou-se pela utilização da API *Layout*. Tal API mostrou-se capaz de receber os diferentes *layouts* de exames de sangue e extrair de todos eles os seus respectivos textos.

Porém, quando o assunto se trata de sistemas automatizados para extração de informação de documentos, apenas a etapa de OCR, acaba por não ser suficiente em alguns casos. Uma vez que o OCR extrai somente textos esparsos, isto é, sem informação semântica atrelada. Por isso, para auxiliar nesses processos, utiliza-se a técnica conhecida como *form understanding* e apresentada anteriormente no capítulo de fundamentação teórica.

6.1.3 Variáveis

Neste caso, uma vez que o conteúdo em texto está disponível, após a utilização da API *Layout* do *Azure*, o próximo passo consiste em aplicar a técnica de *form understanding* a fim de encontrar as variáveis de interesse.

Porém, antes de iniciar o processo de aplicação da técnica *form understanding* e obtenção dos pares-chave, faz-se necessário que se realize um passo intermediário que consiste em retirar a acentuação e converter letras maiúsculas em minúsculas para todas as palavras do texto. Essa etapa é importante para padronizar as palavras e facilitar a busca da variável.

Por se tratar de extração de dados de exames de sangue o escopo de variáveis de interesse pode ser definido de acordo com as necessidades do sistema que esta sendo desenvolvido. Como o objetivo deste trabalho consiste em validar as técnicas de extração

para exames de sangue, extrair o máximo possível de variáveis de cada um dos diferentes exames testados auxilia nesse processo de validação. Sendo assim, optou-se por considerar como variável de interesse todas aquelas variáveis que são encontradas em exames de sangue de rotina, totalizando 37 variáveis analisadas. As variáveis a seguir representam algumas variáveis ou grupos de variáveis que compõem as 37 variáveis analisadas:

- Hemograma Completo;
- Creatinina;
- Ureia;
- Colesterol;
- Ácido Úrico;
- Potássio;
- Vitaminas;
- Glicose;
- Insulina;
- Hemoglobina Glicada (HbA1c);
- TGO;
- TGP;
- Albuminuria

É importante destacar que se utilizou da estrutura *keyValuePairs* (pares chave-valor) para associar cada variável de interesse ao seu respctivo valor, ou seja, as variáveis de interesse representam a chave e o valor representa o próprio número extraído do exame de sangue.

Definidas as variáveis de interesse e realizado o processo de retirada de acentos e conversão de letras maiúsculas em minúsculas, o próximo passo consiste na aplicação da técnica *form understanding* propriamente dita, ou seja, todo o texto será percorrido a fim de verificar a existência das variáveis de interesse e seus respectivos valores. A Figura 31 resume através de um fluxograma este processo de verificação do texto extraído. Destaca-se que, para cada uma das variáveis de interesse, o fluxograma é reinicializado, ou seja, cada uma das variáveis de interesse percorre o fluxograma de forma individual. Dessa forma, o processo apresentado pelo fluxograma se repete tantas vezes quantas forem as variáveis de interesse a serem analisadas.

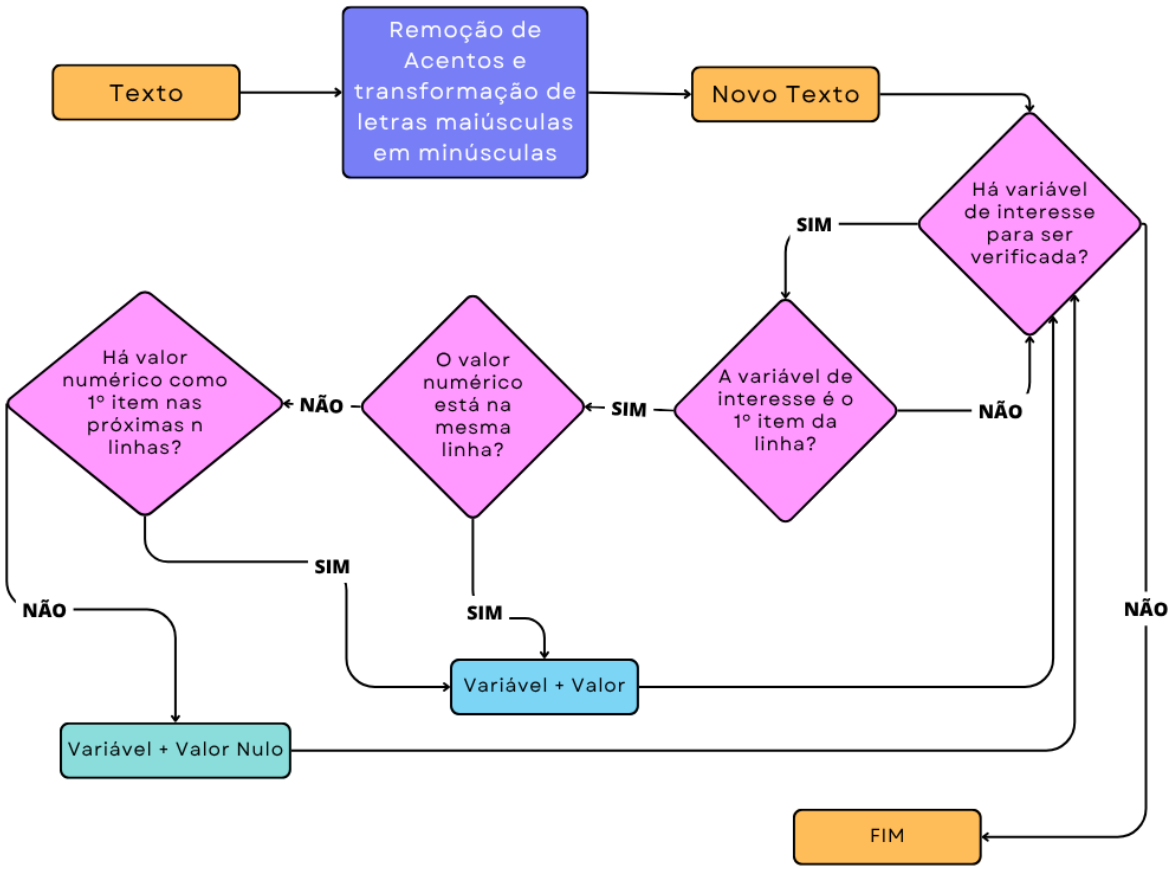


Figura 31: Fluxograma do passo a passo para detecção das variáveis de interesse
Fonte: Autora.

A verificação inicia percorrendo cada uma das linhas a fim de observar se em alguma das linhas a variável de interesse encontra-se localizada como primeiro item da linha e se também há um valor numérico presente naquela linha. Caso essa verificação seja verdadeira, o valor numérico da linha é retornado como o valor da variável e o algoritmo se encerra.

Por outro lado, se não for retornado nenhum valor, segue-se para o próximo passo que consiste em verificar para cada linha se a variável de interesse está como primeiro item. Se para alguma linha essa verificação for verdadeira, o algoritmo irá verificar nas próximas 10 linhas se há algum valor numérico presente como primeiro item. Sendo essa segunda verificação verdadeira, este valor numérico será retornado como o valor obtido.

Se o quarto passo for concluído sem retornar nenhum valor, o algoritmo retornará um valor nulo, indicando que o valor da variável não foi encontrado.

6.1.4 Resultado

Por fim, após a detecção de todas as variáveis de interesse e seus respectivos valores para cada um dos PDFs analisados, todas as informações são compiladas e enviadas para um arquivo .csv a fim de apresentar de forma mais clara as informações extraídas de cada

um dos PDFs. Neste arquivo cada uma das variáveis de interesse representam as colunas da tabela, enquanto que cada um dos PDFs representam as linhas.

6.2 Estudo de caso

A extração de informação de dados de exame de sangue consiste em uma das principais ferramentas presentes no aplicativo.

A extração automática dos dados oriundos dos exames de sangue anexados pelos usuários na ferramenta permite que tanto o paciente quanto os profissionais de saúde à ele vinculados tenham acesso as informações de cada um dos exames na tela no aplicativo de forma mais clara, sem precisar abrir cada um dos documentos.

Além disso, a partir da extração automática dos dados dos exames enviados pelo usuário, a ferramenta possui a funcionalidade de apresentar, através de um gráfico, a variação de cada uma destas variáveis ao longo do tempo.

Assim como descrito na metodologia, os resultados deste estudo podem ser divididos em três etapas, como mostra a Figura 32: a primeira delas apresenta os resultados obtidos a partir da utilização da ferramenta *Azure* para extração do texto dos exames de sangue; a segunda, apresenta os resultados após a aplicação da técnica *form understanding* para a localização e obtenção dos valores das variáveis de interesse; a terceira e última etapa representa a geração de um arquivo .*Comma Separated Values* (.csv) com todos os exames analisados e seus respectivos valores para cada uma das variáveis de interesse. A seguir, cada uma das etapas será melhor descrita e terá seus resultados apresentados.

6.2.1 Utilização da ferramenta *Azure*

O modelo de análise de documentos escolhido, como dito anteriormente, foi o modelo de *Layout*, que consiste em uma API avançada de análise de documentos baseada em aprendizado de máquina, disponível na nuvem. Esta API permite analisar documentos em diferentes formatos e retornar representações de dados estruturados desses documentos. Para extrair as informações de texto, tabelas, marcas de seleção e estrutura do documento, a API combina recursos de OCR com modelos de aprendizado profundo [72].

A Figura 33 apresenta um exemplo, fornecido pela própria *Microsoft*, de documento processado utilizando o modelo *layout* para análise de documentos.

A API pode ser usada em diferentes linguagens de programação, neste estudo foi escolhido usar a linguagem *Python*. No entanto, para que tal API seja utilizada, alguns pré-requisitos devem ser considerados, de acordo com a documentação disponível em [72], eles são:

- Assinatura do *Azure*;
- *Python* 3.7 ou posterior;



Figura 32: Etapas do processo
Fonte: Autora.

- Versão mais recente do *Visual Studio Code* ou de um *Integrated Development Environment* (IDE) da preferência do programador;
- Recurso de Inteligência de Documento ou Serviços de IA do *Azure*;
- Chave e Ponto de extremidade do recurso implementado.

Caso se deseje acessar vários serviços da IA *Azure* faz-se necessário que se crie um recurso de serviços, mas como neste caso a ideia consiste somente em utilizar o serviço de inteligência de documentos, foi necessário criar somente um recurso do *Document Intelligence*. Após a implantação do recurso pode-se obter então as informações da chave e do ponto de extremidade deste recurso, como mostra a Figura 34. Tais informações permitem que se possa conectar o sistema que se está desenvolvendo com a API do *Document Intelligence*, ou seja, para interagir com o serviço do *Document Intelligence* é preciso criar uma instância da classe *DocumentAnalysisClient* utilizando as informações da chave e do ponto de extremidade obtidas anteriormente [73].

Após a definição da chave e do ponto de extremidade, são utilizadas basicamente duas funções principais disponibilizadas pela ferramenta: a função *DocumentAnalysisClient*

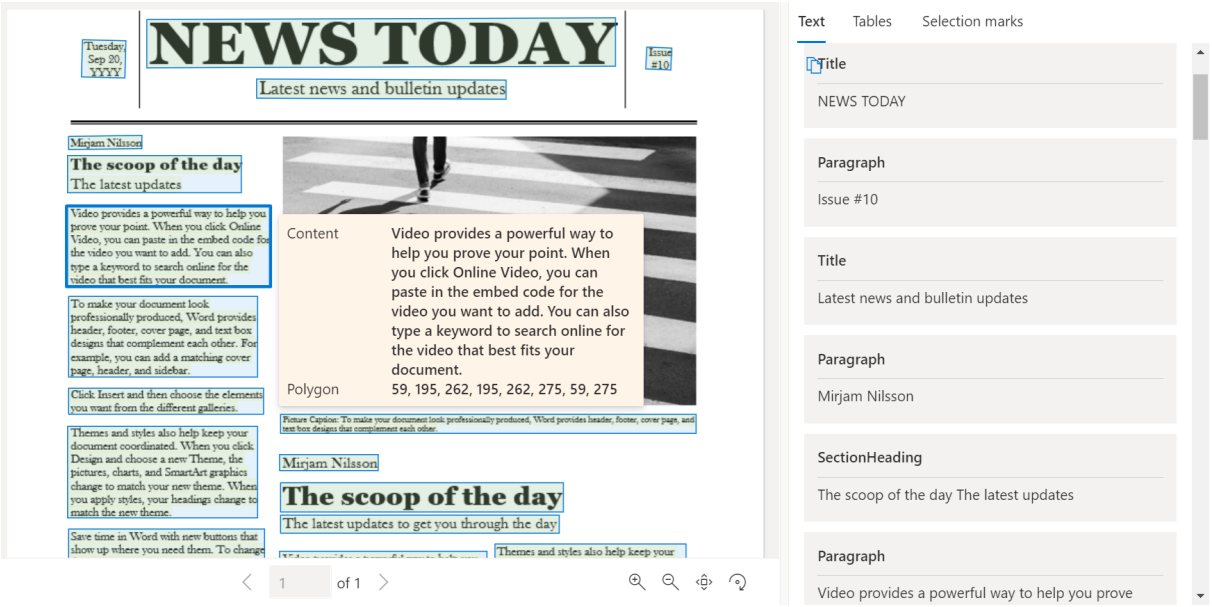


Figura 33: Exemplo de documento processado utilizando o modelo *layout*
Fonte: [72]

que fornece as operações necessárias para analisar os documentos fornecidos a partir da utilização de modelos pré-construídos e personalizados; e a função *analyze_layout* que é utilizada para analisar o documento. O recurso *prebuilt-layout* é utilizado para extrair texto e localizações de texto, tabelas, marcas de seleção e informações de estrutura de documentos e imagens.

Um exemplo de saída do sistema após a utilização destes recursos é apresentado pela Figura 35.

6.2.2 Utilização da técnica *Form Understanding*

Após a extração do texto dos PDFs de exame de sangue, o próximo passo consiste em aplicar a técnica conhecida como *Form Understanding* a fim de encontrar as variáveis de interesse. Para isso, foi desenvolvido um algoritmo que recebe como entrada o texto extraído a partir da utilização da API *Azure* e percorre tal texto seguindo a metodologia apresentada anteriormente pela Figura 10, onde n é igual a 10 e representa o critério de parada.

A Figura 36 apresenta um exemplo da saída do sistema após a aplicação da técnica *Form Understanding* para a extração de algumas das variáveis de interesse apresentadas anteriormente. Além disso, a saída do sistema apresenta a resposta para algumas variáveis de interesse considerando diferentes arquivos, ou seja, diferentes laboratórios e, consequentemente, diferentes *layouts* de documentos.

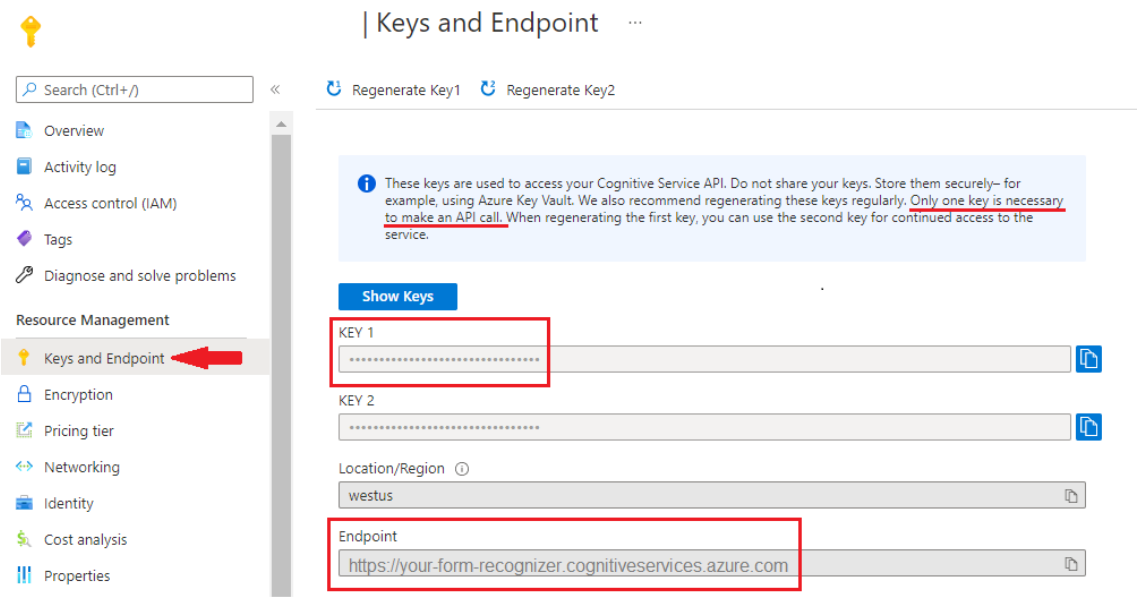


Figura 34: Exemplo de localização das *Keys and Endpoint*
Fonte: [74]

6.2.3 Geração de arquivo .csv

Tão logo as informações das variáveis de interesse são obtidas um arquivo .csv é gerado com as informações extraídas, conforme mostra a Figura 37. Neste arquivo, como descrito anteriormente, as variáveis de interesse representam as colunas da tabela, enquanto que as linhas representam os diferentes arquivos enviados na entrada.

6.2.4 Exemplo de aplicação para 3 laboratórios diferentes

Com o objetivo de mostrar que o sistema desenvolvido realmente retorna no arquivo .csv os valores que constam no PDF original utilizado como entrada para o sistema, são apresentados os resultados obtidos para 3 laboratórios diferentes (Gram, Analisa e Sérgio Franco), como mostram as Figuras 38, 39 e 40. Vale ressaltar que por uma questão de sigilo quaisquer informações que pudessem remeter aos pacientes e/ou médicos dos exames foram encobertas.

Tais documentos foram colocados como entrada no sistema e passaram por todas as etapas descritas anteriormente, ou seja, os três tiveram seus textos extraídos utilizando a ferramenta *Azure* (Figuras 41, 42 e 43).

Em seguida, os textos extraídos foram percorridos a fim de encontrar as variáveis de interesse (ver Figura 44) que são apresentadas no arquivo .csv gerado, como mostra a Figura 45.

Por fim, as Figuras 46, 47 e 48 apresentam os resultados obtidos para cada um dos 3 laboratórios considerados.

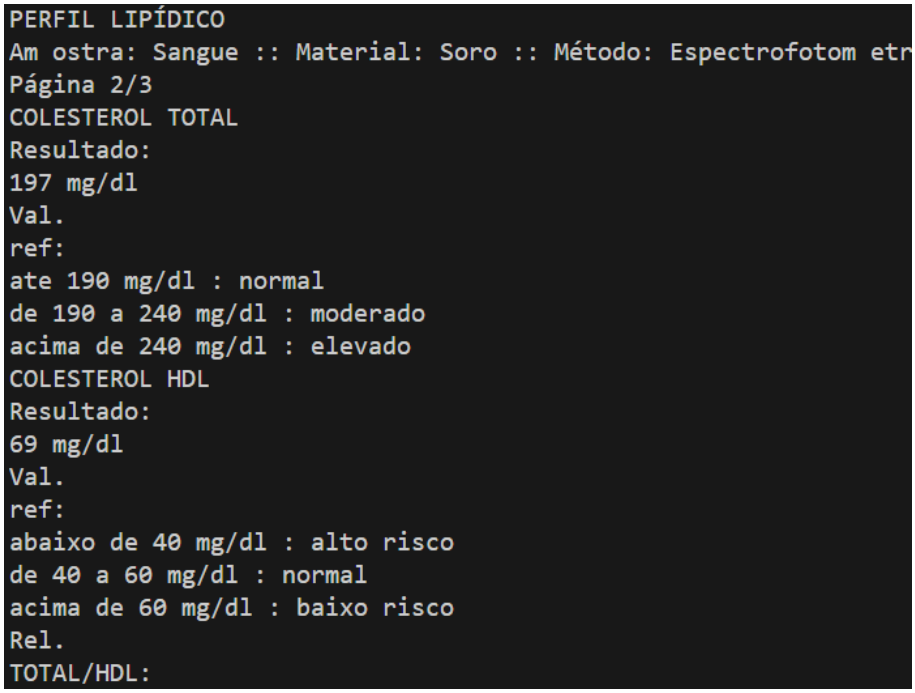


Figura 35: Exemplo parcial de extração de dados de exames de sangue a partir da utilização da ferramenta *Azure* da *Microsoft*

Fonte: Autora.



Figura 36: Exemplo de extração de algumas variáveis de interesse de vários exames de sangue

Fonte: Autora.

6.3 Análise dos Resultados

A presente seção tem como objetivo apresentar as análises realizadas a fim de quantificar o desempenho apresentado pela metodologia proposta. Para isso, são realizadas análises de desempenho quanto aos exames de sangue e as variáveis de interesse, bem como a análise da complexidade.

Os resultados apresentados anteriormente pelo estudo de caso representam exemplos de todo o processo de extração que foi realizado. Porém, os resultados apresentados neste capítulo tratam-se de um recorte, pois a quantidade de laboratórios e exames de sangue é superior ao que foi apresentado pelo estudo de caso. No total, foram analisados 210 exames de sangue diferentes, sendo 133 deles arquivos PDFs originais dos laboratórios e 77 arquivos escaneados.

	creatinina	glicose	bnp	ureia	tgo	tgp	colesterol total	
Analisa1-1-2.pdf							197	
Analisa1-3.pdf	0,8	88						
Analisa1.pdf							197	
Analisa5-1-2.pdf							209	
Analisa5-3-4.pdf		103			29	23		
Analisa5-5.pdf		280,5						
PerezLimardo2-1-2.pdf	0,8			35,5	15	19,3		
SergioFranco-1-2.pdf		107						
SergioFranco-3-4.pdf	0,67	86		37				
SergioFranco-5.pdf							175	
UnimedCaxias-1-2.pdf		91			21	16		
UnimedCaxias-3-4.pdf						24		
Weinmann-1-2.pdf			7	21				
Weinmann-3-4.pdf	0,87							
Weinmann-5.pdf	0,87							

Figura 37: Exemplo de arquivo gerado a partir do processo de extração das variáveis de interesse de vários exames de sangue

Fonte: Autora.

Anteriormente, foram apresentados exemplos de 3 laboratórios distintos, porém, mais laboratórios foram testados, são eles:

- Wilde;
- Lunav;
- Torres;
- Gram;
- Hemolab;
- Birck;
- Santa Casa de Rio Grande;
- Hospital Universitário (HU) - FURG;
- DB Diagnósticos;
- Santa Casa de Pelotas;
- Santa Luzia;
- Antonello;
- Hermes Pardini;
- Mont'Serrat;
- DR Thofehrn;


laboratório Gram análises clínicas
desde 1990
Selo de Qualidade
Sistema Nacional de Acreditação DICQ
patrocinado pela Sociedade Brasileira de Análises Clínicas

Paciente [redacted]
Médico [redacted]
Convênio Unimed Litoral Sul
Procedência Presidente Vargas
Destino Resultado na Internet
Alvará de Localização 414896
00013211
Idade.....26 Anos
Data Req.....16/11/2021
Data Emissão 13/09/2022

Hemograma

Material : Sangue | Método : Citometria de Fluxo / SYSMEX XS-1000i | Data de Coleta: 16/11/2021 | Hora de Coleta: 09:43

			Valores de Referência: V.R.: Masculino	V.R.: Feminino
Eritrograma				
Eritrócitos	*4,42 milhões/mm³*		4.50 - 5,90 milhões/mm³	4.00 - 5,20 milhões/mm³
Hemoglobina	*12,3 g/dL*		14.0 - 18,0 g/dL	12.0 - 16,0 g/dL
Hematócrito	*35,6 %*		41,0 - 54.0 %	36,0 - 45.0 %
R.D.W.	*13,6 %*		10,0 - 15,0 %	10,0 - 15,0 %
V.C.M.	*80,5 µ³*		80,0 - 96.0 µ³	80.0 - 96,0 µ³
H.C.M.	*27,8 pg*		26,0 - 34,0 pg	26,0 - 34,0 pg
C.H.C.M.	*34,6 %*		31,0 - 37,0 %	31,0 - 37,0 %
Leucograma				
Leucócitos		*7.410 /mm³*	4.000 - 10.000/mm³	
Bastonetes	*1,0 %*	*74 /mm³*	0 - 400/mm³	
Segmentados	*50,2 %*	*3.720 /mm³*	1.800 - 7.500/mm³	
Eosinófilos	*2,2 %*	*163 /mm³*	0 - 450/mm³	
Basófilos	*0,3 %*	*22 /mm³*	0 - 100/mm³	
Linfócitos	*40,6 %*	*3.008 /mm³*	1.200 - 5.280/mm³	
Monócitos	*5,7 %*	*422 /mm³*	80 - 800/mm³	
Plaquetas	*236 mil/mm³*		150 - 450 mil/mm³	

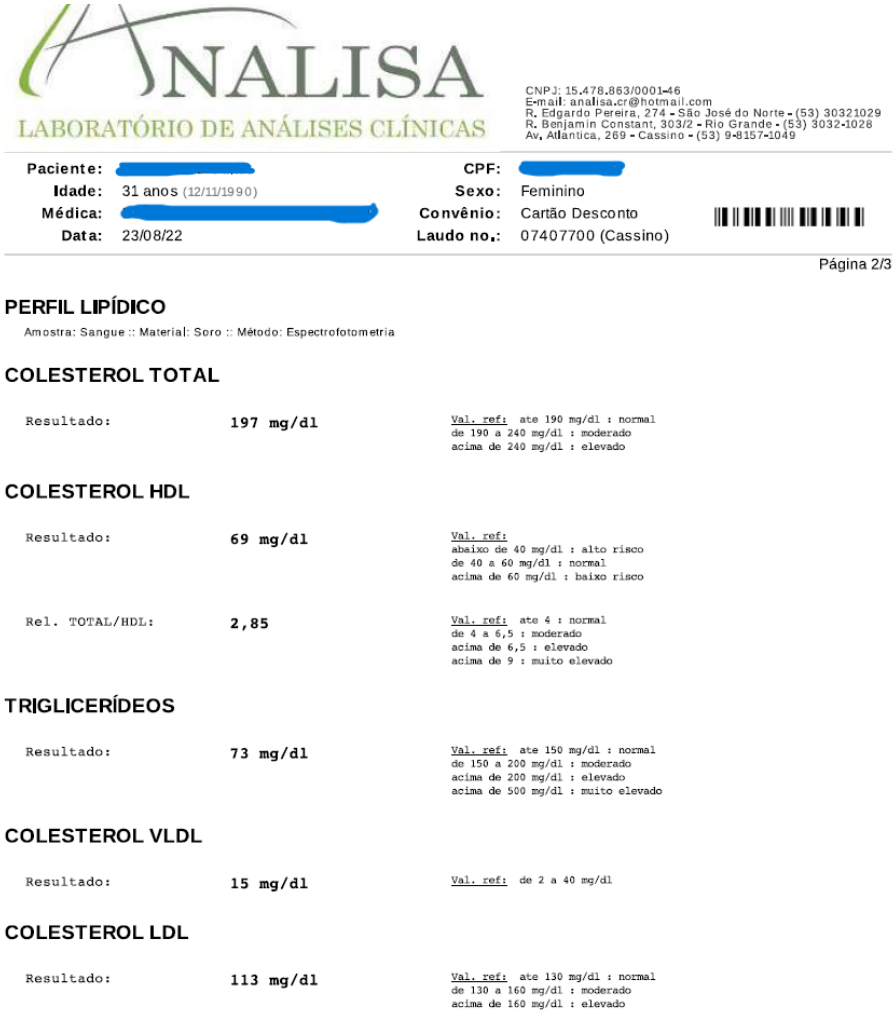
Data de conferência: 16/11/2021

Figura 38: Exemplo de *layout* do laboratório Gram
Fonte: Autora.

- Hospital Fêmina S/A;
- HNSA;
- Unimed Pelotas;
- Unimed Caxias;
- Sergio Franco;
- Perez Limardo;
- Weinmann

Ressalta-se, ainda, que para um mesmo laboratório foi possível observar diferentes *layouts*.

A fim de medir o desempenho da metodologia adotada para a realização da extração de informação dos exames de sangue, optou-se por calcular *Precision*, *Recall* e *F1 Score* [75], definidas e apresentadas anteriormente, para dois testes distintos: o primeiro deles,



ANALISA
LABORATÓRIO DE ANÁLISES CLÍNICAS

CNPJ: 15.478.863/0001-46
E-mail: analisa.cr@hotmail.com
R. Edgardo Pereira, 274 - São José do Norte - (53) 30321029
R. Benjamin Constant, 303/2 - Rio Grande - (53) 3032-1028
Av. Atlantica, 269 - Cassino - (53) 9-8157-1049

Paciente: [redacted]
Idade: 31 anos (12/11/1990)
Médica: [redacted]
Data: 23/08/22

CPF: [redacted]
Sexo: Feminino
Convênio: Cartão Desconto
Laudo no.: 07407700 (Cassino)

Página 2/3

PERFIL LIPÍDICO

Amostra: Sangue :: Material: Soro :: Método: Espectrofotometria

COLESTEROL TOTAL

Resultado:	197 mg/dl	Val. ref: ate 190 mg/dl : normal de 190 a 240 mg/dl : moderado acima de 240 mg/dl : elevado

COLESTEROL HDL

Resultado:	69 mg/dl	Val. ref: abaixo de 40 mg/dl : alto risco de 40 a 60 mg/dl : normal acima de 60 mg/dl : baixo risco
Rel. TOTAL/HDL:	2,85	Val. ref: ate 4 : normal de 4 a 6,5 : moderado acima de 6,5 : elevado acima de 9 : muito elevado

TRIGLICERÍDEOS

Resultado:	73 mg/dl	Val. ref: ate 150 mg/dl : normal de 150 a 200 mg/dl : moderado acima de 200 mg/dl : elevado acima de 500 mg/dl : muito elevado

COLESTEROL VLDL

Resultado:	15 mg/dl	Val. ref: de 2 a 40 mg/dl

COLESTEROL LDL

Resultado:	113 mg/dl	Val. ref: ate 130 mg/dl : normal de 130 a 160 mg/dl : moderado acima de 160 mg/dl : elevado

Figura 39: Exemplo de *layout* do laboratório Analisa
Fonte: Autora.

considerava estas métricas para cada um dos 210 exames de sangue analisados; enquanto que o segundo teste, utilizava as métricas para avaliar o desempenho para cada uma das variáveis de interesse consideradas.

Além disso, em cada um dos testes foram calculadas as métricas considerando o total de 210 exames, mas também, com a separação destes exames em originais e imagens, para que se pudesse dessa forma analisar se o desempenho da metodologia adotada possui diferença significativa quando se tinha um arquivo escaneado ao invés do arquivo original disponibilizado pelos laboratórios.

6.3.1 Análise de desempenho considerando cada um dos exames de sangue

Nesta primeira análise, o objetivo consiste em definir o desempenho apresentado pela metodologia proposta para cada um dos exames de sangue. Para isso, foram utilizadas as métricas de avaliação apresentadas anteriormente: *Precision, Recall* e *F1 Score*. A seguir são apresentados, de forma detalhada, a maneira como cada uma das métricas foi calculada e quais variáveis foram consideradas.

- *Precision*: Calculado a partir da divisão do total de variáveis de interesse presente no exame extraídas corretamente (VP) pelo total de variáveis de interesse presente no exame extraídas;

- *Recall*: Calculado a partir da divisão do total de variáveis de interesse presente no exame extraídas corretamente (VP) pelo total de variáveis de interesse presente no exame;

- *F1 Score*: Calculado a partir da divisão de duas vezes o valor entre o *precision* e o *recall* pela soma do *precision* e do *recall*.

A fim de facilitar o entendimento acerca dos resultados, optou-se por apresentar todos os resultados em tabelas. Porém, a fim de otimização de espaço, optou-se por apresentar a seguir somente a Tabela 18 que mostra apenas o resultado para alguns exames. As tabelas com os dados de todos os exames podem ser encontradas no Anexo B.

Além disso, os 210 exames foram separados em dois grupos: o primeiro deles apresenta aqueles exames de PDFs originais disponibilizados pelos laboratórios; enquanto que o segundo grupo apresenta os exames tidos como "não originais", pois se tratam daqueles exames que foram escaneados e/ou fotografados. A Tabela 19 apresenta os exames que pertencem ao grupo 1, enquanto que a Tabela 20 representa os exames que pertencem ao grupo 2.

Dessa forma, para evitar a repetição dos dados apresentados nas tabelas anteriores, optou-se por apresentar somente os resultados de média e desvio padrão calculados para cada um dos dois grupos, como mostra a Tabela 21.

Analisando a tabela apresentada pode-se observar que existe uma queda no rendimento do sistema proposto quando se considera a extração de informação de imagens. Isto acontece devido ao fato de algumas vezes a cópia/foto/digitalização não ter sido realizada da maneira adequada, ou seja, quando se trata da extração de informação de arquivos que não são os PDFs originais, a qualidade do arquivo analisado interfere no processo de análise e extração dos dados de interesse.

6.3.2 Análise de desempenho considerando cada uma das variáveis de interesse

Além da análise apresentada anteriormente, optou-se por realizar a análise considerando cada uma das variáveis de interesse, a fim de realizar uma melhor avaliação do sistema proposto. Tal análise foi realizada com o intuito de observar de forma mais específica se alguma das variáveis de interesse apresentavam resultado muito divergente das outras variáveis.

Para o cálculo das métricas de avaliação foram considerados as seguintes variáveis:

- *Precision*: Calculado a partir da divisão do total de vezes que a variável foi extraída corretamente pelo total de vezes que a variável foi extraída;

- *Recall*: Calculado a partir da divisão do total de vezes que a variável foi extraída corretamente pelo total de exames que tinha a variável;

- *F1 Score*: Calculado a partir da divisão de duas vezes o valor entre o *precision* e o *recall* pela soma do *precision* e do *recall*.

As Tabelas 22 e 23 apresentam não só os valores obtidos de *precision*, *recall* e *F1 Score* para cada uma das 37 variáveis de interesse analisadas, mas também a média e o desvio padrão destas métricas.

Assim como na análise apresentada anteriormente, também optou-se por realizar este teste considerando os grupos 1 e 2, para que se pudesse verificar a influência que um arquivo não original podia ter em alguma variável de interesse específica. A Tabela 24 apresenta os resultados obtidos.

Diferente dos resultados apresentados para o teste anterior, quando se trata de uma análise por variável de interesse observa-se que os resultados obtidos para o segundo grupo são melhores do que os obtidos para o primeiro grupo.

Ressalta-se que, para ambos os testes apresentados, não se pode ignorar o fato de que as amostras consideradas para a divisão em grupo 1 e grupo 2 são diferentes, o que pode influenciar nos resultados obtidos.

6.3.3 Análise de Complexidade

Para analisar a complexidade computacional do algoritmo de extração proposto, seja n o número de documentos PDF disponíveis no diretório de entrada, m o tamanho médio de cada documento (medido em termos do número de páginas ou da quantidade de conteúdo textual extraído) e k o número de parâmetros laboratoriais a serem extraídos de cada documento.

Inicialmente, o algoritmo percorre todos os arquivos PDF contidos no diretório de destino, resultando em uma dependência linear em relação ao número de documentos, isto é, $O(n)$. Para cada documento, o método invoca a rotina de processamento responsável pela leitura e extração das informações textuais do PDF. Essa etapa constitui a operação computacionalmente mais custosa, pois exige a análise do conteúdo do documento antes que o processo de extração possa ser realizado.

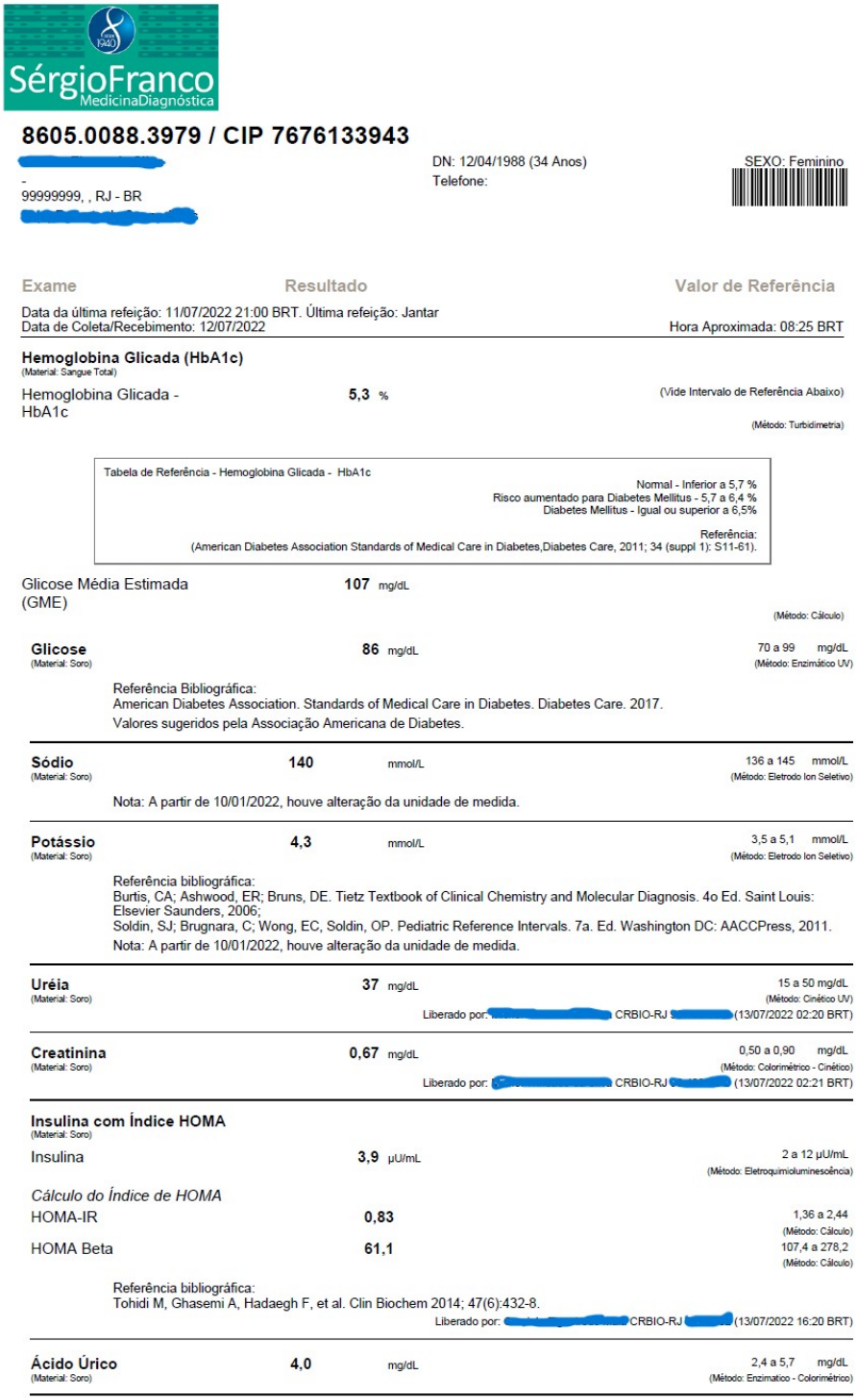
Posteriormente, o algoritmo executa uma série de buscas para identificar e extrair os parâmetros laboratoriais de interesse. Considerando que aproximadamente (k) atributos laboratoriais são avaliados para cada documento e que cada busca pode requerer a varredura do conteúdo extraído, o custo computacional associado a um único documento pode ser expresso como $O(k \cdot m)$.

Portanto, a complexidade temporal total do algoritmo é dada por: $T(n) = O(n \cdot k \cdot m)$.

Como (k) corresponde a um número fixo de parâmetros laboratoriais e não aumenta

com o tamanho da entrada, ele pode ser tratado como um fator constante. Consequentemente, a complexidade temporal pode ser simplificada para: $T(n) = O(n \cdot m)$.

Em relação ao consumo de memória, o algoritmo armazena os atributos extraídos em uma estrutura de dicionário e acumula os registros extraídos antes de exportá-los para um arquivo CSV. Como o número de atributos é fixo, a memória necessária para processar um documento individual permanece constante, isto é, $O(1)$. Entretanto, como os resultados extraídos de todos os documentos processados são mantidos em memória até a geração do arquivo CSV, a complexidade espacial total cresce linearmente com o número de documentos, resultando em uma complexidade espacial de: $S(n) = O(n)$.



SérgioFranco MedicinaDiagnóstica

8605.0088.3979 / CIP 7676133943

[redacted]
-
99999999, , RJ - BR
[redacted]

DN: 12/04/1988 (34 Anos)
Telefone:

SEXO: Feminino

Exame	Resultado	Valor de Referência
Data da última refeição: 11/07/2022 21:00 BRT. Última refeição: Jantar Data de Coleta/Recebimento: 12/07/2022		Hora Aproximada: 08:25 BRT
Hemoglobina Glicada (HbA1c) (Material: Sangue Total)		
Hemoglobina Glicada - HbA1c	**5,3** %	(Vide Intervalo de Referência Abaixo) (Método: Turbidimetria)

Tabela de Referência - Hemoglobina Glicada - HbA1c

Normal - Inferior a 5,7 %
Risco aumentado para Diabetes Mellitus - 5,7 a 6,4 %
Diabetes Mellitus - Igual ou superior a 6,5%

Referência:
(American Diabetes Association Standards of Medical Care in Diabetes,Diabetes Care, 2011; 34 (suppl 1): S11-61).

Glicose Média Estimada (GME)	**107** mg/dL	(Método: Cálculo)
Glicose (Material: Soro)	**86** mg/dL	70 a 99 mg/dL (Método: Enzimático UV)

Referência Bibliográfica:
American Diabetes Association. Standards of Medical Care in Diabetes. Diabetes Care. 2017.
Valores sugeridos pela Associação Americana de Diabetes.

Sódio (Material: Soro)	**140** mmol/L	136 a 145 mmol/L (Método: Eletrodo Ion Seletivo)

Nota: A partir de 10/01/2022, houve alteração da unidade de medida.

Potássio (Material: Soro)	**4,3** mmol/L	3,5 a 5,1 mmol/L (Método: Eletrodo Ion Seletivo)

Referência bibliográfica:
Burtis, CA; Ashwood, ER; Bruns, DE. Tietz Textbook of Clinical Chemistry and Molecular Diagnosis. 4o Ed. Saint Louis: Elsevier Saunders, 2006;
Soldin, SJ; Brugnara, C; Wong, EC, Soldin, OP. Pediatric Reference Intervals. 7a. Ed. Washington DC: AACCPress, 2011.
Nota: A partir de 10/01/2022, houve alteração da unidade de medida.

Uréia (Material: Soro)	**37** mg/dL	15 a 50 mg/dL (Método: Cinético UV)
	Liberado por: [redacted] CRBIO-RJ [redacted] (13/07/2022 02:20 BRT)	
Creatinina (Material: Soro)	**0,67** mg/dL	0,50 a 0,90 mg/dL (Método: Colorimétrico - Cinético)
	Liberado por: [redacted] CRBIO-RJ [redacted] (13/07/2022 02:21 BRT)	
Insulina com Índice HOMA (Material: Soro)		
Insulina	**3,9** µU/mL	2 a 12 µU/mL (Método: Eletroquimioluminescência)
Cálculo do Índice de HOMA		
HOMA-IR	**0,83**	1,36 a 2,44 (Método: Cálculo)
HOMA Beta	**61,1**	107,4 a 278,2 (Método: Cálculo)

Referência bibliográfica:
Tohidi M, Ghasemi A, Hadaegh F, et al. Clin Biochem 2014; 47(6):432-8.

Liberado por: [redacted] CRBIO-RJ [redacted] (13/07/2022 16:20 BRT)

Ácido Úrico (Material: Soro)	**4,0** mg/dL	2,4 a 5,7 mg/dL (Método: Enzimatico - Colorimétrico)

Figura 40: Exemplo de *layout* do laboratório Sérgio Franco
Fonte: Autora.

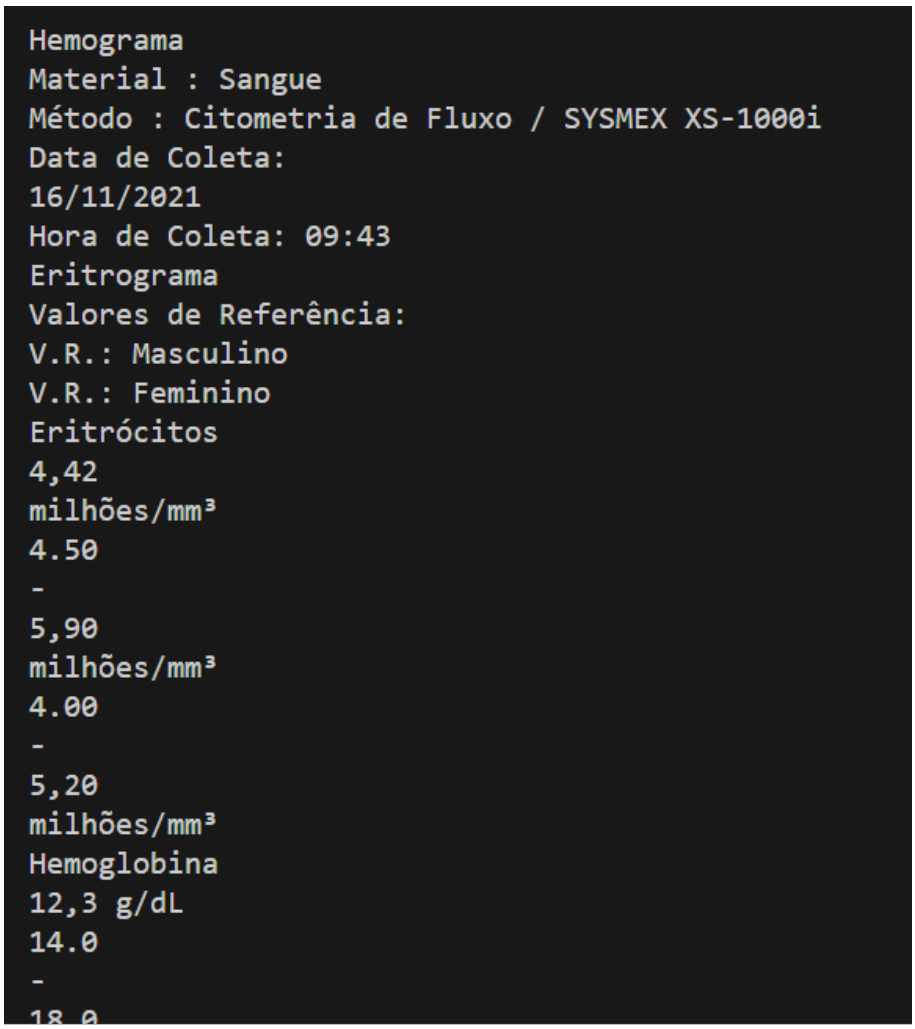


Figura 41: Texto extraído para o exemplo de *layout* do laboratório Gram
Fonte: Autora.

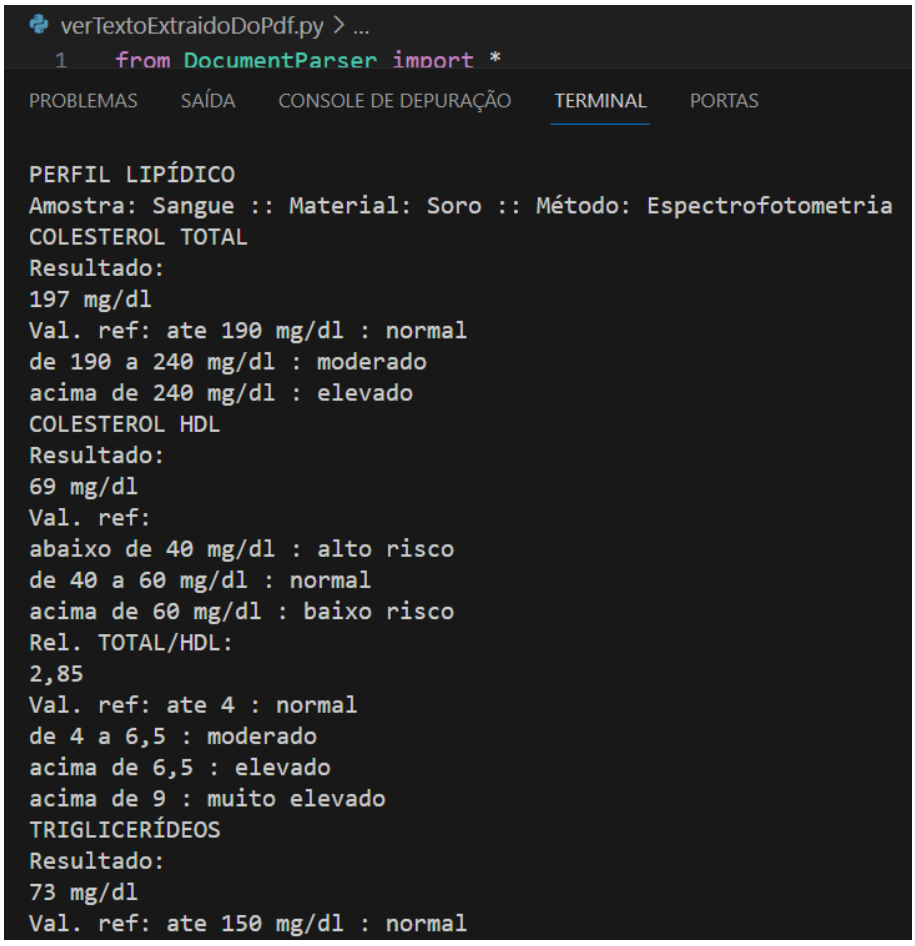


Figura 42: Texto extraído para o exemplo de *layout* do laboratório Analisa
Fonte: Autora.

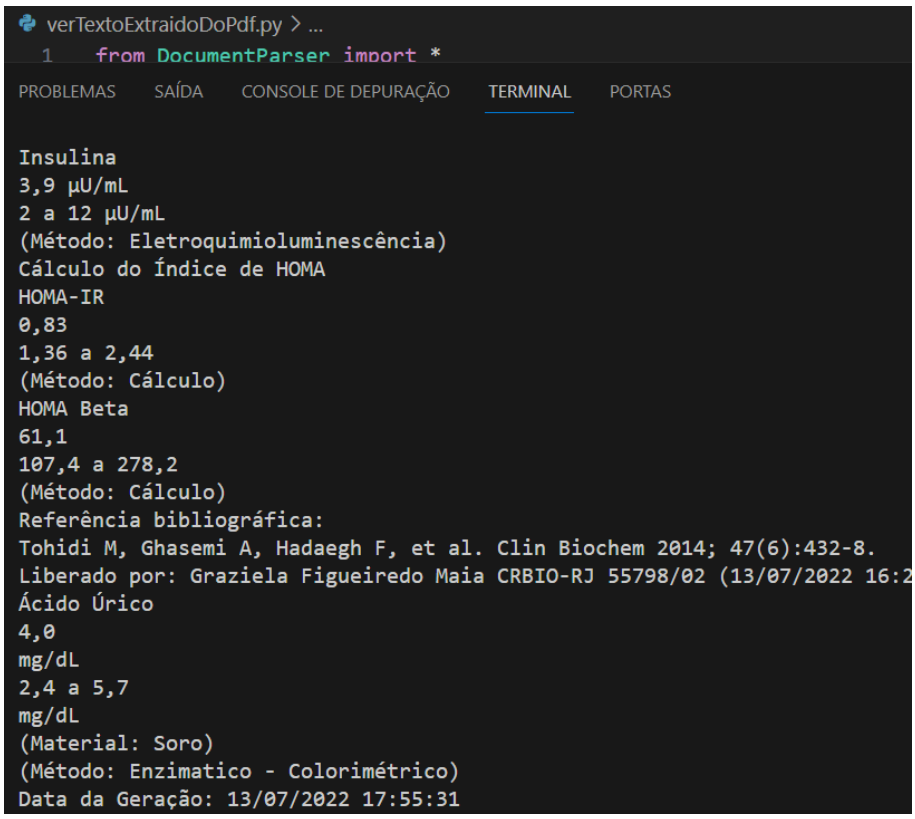


Figura 43: Texto extraído para o exemplo de *layout* do laboratório Sérgio Franco
Fonte: Autora.

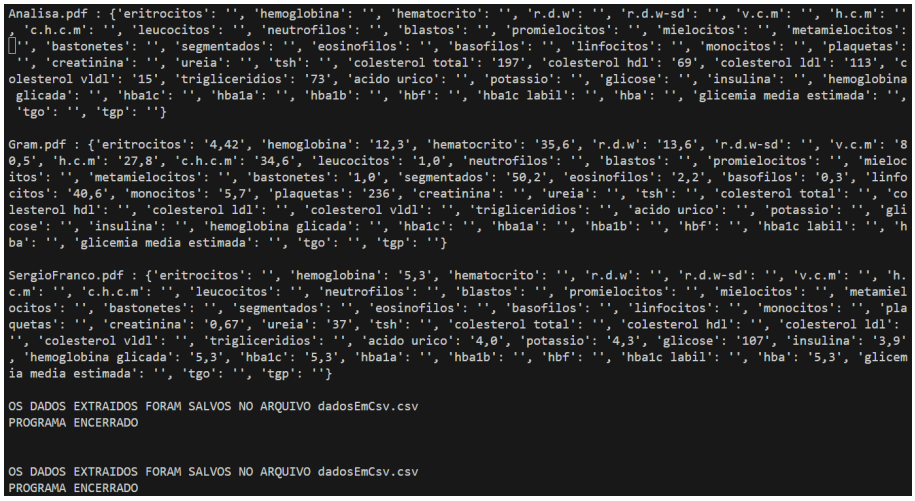


Figura 44: Variáveis de interesse obtidas a partir da utilização da técnica *form understanding* para os três *layouts*
Fonte: Autora.

	A	B	C	D
7		Analisa.pdf	Gram.pdf	SergioFranco.pdf
8	eritrocitos		4,42	
9	hemoglobina		12,3	5,3
10	hematocrito		35,6	
11	r.d.w		13,6	
12	r.d.w-sd			
13	v.c.m		80,5	
14	h.c.m		27,8	
15	c.h.c.m		34,6	
16	leucocitos		1	
17	neutrofilos			
18	blastos			
19	promielocitos			
20	mielocitos			
21	metamielocitos			
22	bastonetes		1	
23	segmentados		50,2	
24	eosinofilos		2,2	
25	basofilos		0,3	
26	linfocitos		40,6	
27	monocitos		5,7	
28	plaquetas		236	
29	creatinina			0,67
30	ureia			37
31	tsh			
32	colesterol total	197		
33	colesterol hdl	69		
34	colesterol ldl	113		
35	colesterol vldl	15		
36	trigliceridios	73		
37	acido urico			4
38	potassio			4,3
39	glicose			107
40	insulina			3,9
41	hemoglobina glicada			5,3

dadosEmCsv

Figura 45: Variáveis de interesse obtidas a partir da utilização da técnica *form understanding* para os três *layouts* apresentadas no .csv gerado

Fonte: Autora.

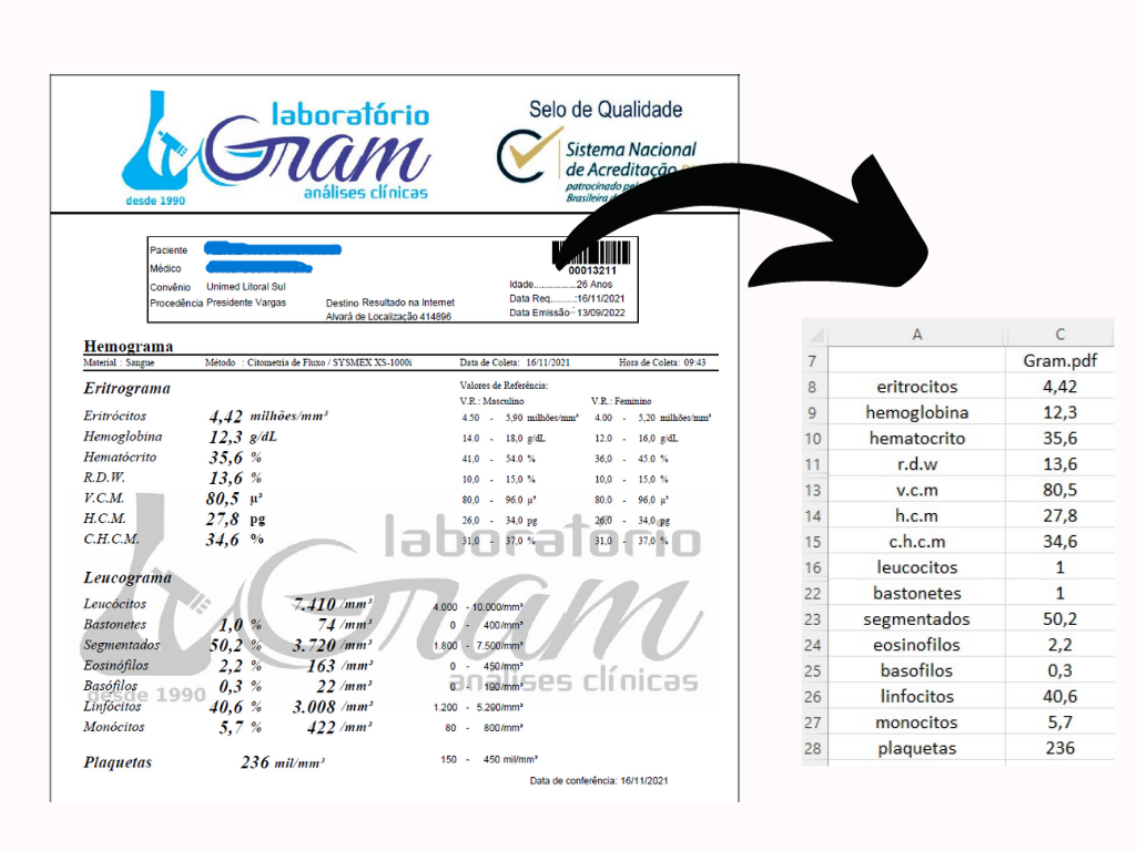


Figura 46: Comparação do resultado obtido para o exemplo de *layout* do laboratório Gram
Fonte: Autora.

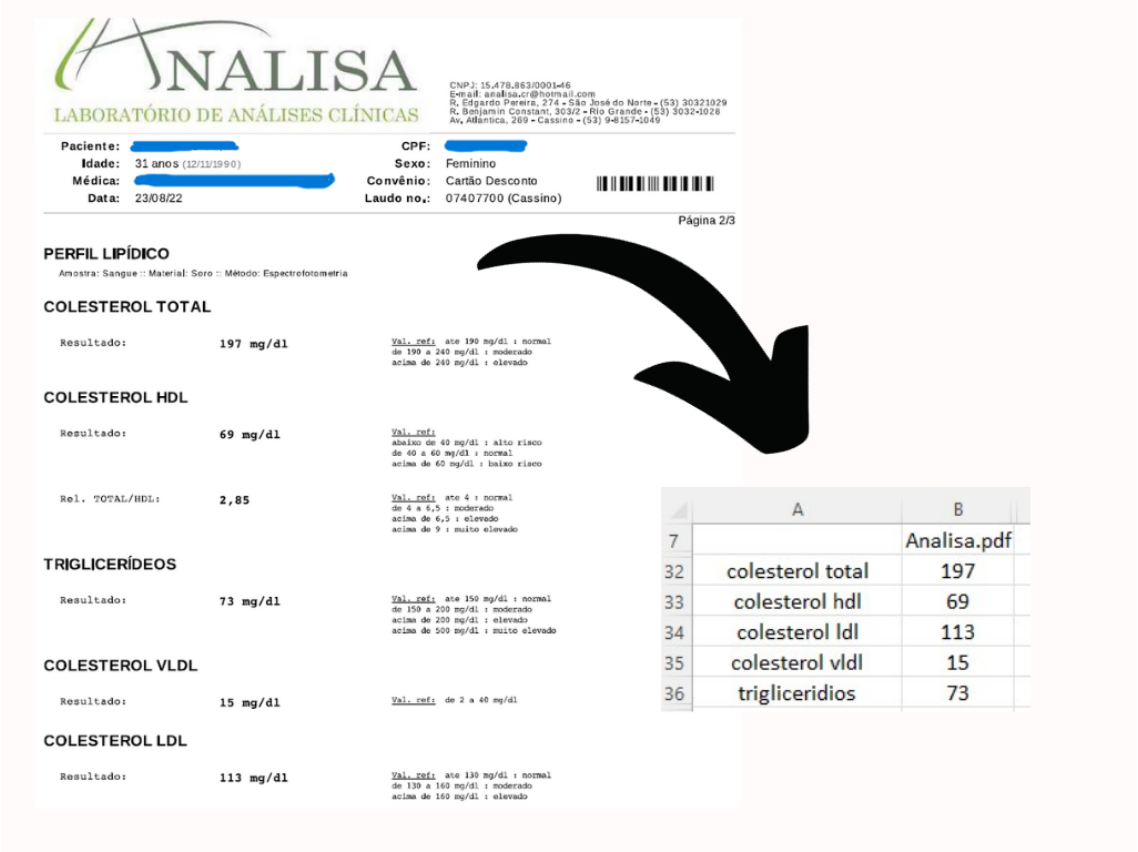


Figura 47: Comparação do resultado obtido para o exemplo de *layout* do laboratório Analisa

Fonte: Autora.

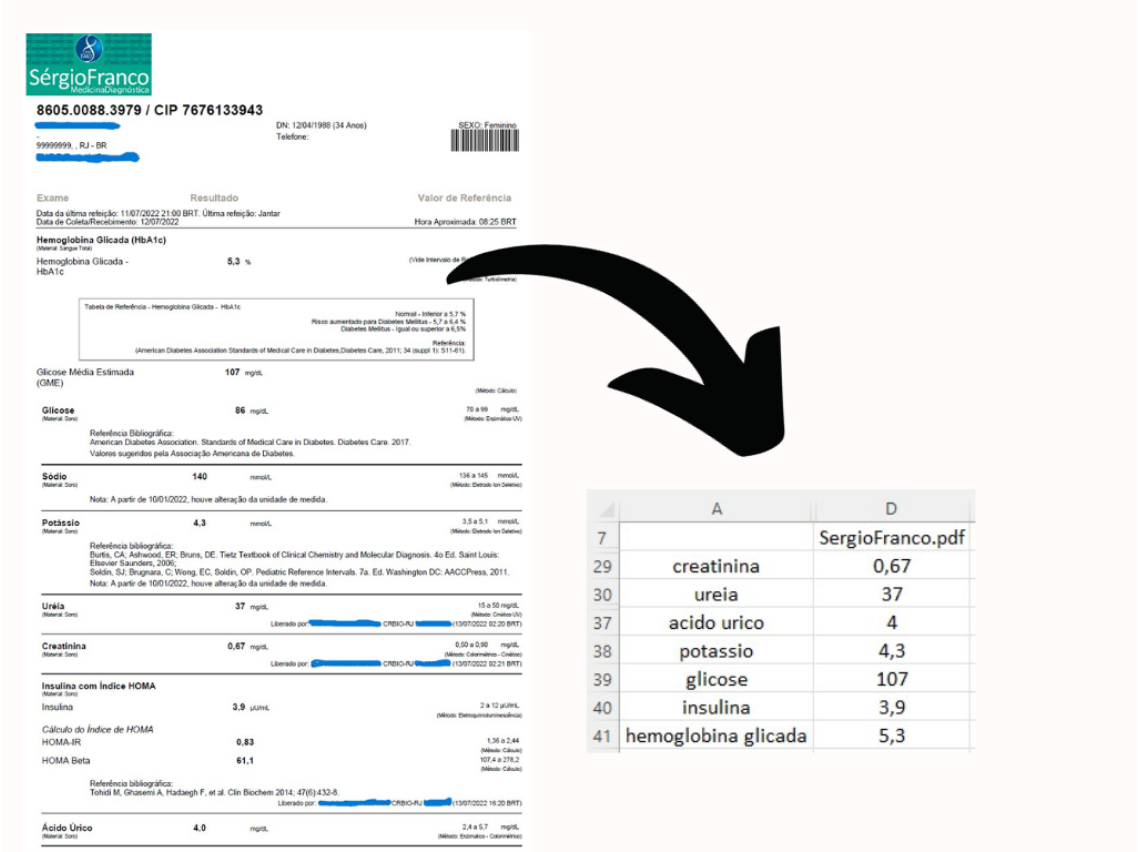


Figura 48: Comparação do resultado obtido para o exemplo de *layout* do laboratório Sérgio Franco

Fonte: Autora.

Tabela 18: Resultados obtidos considerando alguns exames de sangue

Exame	Total de variáveis de interesse presente no exame	Total de variáveis de interesse extraídas	Total de variáveis de interesse presentes no exame NÃO extraídas	Total de variáveis de interesse presente no exame extraídas	Total de variáveis de interesse presente no exames extraídas corretamente	Total de variáveis de interesse presente no exame extraídas de maneira errada	Total de variáveis de interesse extraídas NÃO presentes no exame	Precision	Recall	F1 Score
1	18	18	0	18	15	3	0	83%	83%	83%
2	21	21	0	21	17	4	0	81%	81%	81%
3	23	23	0	23	20	3	0	87%	87%	87%
4	24	24	0	24	22	2	0	92%	92%	92%
5	21	21	0	21	20	1	0	95%	95%	95%
6	19	19	0	19	17	2	0	89%	89%	89%
7	23	21	2	21	20	1	0	95%	87%	91%
...	...	...	...	...	...	...	...	...	...	...
198	28	28	0	28	22	6	0	79%	79%	79%
199	27	27	0	27	24	3	0	89%	89%	89%
200	18	18	0	18	13	5	0	72%	72%	72%

Fonte: Autora.

Tabela 19: Relação de exames de sangue pertencentes ao Grupo 1

1	13	25	37	49	61	73	85	97	109	132	210
2	14	26	38	50	62	74	86	98	110	133	
3	15	27	39	51	63	75	87	99	111	134	
4	16	28	40	52	64	76	88	100	112	135	
5	17	29	41	53	65	77	89	101	113	136	
6	18	30	42	54	66	78	90	102	114	137	
7	19	31	43	55	67	79	91	103	115	138	
8	20	32	44	56	68	80	92	104	127	139	
9	21	33	45	57	69	81	93	105	128	140	
10	22	34	46	58	70	82	94	106	129	141	
11	23	35	47	59	71	83	95	107	130	142	
12	24	36	48	60	72	84	96	108	131	143	

Fonte: Autora.

Tabela 20: Relação de exames de sangue pertencentes ao Grupo 2

116	144	155	166	177	188	199
117	145	156	167	178	189	200
118	146	157	168	179	190	201
119	147	158	169	180	191	202
120	148	159	170	181	192	203
121	149	160	171	182	193	204
122	150	161	172	183	194	205

Fonte: Autora.

Tabela 21: Resumo das métricas considerando cada um dos três testes realizados

Grupo	Amostras	Média *Precision*	Média *Recall*	Média *F1 Score*	Desvio Padrão *Precision*	Desvio Padrão *Recall*	Desvio Padrão *F1 Score*
1	133	90%	88%	90%	15%	16%	11%
2	77	87%	85%	87%	15%	15%	11%
Total	210	89%	87%	89%	15%	16%	11%

Fonte: Autora.

Tabela 22: Resultados obtidos para cada uma das variáveis de interesse considerando todos os exames de sangue

Variável de Interesse	**Total de vezes que a variável foi extraída corretamente**	**Total de vezes que a variável foi extraída**	**Total de exames que tinham a variável de interesse**	***Precision***	***Recall***	***F1 Score***
eritrócitos	162	177	174	92%	93%	92%
hemoglobina	163	185	174	88%	94%	91%
hematócrito	168	172	174	98%	97%	97%
r.d.w	153	174	174	88%	88%	88%
r.d.w-sd	1	1	1	100%	100%	100%
v.c.m	166	173	174	96%	95%	96%
h.c.m	169	172	174	98%	97%	98%
c.h.c.m	161	176	174	91%	93%	92%
leucócitos	136	177	174	77%	78%	77%
neutrófilos	51	51	51	100%	100%	100%
blastos	42	42	42	100%	100%	100%
promielócitos	35	35	36	100%	97%	99%
mielócitos	49	49	49	100%	100%	100%
metamielócitos	58	59	58	98%	100%	99%
bastonetes	126	141	144	89%	88%	88%
segmentados	129	130	131	99%	98%	99%
eosinófilos	173	174	174	99%	99%	99%
basófilos	130	135	135	96%	96%	96%
linfócitos	170	174	174	98%	98%	98%
monócitos	168	174	174	97%	97%	97%
plaquetas	151	173	166	87%	91%	89%

Fonte: Autora.

Tabela 23: Resultados obtidos para cada uma das variáveis de interesse considerando todos os exames de sangue (continuação)

Variável de Interesse	**Total de vezes que a variável foi extraída corretamente**	**Total de vezes que a variável foi extraída**	**Total de exames que tinham a variável de interesse**	***Precision***	***Recall***	***F1 Score***
creatinina	86	117	124	74%	69%	71%
ureia	61	77	88	79%	69%	74%
tsh	78	93	97	84%	80%	82%
colesterol total	91	99	114	92%	80%	85%
colesterol hdl	88	108	109	81%	81%	81%
colesterol ldl	87	97	98	90%	89%	89%
colesterol vldl	70	81	82	86%	85%	86%
triglicerídios	94	112	112	84%	84%	84%
acido urico	28	38	38	74%	74%	74%
potássio	19	35	39	54%	49%	51%
glicose	80	132	137	61%	58%	59%
insulina	12	13	13	92%	92%	92%
hemoglobina glicada	53	61	62	87%	85%	86%
glicemia media estimada	34	34	34	100%	100%	100%
tgo	64	71	84	90%	76%	83%
tgp	55	71	83	77%	66%	71%
MÉDIA				**89%**	**87%**	**88%**
DESVIO PADRÃO				**11%**	**13%**	**12%**

Fonte: Autora.

Tabela 24: Resumo das métricas considerando as variáveis de interesse para cada um dos três testes realizados

Grupo	Amostras	Média *Precision*	Média *Recall*	Média *F1 Score*	Desvio Padrão *Precision*	Desvio Padrão *Recall*	Desvio Padrão *F1 Score*
1	133	88%	82%	84%	18%	20%	19%
2	77	89%	87%	88%	14%	15%	14%
Total	210	89%	87%	88%	11%	13%	12%

Fonte: Autora.

7 MODELAGEM NEURO-*FUZZY* PARA APOIO À DECISÃO CLÍNICA DE DM2

O presente capítulo apresenta os bancos de dados, a metodologia de modelagem e os resultados experimentais associados aos dois níveis do sistema neuro-*fuzzy* proposto para apoio à decisão clínica no contexto do Diabetes *Mellitus* Tipo 2 (DM2). Para evitar ambiguidades clínicas, adota-se, ao longo deste capítulo, a seguinte distinção conceitual: o Nível 01 corresponde a uma implementação *fuzzy* dos critérios laboratoriais de classificação glicêmica, com base em glicemia de jejum e hemoglobina glicada (HbA1c); já o Nível 02 corresponde a uma estratificação transversal do risco metabólico atual, construída a partir de um perfil clínico instantâneo, sem pretensão de estimar incidência futura ou tempo até o desenvolvimento da doença.

7.1 Banco de Dados para o apoio à decisão clínica de DM2

Esta seção tem como objetivo apresentar os bancos de dados utilizados para realizar o processo de apoio à decisão clínica de DM2, para que em seguida possa ser apresentada a metodologia adotada para realizar tal processo.

A proposta inicial do estudo consistia na construção de um banco de dados específico, formado a partir do envio espontâneo de exames de sangue pelos participantes. Para viabilizar essa etapa, foi necessária a submissão do projeto ao Comitê de Ética em Pesquisa, uma vez que o estudo envolveria a utilização de dados reais e sensíveis. Assim, o projeto intitulado **Repositório Dinâmico de Informações Pregressas e Atuais de Saúde do Usuário** foi submetido ao comitê e posteriormente aprovado, autorizando o uso de dados fornecidos voluntariamente pelos participantes. A consulta ao projeto e à sua respectiva aprovação pode ser realizada por meio do número de identificação **CAAE 60939822.5.0000.5324**.

Após a análise dos dados inicialmente recebidos, constatou-se que a quantidade de informações coletadas não era suficiente para a formação do banco de dados pretendido. Diante disso, tornou-se necessário identificar outra fonte de dados que pudesse suprir as necessidades do estudo. Para permitir o uso de novos conjuntos de dados, foi solicitada

uma emenda ao Comitê de Ética, a qual, após aprovada, possibilitou a formalização de parceria com o Hospital Universitário Dr. Miguel Riet Corrêa Jr. (HU-FURG/Ebserh), responsável pelo fornecimento de dados disponíveis em seu banco institucional.

Entretanto, o banco de dados disponibilizado pelo hospital apresentou limitações quanto ao número de registros e às variáveis necessárias para a categorização transversal do risco de desenvolvimento de diabetes *mellitus*. Considerando que muitos dos exames requeridos são específicos para investigação da doença, foi necessário buscar uma alternativa que permitisse ampliar o conjunto de dados. Dessa forma, optou-se pela utilização do simulador *Sint Multiverse*, desenvolvido no doutorado de Mateus Jung (PPGMC, FURG, em andamento sob a orientação da Profa. Graçaliz Dimuro, publicado em [76]), capaz de gerar, em grande escala, dados sintéticos contendo as variáveis essenciais ao processo de categorização do risco de desenvolvimento do DM2.

A seguir, serão apresentados detalhadamente os dois bancos de dados utilizados nas modelagens: o primeiro, composto por dados reais provenientes do hospital parceiro; e o segundo, formado por dados sintéticos gerados pelo simulador *Sint Multiverse*.

7.1.1 Dados Reais: Hospital Universitário (HU)

Os dados recebidos do HU referem-se às informações coletadas de pacientes atendidos em diversos setores da instituição. Entretanto, o processo de digitalização e registro eletrônico de dados no hospital é relativamente recente, o que explica o tamanho reduzido do banco disponibilizado em comparação ao inicialmente previsto.

Todos os dados fornecidos foram previamente tratados pela equipe do hospital, garantindo a remoção de informações que permitissem a identificação dos pacientes. Além disso, os dados foram organizados em arquivos separados, cada um correspondente a um tipo específico de exame. De forma resumida, os arquivos recebidos foram:

- Anamnese;
- Evolução clínica;
- Colesterol HDL;
- Colesterol LDL;
- Colesterol total;
- Colesterol VLDL;
- Creatinina;
- GAMA GT;
- Glicose;

- Hemograma completo;
- Hemoglobina glicada;
- Potássio;
- TGO;
- TGP;
- Triglicerídeos;
- Ureia;
- Informações básicas dos pacientes (sexo, raça e idade).

Ao todo, foram recebidas informações referentes a 16.799 pacientes. No entanto, nem todos possuíam registros completos para todos os exames, e muitos apresentavam múltiplas coletas de um mesmo tipo de exame, situação comum entre pacientes hospitalizados que realizam exames diariamente. Assim, tornou-se necessário aplicar procedimentos de tratamento e filtragem a fim de preparar os dados para a modelagem *fuzzy*, descrita posteriormente.

Como já se conheciam previamente as variáveis necessárias para cada uma das modelagens, optou-se por organizar os dados considerando separadamente as duas abordagens propostas.

Para a modelagem destinada à classificação diagnóstica laboratorial do paciente, somente dois exames eram necessários: glicose e hemoglobina glicada (HbA1c).

Durante o processo de filtragem, foram identificados 11.219 registros de glicose e 5.524 registros de hemoglobina glicada. Entretanto, o interesse do estudo concentrava-se na intersecção desses conjuntos, ou seja, nos pacientes que possuíam ambos os exames. Essa etapa resultou em 3.739 pacientes com medições disponíveis para glicose e HbA1c.

Contudo, além de identificar pacientes com ambos os exames, era necessário determinar se esses indivíduos eram diabéticos, pré-diabéticos ou não diabéticos, a fim de utilizar essa classificação como *target* no treinamento e avaliação da modelagem. Para isso, foram analisados os *dataframes* de anamnese e evolução clínica, buscando indícios de diabetes, tais como registros médicos, diagnósticos descritos ou uso de medicamentos típicos do tratamento do DM2. A Figura 49 ilustra a intersecção necessária entre os diferentes conjuntos de dados.

Após essa etapa, verificou-se que 3.577 dos 3.739 pacientes possuíam todas as informações necessárias, incluindo inclusive, a classificação diagnóstica.

Como mencionado anteriormente, alguns pacientes apresentavam múltiplas coletas para um mesmo exame. Para representar melhor o estado clínico desses indivíduos, optou-

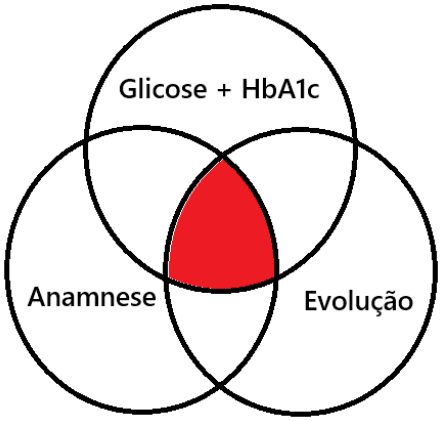


Figura 49: Diagrama representando a intersecção dos dados necessários para a construção do banco destinado à modelagem de classificação diagnóstica laboratorial
Fonte: Autora.

se por utilizar a média dos valores de glicose e hemoglobina glicada quando mais de uma medição estava disponível.

Concluída a filtragem, procedeu-se à organização dos dados para utilização na modelagem. Para isso, foram criados dois arquivos no formato `.csv`: o primeiro, denominado **data**, ilustrado na Figura 50, contendo as médias de glicose e hemoglobina glicada de cada um dos 3.577 pacientes; e o segundo, denominado ***target***, ilustrado na Figura 51, contendo a classificação diagnóstica (diabético, pré-diabético ou não diabético).

	A	B	C
1	**COD_PACIENTE**	**MEDIA_GLICOSE**	**MEDIA_HBA1C**
2	13	92	6.8
3	18	91	5.4
4	24	144	7.7
5	26	190	9.6
6	30	115	6
7	38	105	5
8	44	131	5.8
9	45	176.2	9.3
10	50	96	5.8
11	53	130	6.9
12	55	116	6.2
13	58	94	6.3
14	60	128	6
15	62	95	5.1
16	66	204	9.9
17	67	97	5.4

Figura 50: Exemplo do arquivo **data**, contendo os valores médios de glicose e hemoglobina glicada dos pacientes.
Fonte: Autora.

Ressalta-se que, apesar da SBD e também da ADA [15] (Associação Americana de Diabetes), estabelecerem que o diagnóstico de DM2 requer dois testes alterados (mesma

	B	C	D
1	**Não-Diabetes**	**Pré-Diabetes**	**Diabetes**
2	0	0	1
3	1	0	0
4	0	0	1
5	0	0	1
6	0	1	0
7	0	1	0
8	0	1	0
9	0	0	1
10	0	1	0
11	0	0	1
12	0	1	0
13	0	1	0
14	0	1	0
15	1	0	0
16	0	0	1
17	1	0	0

Figura 51: Exemplo do arquivo ***target***, contendo a classificação diagnóstica dos pacientes. Fonte: Autora.

amostra com dois marcadores diferentes ou repetição do exame em segunda visita), com exceção para os casos de glicemia $\geq$ 200 mg/dL acompanhada de sintomas clássicos (poliúria, polidipsia, perda ponderal), o modelo nível 1 opera sobre uma única medida por paciente.

Portanto, após a utilização das informações providas pelo hospital, conclui-se que o banco de dados proveniente do HU continha informações suficientes para a realização da primeira modelagem.

Por outro lado, verificou-se que o banco não possuía informações suficientes para a segunda modelagem, a qual visava classificar o risco de desenvolvimento de DM2. Isso ocorreu porque o conjunto fornecido pelo hospital não incluía, na maioria dos casos, os dados necessários para o cálculo do IMC, já que peso e altura raramente estavam disponíveis simultaneamente. Além disso, mesmo os poucos pacientes que possuíam essas informações não pertenciam à intersecção desejada com as demais variáveis necessárias, tornando inviável o uso do banco do HU na segunda modelagem.

7.1.1.1 Caracterização da Amostra

A base de dados utilizada neste estudo foi composta por 3.577 indivíduos, contendo informações referentes ao sexo, raça, idade, glicemia média, hemoglobina glicada média e classificação glicêmica.

Em relação ao sexo, observou-se predominância de indivíduos do sexo feminino, cor-

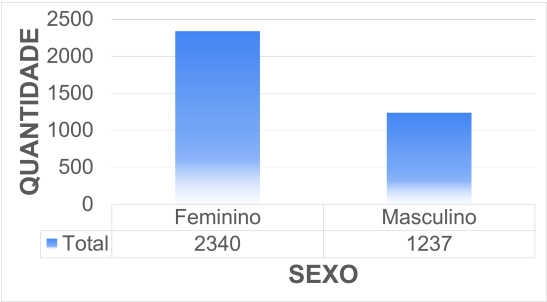
respondendo a 2.340 participantes (65,4%), enquanto 1.237 participantes (34,6%) eram do sexo masculino.

Quanto à raça, a maioria dos indivíduos foi classificada como branca, totalizando 2.422 participantes (67,7%). Os indivíduos autodeclarados pretos corresponderam a 608 registros (17,0%), enquanto 542 participantes (15,2%) não apresentaram declaração racial. Adicionalmente, foram identificados cinco indivíduos classificados como amarelos, representando menos de 0,1% da amostra.

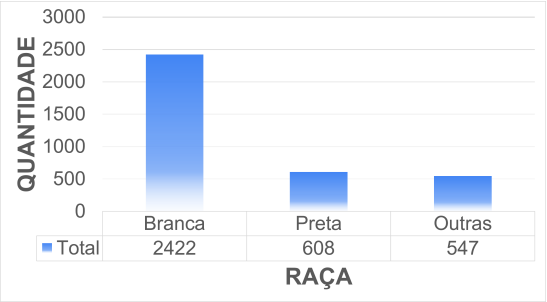
A idade dos participantes variou entre 2 e 96 anos, apresentando média de 56,90 ± 17,02 anos. Considerando os indivíduos adultos, observou-se maior concentração nas faixas etárias entre 50 e 79 anos, com destaque para a faixa de 60 a 69 anos, que apresentou a maior frequência de registros.

Quanto à classificação glicêmica, 1.664 indivíduos (46,5%) foram classificados como pré-diabéticos, 1.068 (29,9%) como diabéticos e 845 (23,6%) como não diabéticos. Essa distribuição evidencia uma predominância de indivíduos com alterações glicêmicas, característica relevante para o desenvolvimento e validação dos modelos propostos neste trabalho.

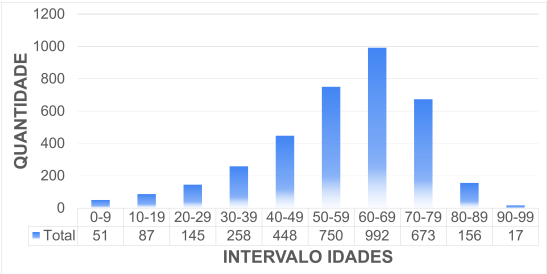
A Figura 52 apresenta a caracterização da amostra utilizada neste estudo, evidenciando a distribuição dos participantes quanto ao sexo, raça, faixa etária e classificação glicêmica.



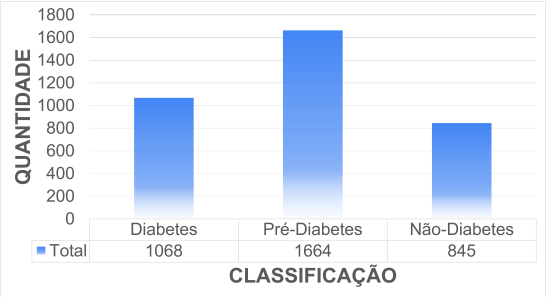
(a) Distribuição dos indivíduos segundo o sexo.



(b) Distribuição dos indivíduos segundo a raça.



(c) Distribuição dos indivíduos por faixa etária.



(d) Distribuição dos indivíduos segundo a classificação glicêmica.

Figura 52: Caracterização da amostra do HU. São apresentadas as distribuições dos participantes quanto ao sexo, raça, faixa etária e classificação glicêmica.

Por outro lado, a Tabela 25 apresenta um resumo das principais características da amostra utilizada neste trabalho.

Tabela 25: Características da amostra utilizada no estudo - Dados do HU.

Característica	**Valor**
Número de indivíduos	3577
Idade (média $\pm$ DP)	$56,90 \pm 17,02$ anos
Idade mínima – máxima	2 – 96 anos
Feminino	2340 (65,4%)
Masculino	1237 (34,6%)
Branca	2422 (67,7%)
Preta	608 (17,0%)
Sem declaração	542 (15,2%)
Amarela	5 (0,1%)
Não-Diabetes	845 (23,6%)
Pré-Diabetes	1664 (46,5%)
Diabetes	1068 (29,9%)

DP: desvio padrão. Os valores são apresentados como média $\pm$ desvio padrão ou número de indivíduos (%).

Além disso, optou-se por apresentar uma caracterização da amostra estratificada segundo a classificação glicêmica, como mostra a Tabela 26. Observa-se que os grupos Pré-Diabetes e Diabetes apresentaram médias de idade superiores à observada no grupo Não-Diabetes, indicando maior concentração de indivíduos com alterações glicêmicas em faixas etárias mais avançadas.

Em relação ao sexo, verificou-se predominância de indivíduos do sexo feminino em todas as classes analisadas, representando 69,2% dos participantes classificados como Não-Diabetes, 65,1% dos Pré-Diabéticos e 62,9% dos Diabéticos.

Quanto à raça, observou-se predominância de indivíduos brancos em todos os grupos, correspondendo a aproximadamente dois terços da amostra em cada classificação glicêmica. As proporções de indivíduos pretos e sem declaração racial permaneceram relativamente estáveis entre as diferentes classes, indicando distribuição demográfica semelhante entre os grupos analisados.

7.1.2 Dados Sintéticos: Simulador *Sint Multiverse*

Diante da necessidade de identificar um *dataset* capaz de fornecer as informações requeridas para a categorização transversal do perfil de risco metabólico associado ao DM2, optou-se pela utilização de um conjunto de dados sintéticos gerado a partir do simulador de população *Sint Multiverse*. Esse simulador é capaz de produzir populações inteiras com base em informações previamente fornecidas e em cálculos probabilísticos, permitindo a geração de um conjunto de dados com o tamanho e as variáveis desejadas.

Tabela 26: Caracterização da amostra segundo a classificação glicêmica - Dados HU.

Característica	**Não-Diabetes**	**Pré-Diabetes**	**Diabetes**	**Total**
n	845	1664	1068	3577
Idade (anos)	$46,64 \pm 19,70$	$59,82 \pm 14,79$	$60,47 \pm 14,62$	$56,90 \pm 17,02$
Sexo feminino	585 (69,2%)	1083 (65,1%)	672 (62,9%)	2340 (65,4%)
Sexo masculino	260 (30,8%)	581 (34,9%)	396 (37,1%)	1237 (34,6%)
Branca	569 (67,3%)	1113 (66,9%)	740 (69,3%)	2422 (67,7%)
Preta	150 (17,8%)	279 (16,8%)	179 (16,8%)	608 (17,0%)
Sem declaração	126 (14,9%)	268 (16,1%)	148 (13,9%)	542 (15,2%)

Valores expressos como média $\pm$ desvio padrão ou número de indivíduos (%). Os percentuais foram calculados em relação ao total de indivíduos de cada classe glicêmica.

Embora o *dataset* sintético tenha sido construído para atender às necessidades da segunda modelagem, era imprescindível que também incluísse as variáveis utilizadas na primeira modelagem, visto que esta segunda etapa depende diretamente dos resultados da primeira. Isso ocorre porque o modelo de risco é treinado exclusivamente com pacientes previamente classificados como pré-diabéticos ou não diabéticos, classificação essa oriunda da modelagem diagnóstica.

Os dados simulados foram disponibilizados em um arquivo JSON contendo as seguintes informações essenciais: identificador, idade, altura, peso, glicose, hemoglobina glicada e presença ou ausência de diabetes. A partir desse conjunto de atributos, foi possível estruturar um *dataset* coerente, robusto e adequado ao treinamento do segundo modelo. Uma análise inicial indicou que o arquivo continha um total de 7.585 pacientes sintéticos.

Antes de sua utilização, o *dataset* passou por um processo de análise e filtragem para adequação às necessidades do estudo. Inicialmente, avaliou-se quais dos 7.585 indivíduos apresentavam todas as variáveis de interesse (glicose, hemoglobina glicada, raça, sexo, peso, altura ou IMC, além da classificação diagnóstica).

Posteriormente, verificou-se a faixa etária dos pacientes. O simulador gera indivíduos ("sints") desde o primeiro ano de vida até sua morte, permitindo a restrição da análise apenas às idades relevantes para o modelo, definidas entre 20 e 90 anos. Após essa filtragem etária e das variáveis essenciais, o número de pacientes foi reduzido para 3.144.

Em seguida, procedeu-se à análise da variável raça, uma vez que o modelo considerava apenas duas categorias: negros e brancos. Aplicado esse filtro, o total foi reduzido para 2.949 pacientes sintéticos.

A partir desse conjunto final, foram estruturados dois *datasets*. O primeiro destinou-se à classificação diagnóstica (diabético, pré-diabético ou não diabético), utilizando os dados de glicose e hemoglobina glicada e o segundo *dataset*, por sua vez, foi formado a partir dos resultados do primeiro, considerando apenas indivíduos classificados como pré-

diabéticos ou não diabéticos. Dessa maneira, dos 2.949 indivíduos, permaneciam 1.352 para a construção do conjunto de dados voltado à modelagem de risco. Esse segundo *dataset* incluía, além da classificação herdada da primeira modelagem, informações de idade, raça, sexo e IMC.

Para cada um dos dois *datasets*, foram geradas duas tabelas: a primeira, denominada **data**, contendo todas as variáveis de entrada utilizadas pelo modelo; e a segunda, denominada ***target***, estruturada como uma matriz binária que associa cada paciente à resposta esperada na classificação. Ressalta-se que, nesta modelagem, a construção do arquivo ***target*** foi realizada a partir da utilização das regras mamdani, uma vez que os dados gerados pelo simulador não apresentavam classificação diagnóstica para os pacientes.

Dessa forma, fez-se necessário que para a construção dos arquivos ***target*** fossem definidas as regras mamdani para ambas as modelagens: a primeira delas considerando a classificação diagnóstica laboratorial em diabético, não diabético e pré diabético e, a segunda, a categorização transversal do risco metabólico.

Sendo assim, para a construção do primeiro arquivo, foi considerado que as regras são elaboradas a partir de dois critérios (G e HbA1c) e que cada critério pode apresentar três classificações (normal, intermediário e alto), conclui-se que são, portanto, 9 regras. São elas:

- Se G é ALTO e HbA1c é ALTO então Y é DIABÉTICO;
- Se G é ALTO e HbA1c é INTERMEDIÁRIO então Y é PRÉ-DIABÉTICO;
- Se G é ALTO e HbA1c é NORMAL então Y é PRÉ-DIABÉTICO;
- Se G é INTERMEDIÁRIO e HbA1c é ALTO então Y é DIABÉTICO;
- Se G é INTERMEDIÁRIO e HbA1c é INTERMEDIÁRIO então Y é PRÉ-DIABÉTICO;
- Se G é INTERMEDIÁRIO e HbA1c é NORMAL então Y é PRÉ-DIABÉTICO;
- Se G é NORMAL e HbA1c é ALTO então Y é DIABÉTICO;
- Se G é NORMAL e HbA1c é INTERMEDIÁRIO então Y é PRÉ-DIABÉTICO;
- Se G é NORMAL e HbA1c é NORMAL então Y é NÃO DIABÉTICO;

Já para a construção do arquivo para a segunda modelagem, foi considerado que as regras são elaboradas a partir de quatro critérios (IMC, C, S e R) e que três deles (C, S e R) apresentam duas classificações e um deles (IMC) apresenta 3 classificações, conclui-se que serão, portanto, 24 regras. São elas:

- Se C é PRÉ-DIABETES e S é FEMININO e R é BRANCO e IMC é ALTO então Y é ALTO;
- Se C é PRÉ-DIABETES e S é FEMININO e R é BRANCO e IMC é INTERMEDIÁRIO então Y é INTERMEDIÁRIO;
- Se C é PRÉ-DIABETES e S é FEMININO e R é BRANCO e IMC é NORMAL então Y é INTERMEDIÁRIO;
- Se C é PRÉ-DIABETES e S é FEMININO e R é NEGRO e IMC é ALTO então Y é ALTO;
- Se C é PRÉ-DIABETES e S é FEMININO e R é NEGRO e IMC é INTERMEDIÁRIO então Y é ALTO;
- Se C é PRÉ-DIABETES e S é FEMININO e R é NEGRO e IMC é NORMAL então Y é ALTO;
- Se C é PRÉ-DIABETES e S é MASCULINO e R é BRANCO e IMC é ALTO então Y é INTERMEDIÁRIO;
- Se C é PRÉ-DIABETES e S é MASCULINO e R é BRANCO e IMC é INTERMEDIÁRIO então Y é BAIXO;
- Se C é PRÉ-DIABETES e S é MASCULINO e R é BRANCO e IMC é NORMAL então Y é BAIXO;
- Se C é PRÉ-DIABETES e S é MASCULINO e R é NEGRO e IMC é ALTO então Y é ALTO;
- Se C é PRÉ-DIABETES e S é MASCULINO e R é NEGRO e IMC é INTERMEDIÁRIO então Y é INTERMEDIÁRIO;
- Se C é PRÉ-DIABETES e S é MASCULINO e R é NEGRO e IMC é NORMAL então Y é INTERMEDIÁRIO;
- Se C é NÃO DIABETES e S é FEMININO e R é BRANCO e IMC é ALTO então Y é BAIXO;
- Se C é NÃO DIABETES e S é FEMININO e R é BRANCO e IMC é INTERMEDIÁRIO então Y é BAIXO;
- Se C é NÃO DIABETES e S é FEMININO e R é BRANCO e IMC é NORMAL então Y é BAIXO;

- Se C é NÃO DIABETES e S é FEMININO e R é NEGRO e IMC é ALTO então Y é ALTO;
- Se C é NÃO DIABETES e S é FEMININO e R é NEGRO e IMC é INTERMEDIÁRIO então Y é INTERMEDIÁRIO;
- Se C é NÃO DIABETES e S é FEMININO e R é NEGRO e IMC é NORMAL então Y é INTERMEDIÁRIO;
- Se C é NÃO DIABETES e S é MASCULINO e R é BRANCO e IMC é ALTO então Y é BAIXO;
- Se C é NÃO DIABETES e S é MASCULINO e R é BRANCO e IMC é INTERMEDIÁRIO então Y é BAIXO;
- Se C é NÃO DIABETES e S é MASCULINO e R é BRANCO e IMC é NORMAL então Y é BAIXO/
- Se C é NÃO DIABETES e S é MASCULINO e R é NEGRO e IMC é ALTO então Y é INTERMEDIÁRIO;
- Se C é NÃO DIABETES e S é MASCULINO e R é NEGRO e IMC é INTERMEDIÁRIO então Y é BAIXO;
- Se C é NÃO DIABETES e S é MASCULINO e R é NEGRO e IMC é NORMAL então Y é BAIXO;

Ressalta-se que, assim como na modelagem anterior, as regras para este modelo são fixas e foram determinadas a partir das condições estabelecidas pela SBD e pela profissional da saúde que acompanha o estudo.

Como as tabelas relativas aos dados reais do HU já foram apresentadas para exemplificar as tabelas utilizadas na primeira modelagem, optou-se aqui por ilustrar apenas as tabelas referentes ao segundo modelo. A Figura 53 apresenta um exemplo do arquivo **data**, contendo todas as variáveis pertinentes à modelagem do risco, enquanto a Figura 54 apresenta um exemplo do arquivo ***target***, que indica, para cada paciente, sua classificação transversal de risco metabólico.

7.1.2.1 Caracterização da Amostra

A base de dados utilizada neste estudo foi composta por 1.352 indivíduos, contendo informações demográficas e antropométricas relevantes para a análise proposta. As variáveis consideradas foram idade, sexo, raça, IMC e classificação glicêmica.

A idade dos participantes variou entre 20 e 97 anos, com média de 39,44 $\pm$ 16,44 anos. O IMC apresentou valores entre 13,84 e 69,01 kg/m^2, com média de 24,05 ± 6,48 kg/^{2}.

	A	B	C	D	E	F
1	COD	IDADE	SEXO	RACA	IMC	CLASSIFICACAO
2	5047	20	1	0	21.21	Normal
3	5048	29	1	0	21.98	Normal
4	5052	88	0	0	23.1	Normal
5	5058	23	1	0	21.23	Normal
6	5065	46	0	0	29.24	Pré Diabetes
7	5067	47	1	0	25.86	Normal
8	5077	77	0	0	23.26	Normal
9	5081	47	0	0	21.56	Normal
10	5091	52	1	0	24.33	Pré Diabetes
11	5092	23	1	1	20.33	Normal
12	5093	63	0	0	30.74	Normal
13	5094	35	1	0	23.28	Normal
14	5096	56	1	0	26.53	Normal
15	5097	46	0	0	22.56	Normal
16	5100	74	0	1	48.13	Pré Diabetes

Figura 53: Exemplo de trecho do arquivo **data** resultante da filtragem dos dados do simulador.

Fonte: Autora.

Em relação ao sexo, a amostra apresentou distribuição equilibrada, sendo composta por 687 indivíduos do sexo feminino (50,8%) e 665 do sexo masculino (49,2%).

A variável raça foi categorizada em dois grupos: indivíduos brancos (n = 1.105; 81,7%) e indivíduos pretos (n = 247; 18,3%). A inclusão dessa variável justifica-se pela possibilidade de diferenças fisiológicas e hematológicas influenciarem os parâmetros analisados e, consequentemente, o desempenho dos modelos desenvolvidos.

Quanto à classificação glicêmica, 1.146 indivíduos (84,8%) foram classificados como normais, enquanto 206 indivíduos (15,2%) foram classificados como pré-diabéticos. Essa distribuição evidencia um desbalanceamento moderado entre as classes, característica comum em bases de dados clínicas utilizadas para desenvolvimento de modelos preditivos.

A Figura 55 apresenta a caracterização da amostra utilizada neste estudo, evidenciando a distribuição dos participantes quanto ao sexo, raça, faixa etária, IMC e classificação glicêmica.

Por outro lado, a Tabela 27 apresenta um resumo das principais características da amostra utilizada neste trabalho.

Além disso, optou-se por apresentar a caracterização da amostra estratificada segundo a classificação glicêmica. Observa-se que os indivíduos classificados como pré-diabéticos apresentaram idade média superior à dos indivíduos normais (44,56 ± 18,57 anos versus 38,52 ± 15,86 anos). O IMC médio também foi ligeiramente maior no grupo pré-diabético (24,48 ± 7,16 kg/m²) quando comparado ao grupo normal (23,97 ± 6,34 kg/m²).

Quanto à distribuição por sexo, o grupo normal apresentou predominância feminina (52,2%), enquanto o grupo pré-diabético apresentou maior proporção de indivíduos do sexo masculino (56,8%). Em relação à raça, observou-se predominância de indivíduos brancos em ambos os grupos, correspondendo a 81,1% dos participantes normoglicêmicos e 85,4% dos pré-diabéticos.

Esses resultados indicam que a amostra é heterogênea em termos demográficos e an-

	A	B	C	D
1	COD	BAIXO	MEDIO	ALTO
2	5047	1	0	0
3	5048	1	0	0
4	5052	1	0	0
5	5058	1	0	0
6	5065	0	1	0
7	5067	1	0	0
8	5077	1	0	0
9	5081	1	0	0
10	5091	1	0	0
11	5092	1	0	0
12	5093	1	0	0
13	5094	1	0	0
14	5096	1	0	0
15	5097	1	0	0
16	5100	0	0	1

Figura 54: Exemplo de trecho do arquivo ***target*** resultante da filtragem dos dados do simulador.

Fonte: Autora.

tropométricos, permitindo avaliar o comportamento dos modelos propostos em diferentes perfis de indivíduos. A Tabela 28 apresenta a caracterização da amostra estratificada pela classificação glicêmica.

7.2 Modelagem Neuro Fuzzy em Dois Níveis

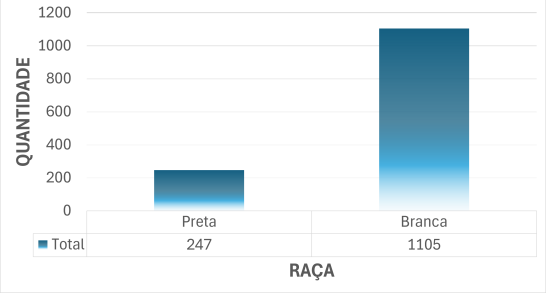
A modelagem proposta foi organizada em dois níveis complementares, ambos fundamentados em um sistema neuro-*fuzzy* do tipo Takagi–Sugeno–Kang (TSK). O primeiro nível realiza uma classificação laboratorial suave, a partir de glicemia de jejum e HbA1c, distinguindo os estados não diabético, pré-diabético e diabético conforme os intervalos

Tabela 27: Características da amostra utilizada no estudo - Dados sintéticos.

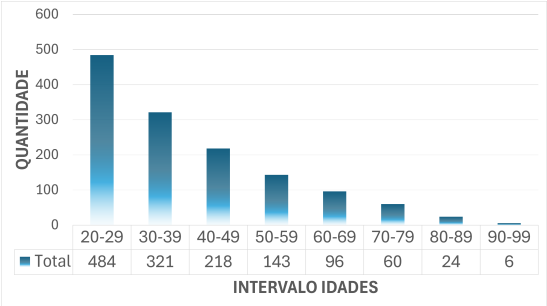
Característica	**Valor**
Número de indivíduos	1352
Idade (média $\pm$ DP)	$39,44 \pm 16,44$ anos
Idade mínima – máxima	20 – 97 anos
IMC (média $\pm$ DP)	$24,05 \pm 6,48$ kg/m^2
IMC mínimo – máximo	13,84 – 69,01 kg/m^2
Sexo feminino	687 (50,8%)
Sexo masculino	665 (49,2%)
Branco	1105 (81,7%)
Preto	247 (18,3%)
Normal	1146 (84,8%)
Pré-diabético	206 (15,2%)



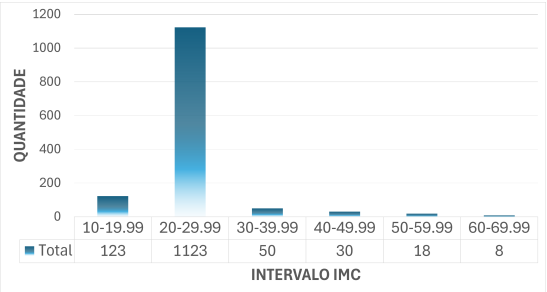
(a) Distribuição por sexo.



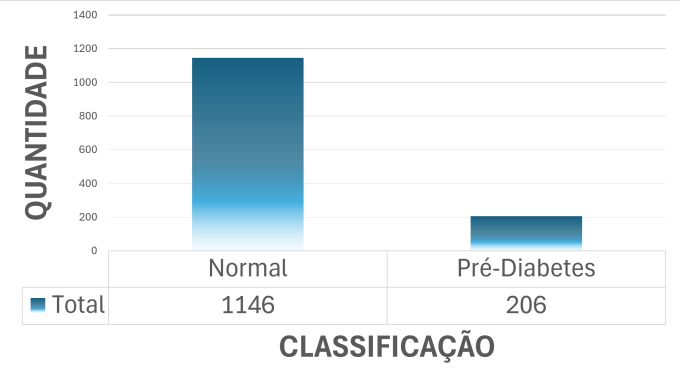
(b) Distribuição por raça.



(c) Distribuição por faixa etária.



(d) Distribuição por faixas de IMC.



(e) Distribuição da classificação glicêmica.

Figura 55: Caracterização da amostra utilizada no estudo. São apresentadas as distribuições referentes ao sexo, raça, faixa etária, índice de massa corporal (IMC) e classificação glicêmica dos 1.352 indivíduos que compõem a base de dados.

Tabela 28: Caracterização da amostra segundo a classificação glicêmica - Dados sintéticos.

Característica	Normal	Pré-Diabetes	Total
n	1146	206	1352
Idade (anos)	$38,52 \pm 15,86$	$44,56 \pm 18,57$	$39,44 \pm 16,44$
IMC (kg/m^2)	$23,97 \pm 6,34$	$24,48 \pm 7,16$	$24,05 \pm 6,48$
Sexo feminino	598 (52,2%)	89 (43,2%)	687 (50,8%)
Sexo masculino	548 (47,8%)	117 (56,8%)	665 (49,2%)
Branco	929 (81,1%)	176 (85,4%)	1105 (81,7%)
Preto	217 (18,9%)	30 (14,6%)	247 (18,3%)

Valores expressos como média $\pm$ desvio padrão ou número de indivíduos (%). IMC: Índice de Massa Corporal.

clínicos adotados. Esse nível não deve ser interpretado como substituto do diagnóstico médico, mas como uma formalização computacional interpretável dos critérios laboratoriais disponíveis.

O segundo nível realiza uma estratificação transversal do risco metabólico atual para indivíduos não diabéticos ou pré-diabéticos, considerando classificação glicêmica prévia, sexo, raça e IMC. Diferentemente de um modelo longitudinal de predição, esse nível não estima a probabilidade futura de desenvolver DM2 em determinado horizonte temporal; ele apenas organiza, de modo interpretável, fatores de risco disponíveis em um determinado momento.

A Figura 56 apresenta um diagrama que resume os dois níveis e as variáveis envolvidas em cada um deles. Posteriormente, cada etapa será descrita em detalhes.

7.2.1 Nível 01: Implementação *Fuzzy* dos Critérios Laboratoriais de Classificação Glicêmica

O primeiro nível tem como objetivo transformar os critérios laboratoriais adotados para avaliação glicêmica em uma classificação *fuzzy* interpretável. Assim, o paciente é classificado em uma das três categorias consideradas neste estudo: não diabético, pré-diabético ou diabético, a partir dos valores de glicemia de jejum e HbA1c.

É importante ressaltar que a expressão "classificação laboratorial", quando empregada neste capítulo, refere-se à classificação laboratorial usada como alvo computacional do modelo. Portanto, o Nível 01 não substitui o diagnóstico clínico de DM2, que depende de avaliação médica, contexto clínico, sintomas, repetição de exames quando necessário e exclusão de diagnósticos diferenciais. Conforme discutido na fundamentação clínica, diretrizes como as da SBD e da ADA recomendam confirmação por dois testes alterados, obtidos na mesma amostra com marcadores diferentes ou repetidos em segunda visita, exceto em casos de hiperglicemia inequívoca acompanhada de sintomas clássicos [14, 77, 15].

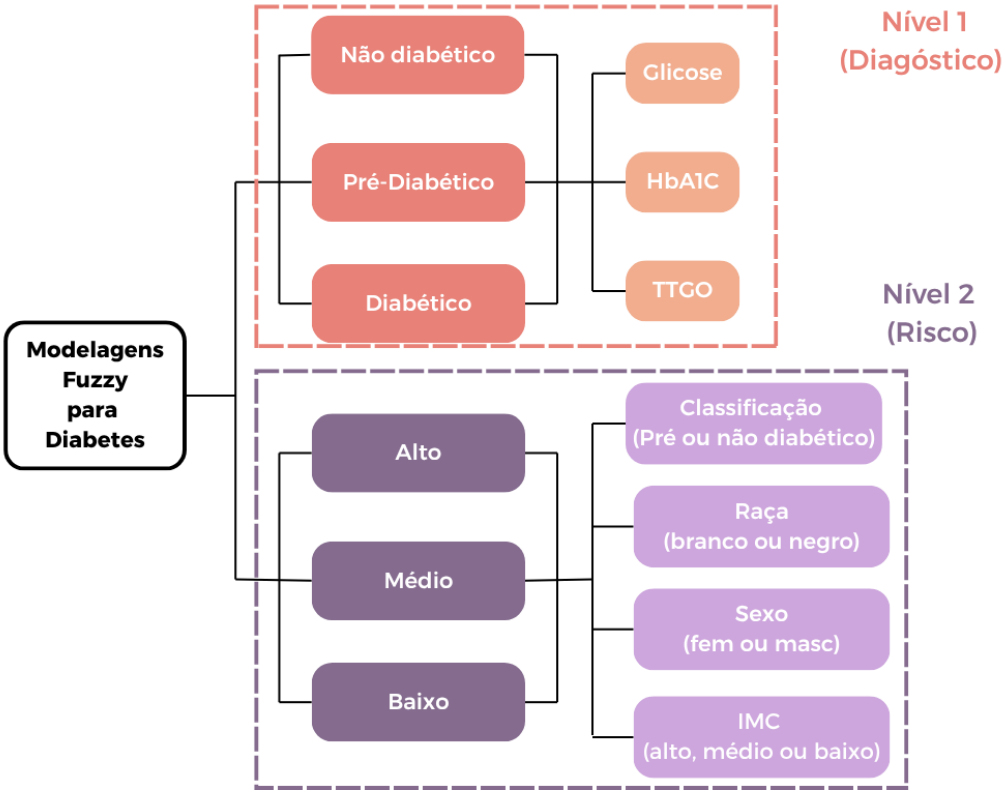


Figura 56: Diagrama que resume as duas modelagens *fuzzy* para DM2
Fonte: Autora.

No banco de dados utilizado neste estudo, cada paciente foi representado por uma única medida agregada de glicemia de jejum e HbA1c, obtida a partir da média dos exames disponíveis. Por esse motivo, o Nível 01 deve ser entendido como uma implementação *fuzzy* dos critérios laboratoriais disponíveis, e não como um protocolo clínico completo de diagnóstico. Essa distinção preserva o mérito da formulação proposta: transformar limiares rígidos em regiões graduais de pertinência, permitindo representar zonas de incerteza próximas aos pontos de corte clínicos.

A Tabela 29, retirada das diretrizes da SBD [77], apresenta um resumo dos critérios laboratoriais para diagnóstico de diabetes *mellitus* e pré-diabetes. Tais dados foram utilizados como base para a modelagem *fuzzy* desenvolvida para o primeiro modelo.

Tabela 29: Critérios laboratoriais para Diabetes *Mellitus* e pré-diabetes

Critérios	**Normal**	**Pré-diabetes**	**Diabetes *Mellitus***
Glicemia de jejum (mg/dl)	< 100	100-125	≥ 126
Glicemia ao acaso (mg/dl) + sintomas	-	-	≥ 200
Glicemia de 1 hora no TTGO (mg/dl)	< 155	155-208	≥ 209
Glicemia de 2 horas no TTGO (mg/dl)	< 140	140-199	≥ 200
HbA1c (%)	$< 5{,}7$	5,7-6,4	$\geq 6{,}5$

Fonte: [77]

Ressalta-se, porém, que todos os métodos possuem limitações metodológicas. Neste caso, as limitações para cada um dos critérios são:

- GJ: como o próprio nome sugere, necessita de jejum. Além disso, sofre interferências decorrentes de condições agudas e tem menor taxa de reprodutibilidade quando comparada à HbA1c;
- TTGO: é mais oneroso, desconfortável e consome mais tempo, mas tem maior sensibilidade. Para sua utilização, deve ser feita alimentação nos três dias antecedentes ao teste, com ingestão de, pelo menos, 150g de carboidratos por dia;
- HbA1c: tem maior custo e não leva em conta a variabilidade individual no fenômeno de glicação proteíca, além de ter menor sensibilidade diagnóstica do que os outros métodos.

Além disso, recomenda-se estabelecer o diagnóstico de diabetes *mellitus* se houver glicemia de jejum maior ou igual a 126 mg/dl e HbA1c maior ou igual a 6,5%, simultaneamente. Neste caso, não há necessidade de realização e avaliação do TTGO. Por outro lado, quando indica-se a realização do TTGO, recomenda-se o uso de TTGO-1h para diagnóstico de DM2 e detecção de pré-diabetes, por ser superior e mais prático do que o TTGO-2h.

Dessa forma, seguindo as diretrizes apresentadas pela SBD [77] e a opinião e auxílio de uma profissional da área, optou-se por classificar os pacientes em não diabético, pré-diabético e diabético a partir da avaliação da glicose e da hemoglobina glicada, ou seja, desconsiderando o TTGO.

A Tabela 30 apresenta o resumo dos dois critérios considerados para a modelagem 01 e seus respectivos valor para cada um dos níveis considerados.

Tabela 30: Critérios adotados para o paciente quanto à classificação laboratorial de Diabetes *Mellitus*, pré-diabetes e não diabético

Critérios	**Normal**	**Pré-diabetes**	**Diabetes *Mellitus***
Glicemia de jejum (mg/dl)	$g < 100$	$100 \leq g \leq 125$	$g > 125$
HbA1c (%)	HbA1c $< 5{,}7$	$5{,}7 \leq$ HbA1c $\leq 6{,}4$	HbA1c $> 6{,}4$

Fonte: [77]

Seguindo a arquitetura baseada no sistema *fuzzy* TSK apresentada e detalhada anteriormente, tem-se os seguintes passos considerados para realização da modelagem:

7.2.1.1 Fuzzificação

O primeiro passo consiste em transformar cada entrada (glicose e HbA1c) em graus de pertinência para cada um dos conjuntos *fuzzy*. As funções de pertinência

mais utilizadas na literatura são as gaussianas ou as trapezoidais. Para esta modelagem optou-se por considerar as funções trapezoidais para ambas entradas.

Variável linguística: Glicose

Conforme discutido anteriormente, a glicose pode ser classificada como alta, intermediária ou normal de acordo com os intervalos apresentados na Tabela 30. No entanto, existe também uma condição conhecida como Hipoglicemia, que corresponde a níveis anormalmente baixos de glicose no sangue (abaixo de 70 mg/dL). Embora a hipoglicemia não tenha sido considerada para fins de classificação, ela foi levada em conta para definir o limite inferior que a glicose pode assumir quando classificada como uma condição normal.

Dessa forma, foi introduzido um ajuste nos limites estabelecidos para a classificação de glicose normal. A Tabela 31 apresenta o intervalo atualizado considerado para a variável glicose.

Tabela 31: Critérios adotados para o paciente quanto à classificação laboratorial de Diabetes *Mellitus*, pré-diabetes e não diabético a partir da consideração do limite de hipoglicemia

Critérios	**Normal**	**Intermediário**	**Alto**
Glicemia de jejum (mg/dl)	$70 < g < 100$	$100 \leq g \leq 125$	$g > 125$

Fonte: [77]

Porém, além dos limites estabelecidos na tabela, optou-se por se considerar uma variação de 5%, calculada com base nos valores mínimo e máximo que a glicose pode assumir no seu intervalo central, que corresponde à faixa de pré-diabetes.

A adoção de uma banda de transição de 5% nas funções de pertinência da glicose e HbA1c e, posteriormente, do IMC, foi fundamentada na variabilidade inerente às medições laboratoriais e às flutuações biológicas observadas nesses parâmetros. Estudos longitudinais, como o Diabetes *Prevention Program* (DPP) [78] e o *Finissh* Diabetes *Prevention Study* (FDPS) [79], bem como as recomendações da IFCC (*International Federation of Clinical Chemistry and Laboratory Medicine*) [80], indicam que a conversão de pré-diabetes para DM2 pode ocorrer em torno de 5-10% ao ano na ausência de intervenção. Além disso, quando se trata de análise da glicose e da HbA1c, as recomendações da IFCC indicam que a glicose apresenta coeficiente de variação analítica em torno de 2% e variabilidade biológica intraindividual próxima de 6,5%, enquanto a HbA1c apresenta variabilidade analítica entre 3% e 5%.

Dessa forma, a utilização de uma banda de 5% em torno dos limiares diagnósticos busca representar explicitamente a região de incerteza existente na prática clínica. Essa abordagem é particularmente adequada à lógica *fuzzy*, pois permite modelar transições graduais entre as classes, evitando mudanças abruptas de classificação decorrentes de

pequenas oscilações analíticas ou biológicas. Por exemplo, um indivíduo com glicemia de jejum de 125 mg/dL pode apresentar valor superior a 126 mg/dL em uma medição subsequente sem que haja alteração clínica significativa. Enquanto abordagens baseadas em limiares rígidos classificariam esses casos em categorias distintas, a modelagem *fuzzy* reconhece essa zona de incerteza e fornece uma representação mais consistente com a realidade clínica e laboratorial.

Dessa forma, considerando que o valor mínimo que a glicose pode assumir neste intervalo é 100 mg/dl e que o valor máximo é 125 mg/dl, tem-se que a variação de 5% deste intervalo é igual a 1,25. Tal variação foi aplicada nos limites inferior e superior de cada uma das classificações, a fim de garantir uma maior precisão nas análises. Considerando-se a forma trapezoidal fez-se necessário que se realizasse o processo de fuzzificação para cada uma das três classificações consideradas para o critério de glicose.

Para definir funções de pertinência trapezoidais, é necessário especificar inicialmente quatro pontos para cada termo linguístico, a partir dos quais são derivadas as equações lineares que definem o subconjunto *fuzzy* correspondente. Para maior clareza, o procedimento adotado para cada um dos três termos linguísticos possíveis da glicose é apresentado separadamente.

- **Glicose - Termo linguístico: NORMAL**
  Para definir os limites, e seguindo a abordagem de variação de 1,25 mg/dL, foram adotados os valores 68,75 e 71,25 como limites inferiores, e 98,75 e 101,25 como limites superiores, definindo os vértices da forma trapezoidal da função de pertinência do subconjunto *fuzzy* que caracteriza o termo linguístico NORMAL da variável glicose. Dessa forma, os pontos para obtenção das equações das retas são:

  - $a_1 = 68{,}75$
  - $a_2 = 71{,}25$
  - $a_3 = 98{,}75$
  - $a_4 = 101{,}25$

  A partir destes pontos, pode-se calcular as duas equações de reta, sendo uma crescente (Equação 23) e outra decrescente (Equação 24).

$$Crescente \rightarrow \frac{x - a_1}{a_2 - a_1} = \frac{x - 68{,}75}{71{,}25 - 68{,}75} = \frac{x - 68{,}75}{2{,}5} \tag{23}$$

$$Decrescente \rightarrow \frac{a_4 - x}{a_4 - a_3} = \frac{101{,}25 - x}{101{,}25 - 98{,}75} = \frac{101{,}25 - x}{2{,}5} \tag{24}$$

  A Figura 57 ilustra a função de pertinência da Glicose NORMAL, definida, para todo $x \in [0, +\infty[$ por:

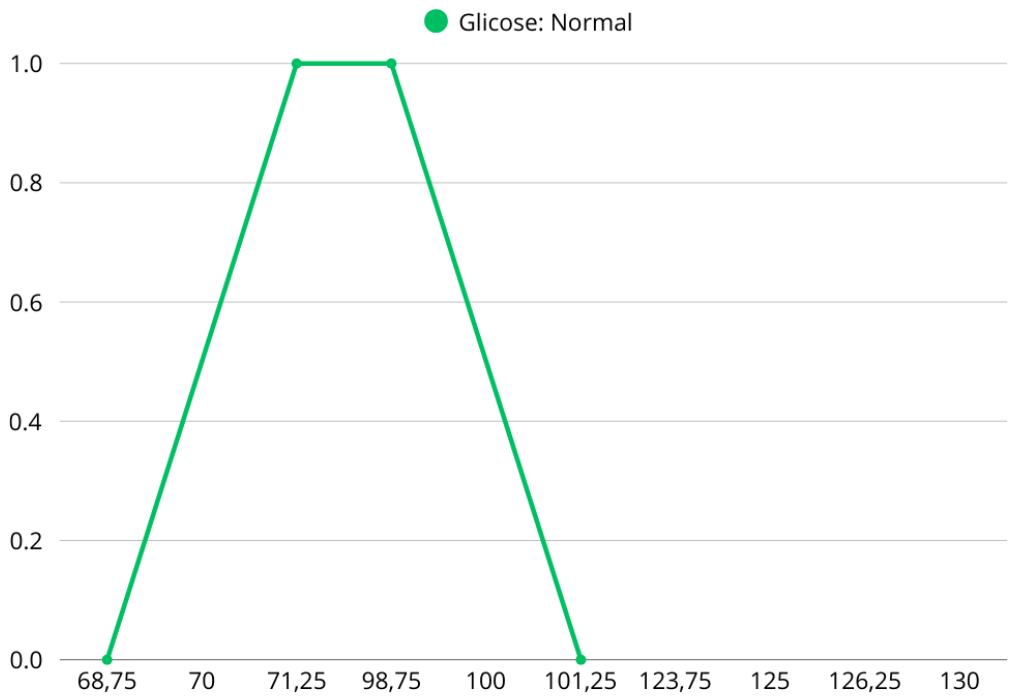


Figura 57: Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável glicose para a classificação normal

Fonte: Autora.

A partir das Equações 23 e 24 e do gráfico apresentado pela Figura 57, pode-se montar a função de pertinência para a classificação normal de glicose. A Equação 25 apresenta a função de pertinência, pode-se observar que a mesma conta com cinco pontos, são eles:

$$\mu_g^N(x) = \begin{cases} 0, & x < 68,75 \\ \frac{x-68,75}{2,5}, & 68,75 \leq x < 71,25 \\ 1, & 71,25 \leq x \leq 98,75 \\ \frac{101,25-x}{2,5}, & 98,75 < x \leq 101,25 \\ 0, & x > 101,25 \end{cases} \tag{25}$$

- **Glicose - Termo linguístico: INTERMEDIÁRIO**
  Seguindo a variação previamente adotada de 1,25 mg/dL, foram considerados os valores 98,75 e 101,25 para o limite inferior, e 123,75 e 126,25 para o limite superior, determinando os vértices da forma trapezoidal da função de pertinência do subconjunto *fuzzy* que caracteriza o termo linguístico INTERMEDIÁRIO da variável glicose. Dessa forma, pode-se concluir que os pontos para obtenção das equações das retas são:

- $a_1 = 98{,}75$
- $a_2 = 101{,}25$
- $a_3 = 123{,}75$
- $a_4 = 126{,}25$

A partir destes pontos, pode-se calcular as duas equações de reta, sendo uma crescente (Equação 26) e outra decrescente (Equação 27).

$$Crescente \rightarrow \frac{x - a_1}{a_2 - a_1} = \frac{x - 98{,}75}{101{,}25 - 98{,}75} = \frac{x - 98{,}75}{2{,}5} \tag{26}$$

$$Decrescente \rightarrow \frac{a_4 - x}{a_4 - a_3} = \frac{126{,}25 - x}{126{,}25 - 123{,}75} = \frac{126{,}25 - x}{2{,}5} \tag{27}$$

A Figura 58 ilustra a função de pertinência da Glicose INTERMEDIÁRIO, definida, para todo $x \in [0, +\infty[$ por:

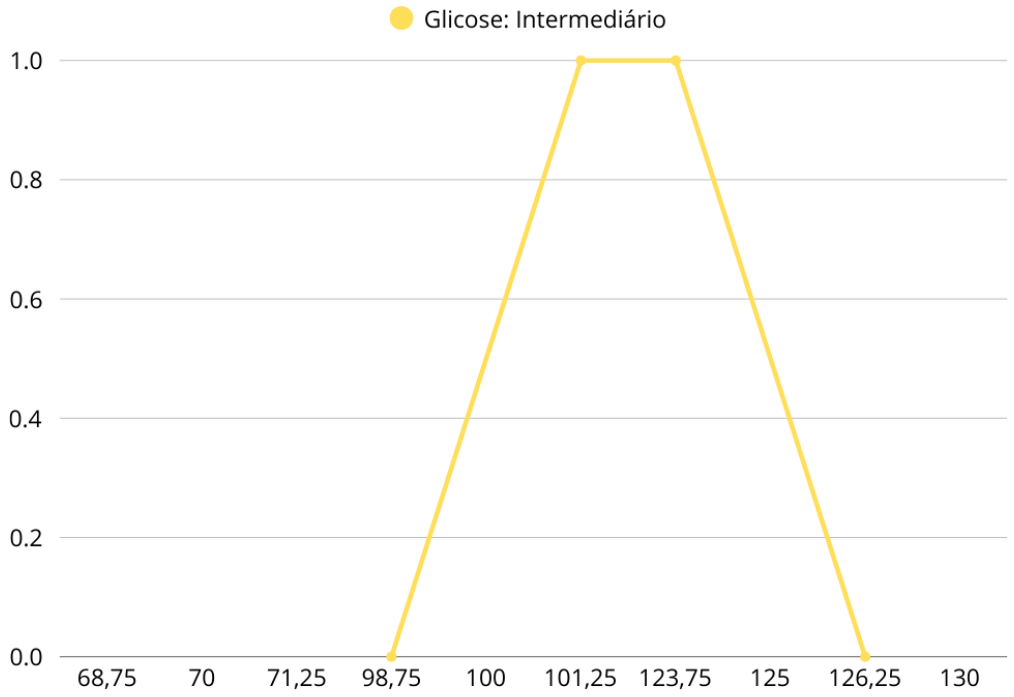


Figura 58: Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável glicose para a classificação intermediário

Fonte: Autora.

A partir das Equações 26 e 27 e do gráfico apresentado pela Figura 58, pode-se montar a função de pertinência para a classificação intermediário de glicose. A

Equação 28 apresenta a função de pertinência, pode-se observar que a mesma conta com cinco pontos, são eles:

$$\mu_g^I(x) = \begin{cases} 0, & x < 98,75 \\ \frac{x-98,75}{2,5}, & 98,75 \leq x < 101,25 \\ 1, & 101,25 \leq x \leq 123,75 \\ \frac{126,25-x}{2,5}, & 123,75 < x \leq 126,25 \\ 0, & x > 126,25 \end{cases} \tag{28}$$

- **Glicose - Termo linguístico: ALTO**
  Diferentemente dos dois casos anteriores de fuzzificação, para esta classificação é necessário definir apenas o limite inferior, ainda seguindo a abordagem de variação de 1,25 mg/dL. Assim, foram considerados os valores 123,75 e 126,25 para o limite inferior, definindo os vértices da forma trapezoidal da função de pertinência do subconjunto *fuzzy* que caracteriza o termo linguístico ALTO. Dessa forma, pode-se concluir que os pontos para obtenção da equação da reta são:

  - a_1 = 123,75
  - a_2 = 126,25

A partir destes pontos, pode-se calcular a equação de reta crescente (Equação 29), única necessária para este processo de fuzzificação.

$$Crescente \rightarrow \frac{x-a_1}{a_2-a_1} = \frac{x-123,75}{126,25-123,75} = \frac{x-123,75}{2,5} \tag{29}$$

A Figura 59 ilustra a função de pertinência da Glicose ALTO, definida, para todo $x \in [0, +\infty[$ por:

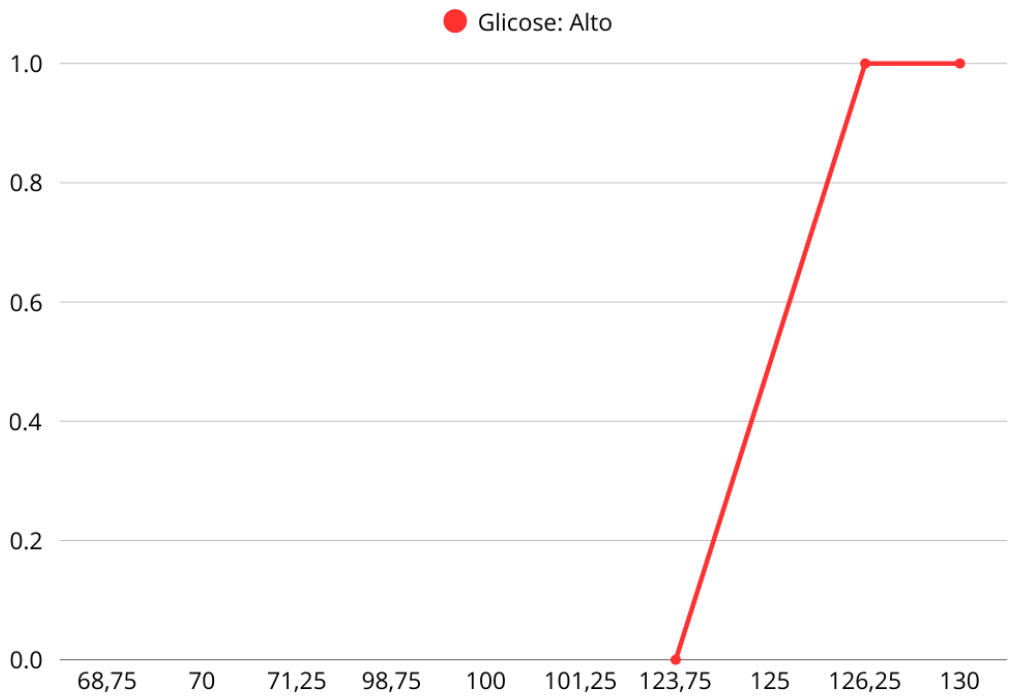


Figura 59: Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável glicose para a classificação alto

Fonte: Autora.

A partir da Equação 29 e do gráfico apresentado pela Figura 59, pode-se montar a função de pertinência para a classificação alto de glicose. A Equação 30 apresenta a função de pertinência, pode-se observar que a mesma conta com três pontos, são eles:

$$\mu_g^A(x) = \begin{cases} 0, & x < 123,75 \\ \frac{x-123,75}{2,5}, & 123,75 \leq x < 126,25 \\ 1, & x > 126,25 \end{cases} \tag{30}$$

A Figura 60 resume a qualificação da variável glicose por meio da combinação dos conjuntos fuzzy trapezoidais.

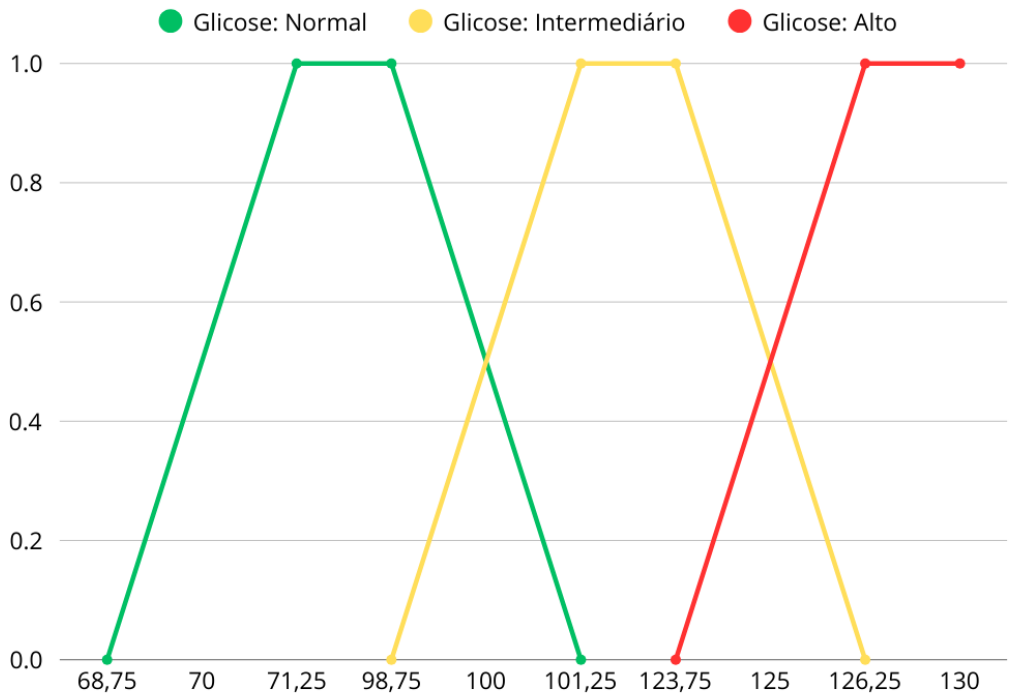


Figura 60: Conjuntos *fuzzy* trapezoidais representando os termos linguísticos da variável glicose a partir das três classificações

Fonte: Autora.

Variável linguística: Hemoglogina Glicada (HbA1c)

A HbA1c, de forma semelhante à glicose, também pode ser classificada como alta, intermediária ou normal, de acordo com os intervalos apresentados na Tabela 30.

Além dos limites estabelecidos na tabela, foi considerada uma variação de 5%, calculada a partir dos valores mínimo e máximo que a HbA1c pode assumir em seu intervalo central, correspondente à faixa de pré-diabetes. Considerando que o valor mínimo de HbA1c nesse intervalo é 5,7 e o valor máximo é 6,4, uma variação de 5% desse intervalo corresponde a 0,035.

Essa variação foi aplicada aos limites inferior e superior de cada classificação, a fim de garantir maior precisão analítica. Dada a forma trapezoidal das funções de pertinência, foi necessário realizar o processo de fuzzificação separadamente para cada uma das três classificações de HbA1c.

Seguiu-se o procedimento previamente descrito para a obtenção das funções de pertinência trapezoidais:

- **HbA1c - Termo linguístico: NORMAL**

  Para definir os limites, seguindo a abordagem de variação de 0,035, foram considerados os valores 5,665 e 5,735 para o limite superior (não sendo necessário definir um limite inferior). A função de pertinência do termo linguístico NORMAL da

HbA1c. Dessa forma, pode-se concluir que os pontos para obtenção das equações das retas são:

- $a_3 = 5{,}665$
- $a_4 = 5{,}735$

A partir destes pontos, pode-se calcular a equação de reta decrescente (Equação 31).

$$Decrescente \rightarrow \frac{a_4 - x}{a_4 - a_3} = \frac{5{,}735 - x}{5{,}735 - 5{,}665} = \frac{5{,}735 - x}{0{,}07} \tag{31}$$

A Figura 61 ilustra a função de pertinência da HbA1c ALTO, definida, para todo $x \in [0, +\infty[$ por:

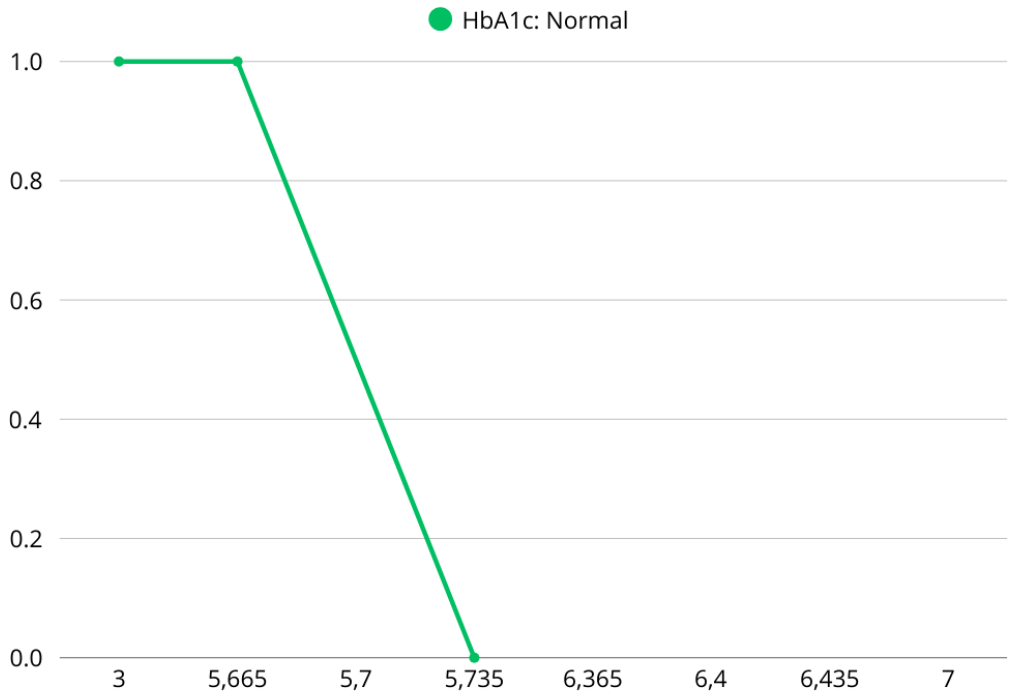


Figura 61: Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável HbA1c para a classificação normal

Fonte: Autora.

A partir da Equação 31 e do gráfico apresentado pela Figura 61, pode-se montar a função de pertinência para a classificação normal de HbA1c. A Equação 32 apresenta a função de pertinência, pode-se observar que a mesma conta com três pontos, são eles:

$$\mu_h^N(x) = \begin{cases} 1, & x < 5,665 \\ \frac{5,735-x}{0,07}, & 5,665 \leq x < 5,735 \\ 0, & x > 5,735 \end{cases} \tag{32}$$

- **HbA1c - Termo linguístico: INTERMEDIÁRIO**
  Para definir os limites, seguindo a abordagem de variação de 0,035, foram considerados os valores 5,665 e 5,735 para o limite inferior, e 6,365 e 6,435 para o limite superior, determinando os vértices da função de pertinência trapezoidal da HbA1c INTERMEDIÁRIA. Dessa forma, pode-se concluir que os pontos para obtenção das equações das retas são:

  - $a_1 = 5,665$
  - $a_2 = 5,735$
  - $a_3 = 6,365$
  - $a_4 = 6,435$

  A partir destes pontos, pode-se calcular as duas equações de reta, sendo uma crescente (Equação 33) e outra decrescente (Equação 34).

$$Crescente \rightarrow \frac{x-a_1}{a_2-a_1} = \frac{x-5,665}{5,735-5,665} = \frac{x-5,665}{0,07} \tag{33}$$

$$Decrescente \rightarrow \frac{a_4-x}{a_4-a_3} = \frac{6,435-x}{6,435-6,365} = \frac{6,435-x}{0,07} \tag{34}$$

  A Figura 62 ilustra a função de pertinência da HbA1c INTERMEDIÁRIO, definida, para todo $x \in [0, +\infty[$ por:

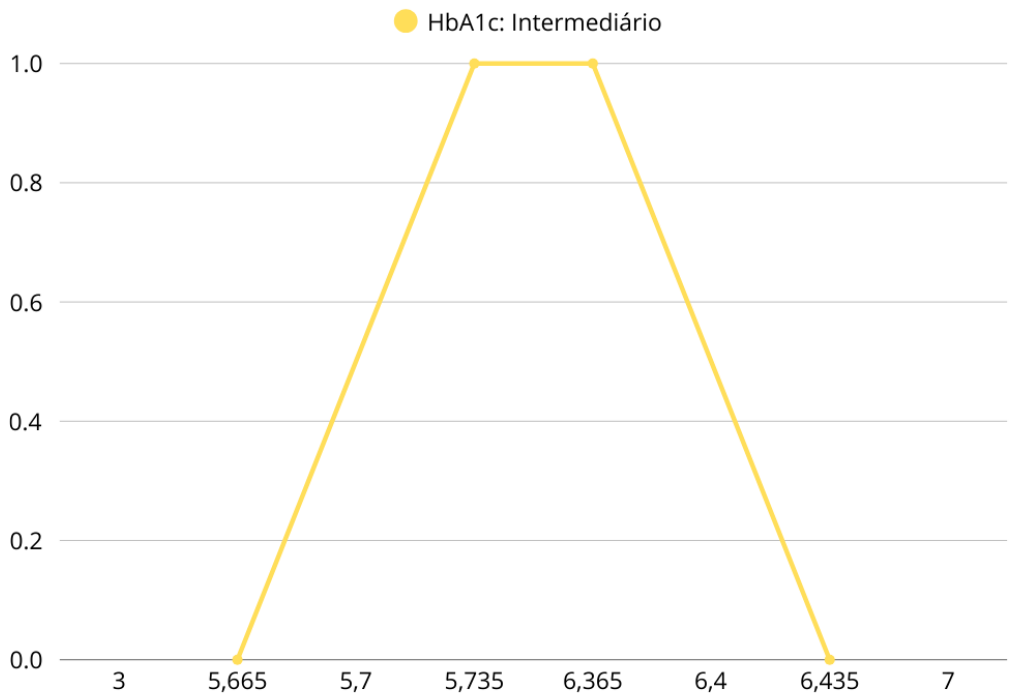


Figura 62: Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável HbA1c para a classificação intermediário

Fonte: Autora.

A partir das Equações 33 e 34 e do gráfico apresentado pela Figura 62, pode-se montar a função de pertinência para a classificação intermediário de HbA1c. A Equação 35 apresenta a função de pertinência, pode-se observar que a mesma conta com cinco pontos, são eles:

$$\mu_h^I(x) = \begin{cases} 0, & x < 5,665 \\ \frac{x-5,665}{0,07}, & 5,665 \leq x < 5,735 \\ 1, & 5,735 \leq x \leq 6,365 \\ \frac{6,435-6,365}{0,07}, & 6,365 < x \leq 6,435 \\ 0, & x > 6,435 \end{cases} \tag{35}$$

- **HbA1c - Termo linguístico: ALTO**
  Neste caso, apenas o limite inferior precisa ser definido. Seguindo a abordagem de variação de 0,035, foram considerados os valores 6,365 e 6,435 para o limite inferior, determinando os vértices da função de pertinência trapezoidal da HbA1c ALTA. Dessa forma, pode-se concluir que os pontos para obtenção da equação da reta são:

  – $a_1 = 6,365$

- a_2 = 6,435

A partir destes pontos, pode-se calcular a equação de reta crescente (Equação 36), única necessária para este processo de fuzzificação.

$$Crescente \rightarrow \frac{x-a_1}{a_2-a_1} = \frac{x-6,365}{6,435-6,365} = \frac{x-6,365}{0,07} \tag{36}$$

A Figura 63 ilustra a função de pertinência da HbA1c ALTO, definida, para todo $x \in [0, +\infty[$ por:

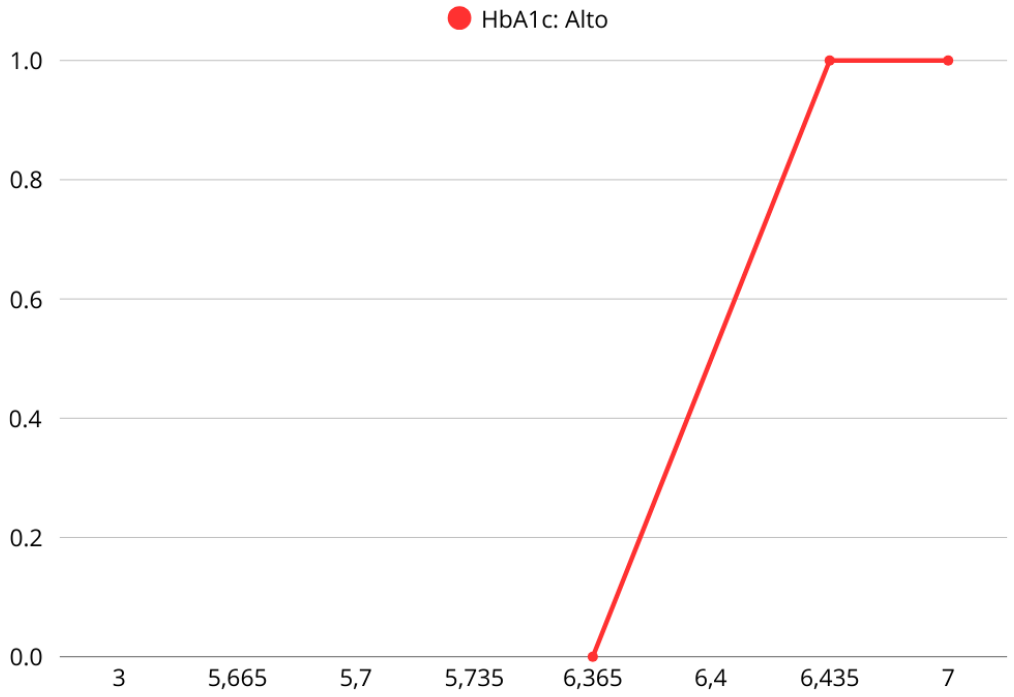


Figura 63: Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável HbA1c para a classificação alto

Fonte: Autora.

A partir da Equação 36 e do gráfico apresentado pela Figura 63, pode-se montar a função de pertinência para a classificação alto de HbA1c. A Equação 37 apresenta a função de pertinência, pode-se observar que a mesma conta com 3 pontos, são eles:

$$\mu_h^A(x) = \begin{cases} 0, & x < 6,365 \\ \frac{x-6,365}{0,07}, & 6,365 \leq x < 6,435 \\ 1, & x > 6,435 \end{cases} \tag{37}$$

Por fim, a representação fuzzy completa da variável HbA1c é ilustrada na Fig. 64.

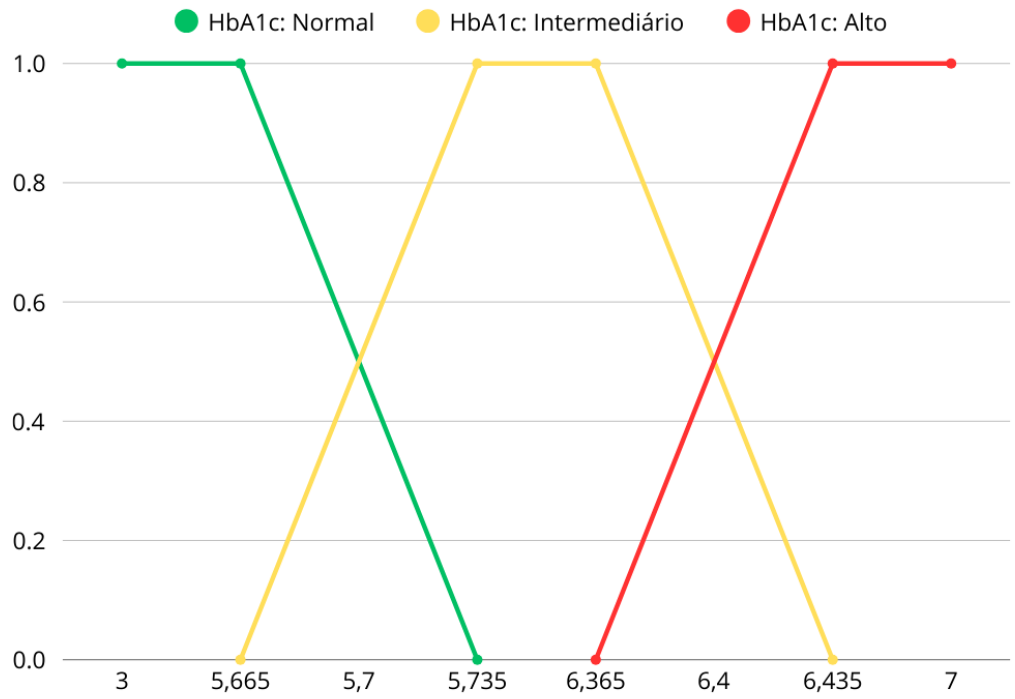


Figura 64: Conjuntos *fuzzy* representando os termos linguísticos da variável HbA1c a partir das três classificações

Fonte: Autora.

7.2.1.2 Ativação das Regras (peso) e Normalização

A força de disparo é calculada utilizando a Eq. (19) e a normalização é realizada pela Eq. (20). O Algoritmo 1 proposto calcula os pesos de ativação das regras *fuzzy* utilizando as funções de pertinência trapezoidais previamente definidas para dois atributos de entrada.

Inicialmente, os dados de entrada são separados em duas variáveis e, para cada regra, são definidos parâmetros trapezoidais específicos. Em seguida, os graus de pertinência de cada entrada são calculados por meio de funções lineares limitadas, garantindo que os valores permaneçam no intervalo [0,1].

Posteriormente, a força de ativação de cada regra é obtida combinando os graus de pertinência individuais por meio de uma t-norma do tipo produto. Esse processo é repetido para todas as regras, resultando em uma matriz de pesos de ativação que representa a contribuição de cada regra para o sistema de inferência *fuzzy*.

7.2.1.3 Agregação e Cálculo da Saída

A contribuição de cada regra para a saída final associada à uma determinada entrada é dada a partir do consequente, também conhecido como a parte do ”então’ de uma regra *fuzzy*.

Ressalta-se, mais uma vez, que neste caso as regras são fixas e foram definidas a partir das diretrizes da SBD [77] e da avaliação de especialistas da área.

Em resumo, a saída final é obtida pela soma das contribuições de cada regra, ponderadas de acordo com seus respectivos graus de ativação. O Algoritmo 1 apresenta o pseudocódigo que resume toda a modelagem fuzzy referente ao nível 01.

Algoritmo 1: Inferência TSK-ANFIS para Classificação Laboratorial

Input: Dados de entrada $x \in \mathbb{R}^{N\times 2}$, Parâmetros das consequentes $c_{i,j}$

Output: Saída final do sistema Y

// 1. Extrair atributos

$x_1 \leftarrow x(:,1),\ x_2 \leftarrow x(:,2);$

// 2. Camadas de Fuzzificação e Ativação

for *cada regra* $i = 1,\ldots,k$ **do**

  // Calcular grau de pertinência (G)

  $\mu_1 \leftarrow \max\left(\min\left(\frac{p_b - x_1}{p_b - p_a}, 1\right), 0\right);$

  // Calcular grau de pertinência (HbA1c)

  $\mu_2 \leftarrow \max\left(\min\left(\frac{p_b - x_2}{p_b - p_a}, 1\right), 0\right);$

  // Calcular a força de ativação (t-norma)

  $wt(:,i) \leftarrow \mu_1 \cdot \mu_2;$

// 3. Camada de Normalização

$S \leftarrow \sum_{j=1}^{k} wt(:,j);$

for *cada regra* $i = 1,\ldots,k$ **do**

  $\overline{wt}(:,i) \leftarrow \frac{wt(:,i)}{S};$

// 4. Camada de Consequente TSK e Agregação

for *cada regra* $i = 1,\ldots,k$ **do**

  $f_i \leftarrow p_{i,0} + p_{i,1}x_1 + p_{i,2}x_2;$

$Y \leftarrow \sum_{i=1}^{k} \overline{wt}(:,i) \cdot f_i;$

return Y

Neste algoritmo, x representa o vetor de entrada, k o número de regras *fuzzy*, μ_i os graus de pertinência, wt_i as forças de ativação das regras, $\overline{wt}_i$ as forças de ativação normalizadas, f_i as saídas das consequentes TSK, $p_{i,j}$ os parâmetros das consequentes e Y a saída final do sistema.

7.2.2 Nível 02: Categorização fuzzy do perfil de risco metabólico associado ao DM2

O segundo nível tem como objetivo classificar um paciente quanto ao risco metabólico atual associado ao DM2. Ou seja, pacientes que ainda não foram diagnosticados diabéticos podem, mesmo assim, apresentar um determinado nível de risco de desenvol-

ver a doença, com base em critérios específicos.

Destaca-se que a literatura clínica é consistente em estabelecer que essa transição é fenômeno longitudinal, com período natural de 5 a 10 anos entre o estado pré-diabético e o DM2 instalado, e taxa de conversão de 5-10 % ao ano sem intervenção. Porém, no caso do modelo implementado para o nível 2, a operação é transversal sobre um *snapshot*, ou seja, sem seguimento, sem tempo até o evento, sem censura. Sendo assim, o nível 2 consiste em uma categorização transversal do risco metabólico atual, com base em perfil instantâneo.

As diretrizes que orientam o diagnóstico do diabetes recomendam o rastreamento para todos os indivíduos com idade igual ou superior a 35 anos, bem como para adultos com sobrepeso ou obesidade que apresentem pelo menos um fator de risco adicional [77]. Além disso, as diretrizes indicam diversas condições nas quais o rastreamento para DM2 deve ser realizado, tais como:

- Idade acima de 35 anos (universal);
- Idade abaixo de 35 anos com sobrepeso ou obesidade, e mais um fator de risco:
  - História familiar de DM2 em parente de primeiro grau;
  - História de doença cardiovascular;
  - Hipertensão arterial;
  - HDL abaixo de 35 mg/dl;
  - Triglicerídeos acima de 250 mg/dl;
  - Síndrome de ovários policísticos;
  - Acantose nigricans;
  - Sedentarismo.
- Pré-diabetes em exame prévio;
- Diabetes gestacional prévio ou recém-nato grande para idade gestacional;
- FINDRISC (*Finnish Diabetes Risk Score*) alto ou muito alto.

Considerando os critérios listados anteriormente, bem como a experiência de um profissional de saúde, as seguintes variáveis foram selecionadas para avaliar o risco de desenvolvimento de Diabetes *Mellitus* Tipo 2:

- Classificação: pré-diabético ou não diabético;
- Sexo: feminino ou masculino;
- Raça: negro ou branco;

- IMC: alto, médio, baixo.

A partir da combinação destas entradas o paciente é classificado quanto ao risco metabólico atual associado ao DM2, sendo ele alto, intermediário ou baixo. Porém, diferente do que aconteceu na modelagem anterior, nesta modelagem somente o IMC apresenta valores específicos para cada uma das três classificações possíveis. As outras três variáveis são consideradas binárias, ou seja, ou apresentam maior ou menor risco dependendo do seu valor.

Especificamente em relação ao IMC, quando se avalia a classificação apresentada pelo estado da arte, encontra-se uma classificação dividida em 6 níveis, como mostra a Tabela 32.

Tabela 32: Classificação do IMC segundo estado da arte

Classificação	**IMC** (kg/m^2)
Abaixo do peso	$< 18,50$
Eutrófico	$18,50 - 24,99$
Sobrepeso	$25,00 - 29,99$
Obesidade grau I	$30,00 - 34,99$
Obesidade grau II	$35,00 - 39,99$
Obesidade grau III	$\geq 40,00$

Porém, quando se avalia tal classificação quanto ao risco de comorbidades, como por exemplo a DM2, considera-se que o risco passa a existir a partir da classificação intitulada sobrepeso, quaisquer valores inferiores à essa classificação são considerados normais e, por outro lado, quaisquer valores superiores à essa classificação são considerados de alto risco.

Dessa forma, a classificação do IMC foi resumida, como é apresentado pela Tabela 33, em três faixas: normal (que inclui abaixo do peso e eutrófico), intermediário (que inclui sobrepeso) e alto (que inclui obesidade grau I, II e III). Esta tabela apresenta o critério considerado para o IMC em relação à cada uma das classificações, pensando-se em uma função trapezoidal.

Tabela 33: Critérios adotados para o IMC quanto ao risco metabólico atual associado ao DM2

Critério	**Normal**	**Intermediário**	**Alto**
IMC (kg/m^2)	IMC < 25	$25 \leq$ IMC ≤ 30	IMC > 30

Fonte: [77]

Por outro lado, as três variáveis de entrada restantes são apresentadas na Tabela 34, onde são associadas ao nível 0 quando se assume uma menor probabilidade de desenvol-

vimento da Diabetes *Mellitus* Tipo 2, e ao nível 1 quando se assume uma maior probabilidade.

Tabela 34: Critérios adotados para as entradas binárias quanto ao risco metabólico atual associado ao DM2

Critério	**Risco Baixo (0)**	**Risco Alto (1)**
Classificação (C)	Não diabetes	Pré-diabetes
Sexo (S)	Masculino	Feminino
Raça (R)	Branco	Negro

Fonte: [77]

Assim como a modelagem anterior, esta também segue a arquitetura baseada no sistema TSK. Antes de apresentar os passos da modelagem, ressalta-se que quando um paciente apresenta risco em uma ou nenhuma entrada avaliada, então o mesmo apresenta perfil de baixo risco metabólico atual associado ao DM2; por outro lado se ele apresenta risco em duas entradas, se considera um perfil de risco metabólico intermediário associado ao DM2; por fim, se ele apresenta risco em três ou quatro entradas, então ele apresenta perfil de alto risco metabólico atual associado ao DM2. Dito isso, os seguintes passos foram seguidos para a realização da modelagem:

7.2.2.1 Fuzzificação

Nesta etapa, cada variável de entrada (IMC, sexo, raça e classificação diagnóstica) é modelada por funções de pertinência. Assim como na etapa de modelagem anterior, uma função de pertinência trapezoidal é adotada para o IMC, enquanto funções de pertinência binárias são consideradas para as demais entradas, uma vez que estas podem assumir apenas dois valores possíveis.

Variável linguística: IMC

Conforme discutido anteriormente, o IMC pode ser classificado como alto, intermediário ou normal de acordo com os intervalos apresentados na Tabela 33. Além dos limites estabelecidos na tabela, foi considerada uma variação de 5%, calculada a partir dos valores mínimo e máximo que o IMC pode assumir em seu intervalo central, ou seja, a faixa intermediária. Considerando que o valor mínimo de IMC nesse intervalo é 25 e o máximo é 30, uma variação de 5% desse intervalo corresponde a 0,25.

Essa variação foi aplicada aos limites inferior e superior de cada classificação, a fim de garantir maior precisão analítica. Dada a forma trapezoidal das funções de pertinência, o processo de fuzzificação foi realizado separadamente para cada uma das três classificações de IMC.

- **IMC - Termo linguístico: NORMAL**

  Para definir os limites, seguindo a abordagem de variação de 0,25, são considerados

os valores 24,75 e 25,25 para o limite superior, não sendo necessário definir um limite inferior neste caso. Dessa forma, os pontos para obtenção das equações das retas são:

- $a_3 = 24{,}75$
- $a_4 = 25{,}25$

A partir destes pontos, pode-se calcular a equação de reta decrescente (Equação 38).

$$Decrescente \rightarrow \frac{a_4 - x}{a_4 - a_3} = \frac{25{,}25 - x}{25{,}25 - 24{,}75} = \frac{25{,}25 - x}{0{,}5} \tag{38}$$

A Figura 65 ilustra a função de pertinência do IMC NORMAL, definida para todo $x \in [0, +\infty[$ por:

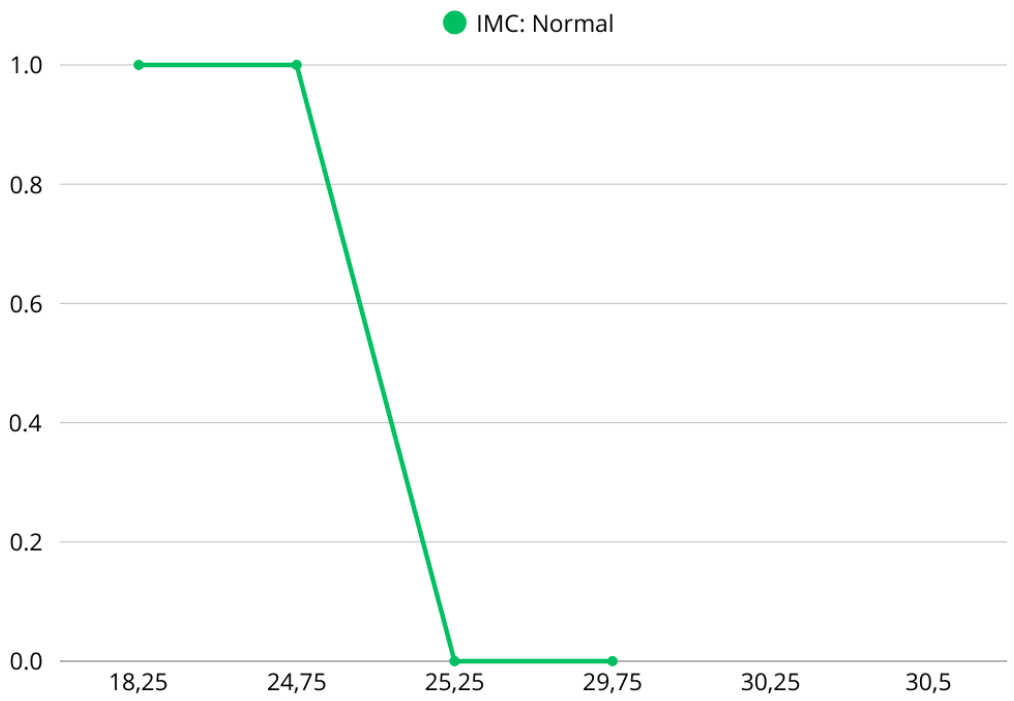


Figura 65: Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável IMC para a classificação normal

Fonte: Autora.

A partir da Equação 38 e do gráfico apresentado pela Figura 65, pode-se montar a função de pertinência para a classificação normal de IMC. A Equação 39 apresenta a função de pertinência, pode-se observar que a mesma conta com três pontos, são eles:

$$\mu^{N}_{IMC}(x) = \begin{cases} 1, & x < 24,75 \\ \frac{25,25-x}{0,5}, & 24,75 \leq x < 25,25 \\ 0, & x > 25,25 \end{cases} \tag{39}$$

- **IMC - Termo linguístico: INTERMEDIÁRIO**
  Para definir os limites, seguindo a abordagem de variação de 0,25, foram considerados os valores 24,75 e 25,25 para o limite inferior, e 29,75 e 30,25 para o limite superior. Dessa forma, pode-se concluir que os pontos para obtenção das equações das retas são:

  - a_1 = 24,75
  - a_2 = 25,25
  - a_3 = 29,75
  - a_4 = 30,25

  A partir destes pontos, pode-se calcular as duas equações de reta, sendo uma crescente (Equação 40) e outra decrescente (Equação 41).

$$Crescente \rightarrow \frac{x-a_1}{a_2-a_1} = \frac{x-24,75}{25,25-24,75} = \frac{x-24,75}{0,5} \tag{40}$$

$$Decrescente \rightarrow \frac{a_4-x}{a_4-a_3} = \frac{30,25-x}{30,25-29,75} = \frac{30,25-x}{0,5} \tag{41}$$

  A Figura 66 ilustra a função de pertinência do IMC INTERMEDIÁRIO, definida para todo $x \in [0, +\infty[$ por:

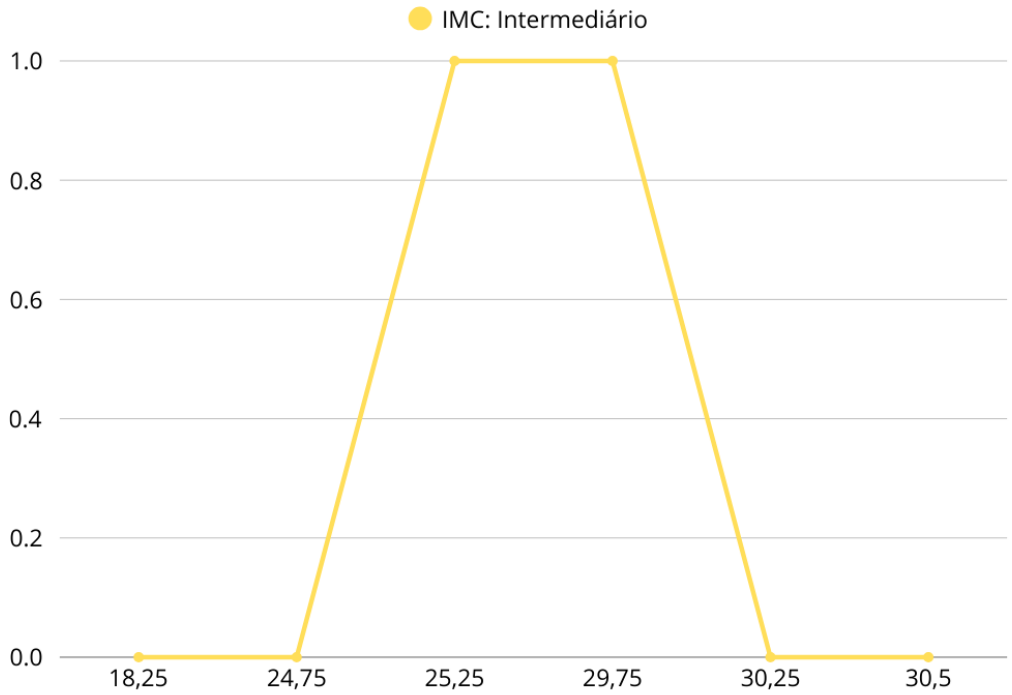


Figura 66: Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável IMC para a classificação intermediário

Fonte: Autora.

A partir das Equações 40 e 41 e do gráfico apresentado pela Figura 66, pode-se montar a função de pertinência para a classificação intermediário de IMC. A Equação 42 apresenta a função de pertinência, pode-se observar que a mesma conta com cinco pontos, são eles:

$$\mu^{I}_{IMC}(x) = \begin{cases} 0, & x < 24,75 \\ \frac{x-24,75}{0,5}, & 24,75 \leq x < 25,25 \\ 1, & 25,25 \leq x \leq 29,75 \\ \frac{30,25-x}{0,5}, & 29,75 < x \leq 30,25 \\ 0, & x > 30,25 \end{cases} \tag{42}$$

- **IMC - Termo linguístico: ALTO**
  Neste caso, apenas o limite inferior deve ser definido. Seguindo a abordagem de variação de 0,25, foram considerados os valores 29,75 e 30,25 para o limite inferior. Dessa forma, pode-se concluir que os pontos para obtenção da equação da reta são:

  – $a_1 = 29,75$

  – $a_2 = 30,25$

A partir destes pontos, pode-se calcular a equação de reta crescente (Equação 43), única necessária para este processo de fuzzificação.

$$Crescente \rightarrow \frac{x-a_1}{a_2-a_1} = \frac{x-29,75}{30,25-29,75} = \frac{x-29,75}{0,5} \tag{43}$$

A Figura 67 ilustra a função de pertinência do IMC ALTO, definida para todo $x \in [0,+\infty[$ por:

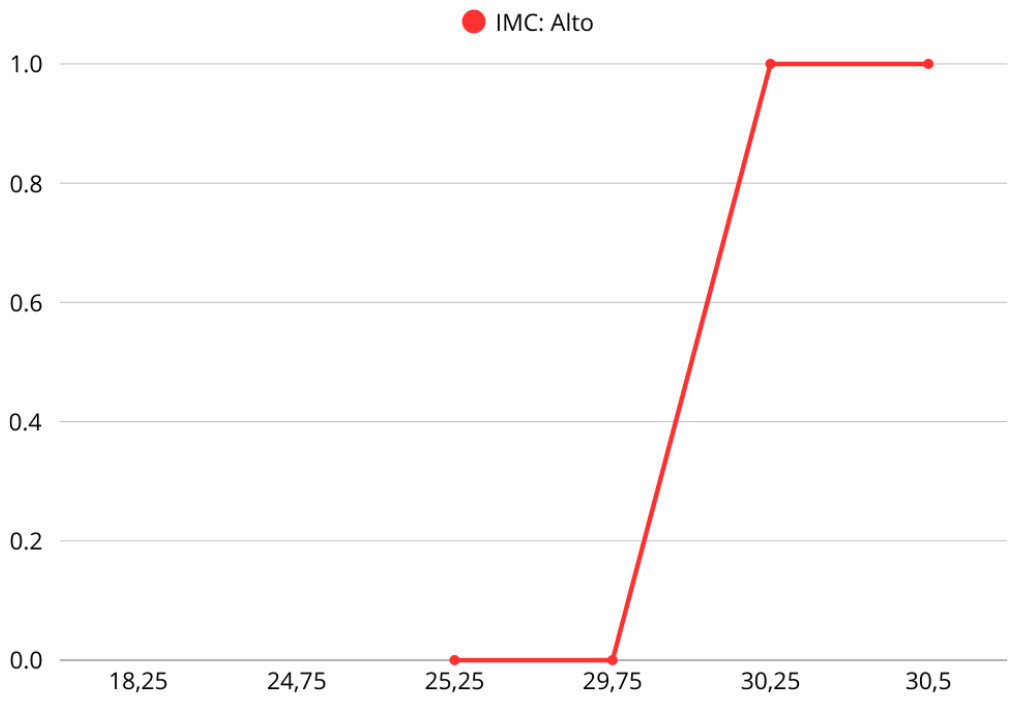


Figura 67: Conjunto *fuzzy* da função trapezoidal representando os termos linguístico da variável IMC para a classificação alto

Fonte: Autora.

A partir da Equação 43 e do gráfico apresentado pela Figura 67, pode-se montar a função de pertinência para a classificação alto de IMC. A Equação 44 apresenta a função de pertinência, pode-se observar que a mesma conta com 3 pontos, são eles:

$$\mu^N_{IMC}(x) = \begin{cases} 0, & x < 29,75 \\ \frac{x-29,75}{0,5}, & 29,75 \leq x < 30,25 \\ 1, & x > 30,25 \end{cases} \tag{44}$$

Por fim, a representação fuzzy completa da variável IMC pode ser obtida pela combinação das três funções de pertinência apresentadas anteriormente, conforme ilustrado na Fig. 68.

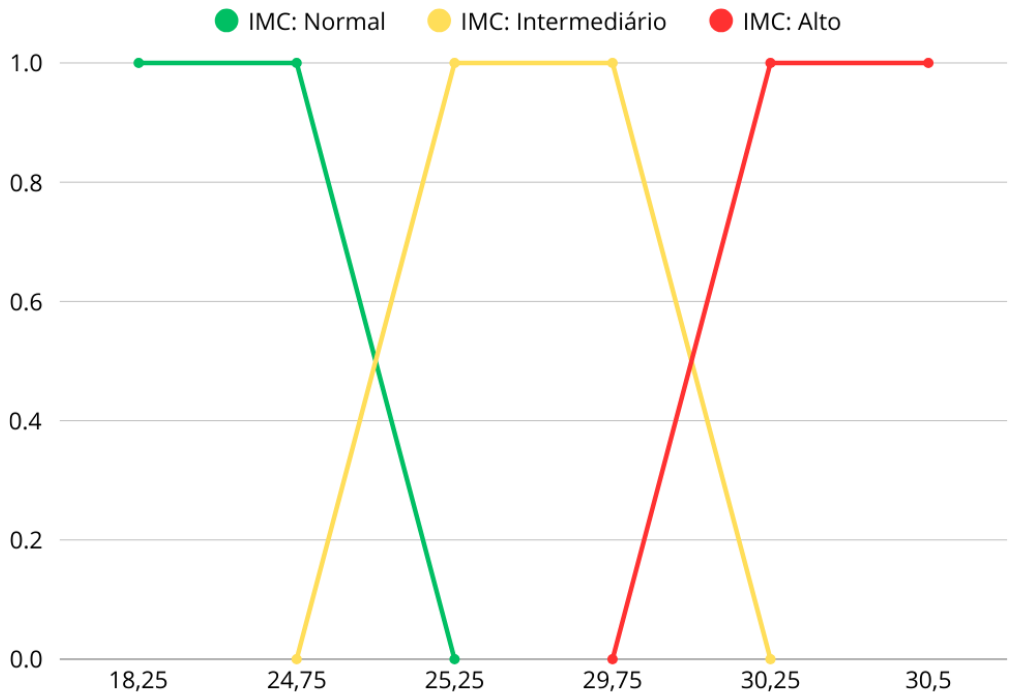


Figura 68: Conjuntos *fuzzy* representando os termos linguísticos da variável IMC a partir das três classificações

Fonte: Autora.

Variável Crisp: Classificação diagnóstica

Conforme discutido anteriormente, a classificação diagnóstica é representada por uma função binária que pode ser definida de acordo com seu nível lógico (baixo ou alto). Nesse caso, diferentemente da variável IMC, o processo de fuzzificação é consideravelmente mais simples e é realizado diretamente no código com base nos valores apresentados na Tabela 34.

Dessa forma, o nível lógico baixo indica uma condição não diabética e é representado pelo valor 0, enquanto o nível lógico alto indica uma condição pré-diabética e é representado pelo valor 1.

Variável Crisp: Sexo

A variável sexo, conforme discutido anteriormente, é representada por uma função binária que pode ser classificada de acordo com seu nível lógico (baixo ou alto). Nesse caso, de forma semelhante à classificação diagnóstica descrita anteriormente, o processo de fuzzificação é simples e é realizado diretamente no código com base nos valores apresentados na Tabela 34.

Assim, o nível lógico baixo indica o sexo masculino e é representado pelo valor 0, enquanto o nível lógico alto indica o sexo feminino e é representado pelo valor 1.

Variável Crisp: Raça

A variável raça, conforme discutido anteriormente, é representada por uma função binária que pode ser classificada de acordo com seu nível lógico (baixo ou alto). Nesse caso, de forma semelhante às variáveis classificação diagnóstica e sexo apresentadas anteriormente, o processo de fuzzificação é simples e é realizado diretamente no código com base nos valores apresentados na Tabela 34.

Assim, o nível lógico baixo indica a raça branca e é representado pelo valor 0, enquanto o nível lógico alto indica a raça negra e é representado pelo valor 1.

7.2.2.2 Ativação das Regras (peso) e Normalização

De acordo com o que foi apresentado anteriormente, o grau de ativação combina os graus de pertinência presentes no antecedente de uma regra fuzzy, resultando em um único valor de ativação. Existem diferentes formas de combinar esses graus de pertinência e, assim como na etapa de modelagem anterior, a t-norma do produto foi adotada neste modelo para realizar a ativação das regras nos antecedentes.

O Algoritmo 2 proposto calcula os pesos de ativação das regras fuzzy para a classificação de risco utilizando funções de pertinência binárias e trapezoidais. Inicialmente, os dados de entrada são separados em quatro atributos correspondentes à classificação diagnóstica, sexo, raça e índice de massa corporal (IMC). Para cada regra, funções de pertinência binárias são aplicadas às variáveis categóricas, enquanto uma função de pertinência trapezoidal é utilizada para modelar o IMC. Em seguida, os graus de pertinência individuais são combinados por meio de uma t-norma do produto para obter o grau de ativação de cada regra. Esse processo é repetido para todas as regras, resultando em uma matriz de pesos de ativação que representa a contribuição de cada regra no sistema de inferência fuzzy.

7.2.2.3 Agregação e Cálculo da Saída

A contribuição de cada regra para a saída final associada a uma determinada entrada, conforme discutido anteriormente, é determinada pelo seu consequente, também referido como a parte “então” de uma regra *fuzzy*.

Ressalta-se, mais uma vez, que, neste caso, as regras são fixas e foram definidas com base nas diretrizes da Sociedade Brasileira de Diabetes [77] e na avaliação de especialistas da área.

Em resumo, a saída final é obtida pela soma das contribuições de cada regra, ponderadas de acordo com seus respectivos graus de ativação. O Algoritmo 2 apresenta o pseudocódigo que resume toda a modelagem fuzzy referente ao nível 02.

Algoritmo 2: Inferência TSK-ANFIS para Classificação de Risco

Input: Dados de entrada $x \in \mathbb{R}^{N\times 4}$, Parâmetros das consequentes $c_{i,j}$

Output: Saída final do sistema Y

// 1. Extrair atributos

$x_1 \leftarrow x(:,1),\ x_2 \leftarrow x(:,2),\ x_3 \leftarrow x(:,3),\ x_4 \leftarrow x(:,4)$;

// 2. Camadas de Fuzzificação e Ativação

for *cada regra* $i = 1,\ldots,k$ **do**

  // Calcular pertinências binárias

  $\mu_1 \leftarrow \mathbb{I}(x_1 = 1),\ \mu_2 \leftarrow \mathbb{I}(x_2 = 1),\ \mu_3 \leftarrow \mathbb{I}(x_3 = 0)$;

  // Calcular grau de pertinência (IMC)

  $\mu_4 \leftarrow \max\left(\min\left(\frac{p_b - x_4}{p_b - p_a}, 1\right), 0\right)$;

  // Calcular a força de ativação (t-norma)

  $wt(:,i) \leftarrow \mu_1 \cdot \mu_2 \cdot \mu_3 \cdot \mu_4$;

// 3. Camada de Normalização

$S \leftarrow \sum_{j=1}^{k} wt(:,j)$;

for *cada regra* $i = 1,\ldots,k$ **do**

  $\overline{wt}(:,i) \leftarrow \frac{wt(:,i)}{S}$;

// 4. Camada de Consequente TSK e Agregação

for *cada regra* $i = 1,\ldots,k$ **do**

  $f_i \leftarrow p_{i,0} + p_{i,1}x_1 + p_{i,2}x_2 + p_{i,3}x_3 + p_{i,4}x_4$;

$Y \leftarrow \sum_{i=1}^{k} \overline{wt}(:,i) \cdot f_i$;

return Y

Neste algoritmo, x representa o vetor de entrada, k o número de regras *fuzzy*, μ_i os graus de pertinência, wt_i as forças de ativação das regras, $\overline{wt}_i$ as forças de ativação normalizadas, f_i as saídas das consequentes TSK, $p_{i,j}$ os parâmetros das consequentes e Y a saída final do sistema.

7.3 Análise dos Resultados da Metodologia Neuro-*Fuzzy* em Dois Níveis

Esta seção tem como objetivo apresentar os resultados obtidos a partir das metodologias descritas anteriormente na Seção 7.2.1 (Nível 01) e na Seção 7.2.2 (Nível 02), nas quais foram detalhadas as abordagens de modelagem para a classificação de pacientes quanto ao diagnóstico e ao risco metabólico de desenvolvimento de Diabetes Mellitus Tipo 2.

Conforme discutido anteriormente, os resultados foram obtidos utilizando dois con-

juntos de dados distintos: um conjunto gerado a partir de dados coletados no HU e outro obtido por meio do uso do simulador populacional **Sint Multiverse**. O Nível 01 utilizou o conjunto de dados do HU, enquanto o Nível 02 empregou o conjunto de dados sintético. Em ambos os casos, os conjuntos de dados foram divididos em dois subconjuntos: um para treinamento e outro para teste. A Figura 69 apresenta um resumo das informações utilizadas para cada um dos níveis.



Figura 69: Resumo dos dados utilizados para cada um dos níveis
Fonte: Autora.

De forma geral, para ambos os níveis, a pergunta central consiste em "Como transformar dados clínicos em uma decisão assistida, interpretável e validável?". Como resultado, espera-se que os modelos sejam capazes de reduzir a ambiguidade entre diagnóstico, risco e validação computacional. A Figura 70 apresenta as etapas desde a chegada dos dados clínicos até a geração de uma avaliação a partir destes dados.



Figura 70: Fluxograma resumo sobre a análise dos resultados
Fonte: Autora.

7.3.1 Protocolo de avaliação experimental

Embora o problema proposto seja intrinsecamente multiclasse, no qual cada paciente pertence exatamente a uma única categoria diagnóstica (diabético, pré-diabético ou não

diabético), a avaliação experimental foi conduzida sob um formalismo de aprendizado multirrótulo.

Especificamente, o problema multiclasse foi reformulado como um problema multirrótulo degenerado por meio de um esquema de codificação one-hot, no qual cada amostra é associada a um vetor binário de rótulos $y \in 0,1^C$ que satisfaz: $\sum_{c=1}^{C} y_c = 1$, onde C representa o número de classes diagnósticas.

Essa representação permite o uso direto de métricas padrão de avaliação multirrótulo, ao mesmo tempo em que preserva a semântica original do problema multiclasse. Nesse cenário restrito, diversas métricas multirrótulo se reduzem às suas correspondentes no contexto multiclasse. Em particular, a *Example-Based Accuracy* torna-se equivalente à acurácia convencional de classificação, enquanto a *Hamming Loss* mede a proporção de amostras classificadas incorretamente.

Além disso, métricas baseadas em ranqueamento, como Coverage, Ranking Loss e OneError, permanecem significativas, uma vez que o sistema *fuzzy* TSK proposto produz saídas contínuas para cada classe, as quais podem ser naturalmente interpretadas como graus de confiança associados a cada categoria diagnóstica.

A avaliação experimental da arquitetura proposta foi conduzida de forma independente para cada um dos dois níveis do sistema ANFIS, respeitando as particularidades dos dados utilizados e os objetivos específicos de cada etapa do processo decisório. Enquanto o Nível 01 foi desenvolvido para realizar a classificação diagnóstica de pacientes com base em critérios clínicos consolidados, o Nível 02 teve como objetivo realizar uma avaliação complementar de risco utilizando variáveis adicionais associadas ao perfil do indivíduo.

7.3.1.1 Protocolo Experimental do Nível 01

O Nível 01 foi concebido para realizar a classificação diagnóstica de indivíduos em três categorias: não diabético, pré-diabético e diabético. Para isso, foram utilizados dados reais provenientes do Hospital Universitário (HU), considerando como variáveis de entrada os valores de glicose plasmática e hemoglobina glicada (HbA1c).

A avaliação do modelo foi realizada por meio de uma estratégia composta por validação cruzada e teste externo independente. Inicialmente, os dados destinados ao treinamento foram submetidos a uma validação cruzada do tipo 5-fold, permitindo estimar a capacidade de generalização do modelo e reduzir possíveis efeitos decorrentes da divisão específica dos dados. Nesse procedimento, o conjunto de treinamento foi particionado em cinco subconjuntos, sendo quatro utilizados para treinamento e um para validação, repetindo-se o processo até que todos os subconjuntos fossem utilizados como conjunto de validação. A Tabela 35 apresenta a divisão dos dados de treino e teste externo para cada uma das classes.

Além da validação cruzada, foi reservado um conjunto de teste externo independente,

Classe	Total	Treino	Teste Externo
Não-Diabético	834	665	169
Pré-Diabético	1664	1417	247
Diabético	1079	918	161
Total	**3577**	**3000**	**577**

Tabela 35: Divisão dos dados do Nível 01 para treino e teste para cada uma das classes

completamente separado das etapas de treinamento e ajuste do modelo. Essa estratégia permitiu avaliar o desempenho do sistema em dados não observados durante o processo de aprendizagem, fornecendo uma estimativa mais realista da capacidade de generalização da abordagem proposta.

7.3.1.2 Protocolo Experimental do Nível 02

O Nível 02 foi desenvolvido para realizar uma avaliação de risco complementar, classificando os indivíduos em três categorias de risco: baixo, médio e alto. Diferentemente do Nível 01, esta etapa utilizou uma base de dados sintética construída a partir das informações obtidas no primeiro nível e de variáveis adicionais relacionadas ao perfil dos indivíduos.

As entradas do modelo foram compostas pela classificação diagnóstica previamente obtida no Nível 01, sexo, raça e índice de massa corporal (IMC). O objetivo desta etapa foi investigar o potencial da arquitetura neuro-*fuzzy* em produzir uma avaliação de risco interpretável, fornecendo uma camada adicional de suporte à decisão clínica.

Considerando a presença de desbalanceamento entre as classes de risco, como poder ser observado pela Tabela 36, especialmente na classe de maior risco, a avaliação experimental foi conduzida sob duas perspectivas distintas. Inicialmente, foi realizado um teste utilizando a distribuição original dos dados sintéticos, preservando as proporções naturais entre as classes. Em seguida, foi conduzido um segundo experimento utilizando um conjunto de teste balanceado, construído especificamente para ampliar a representatividade das classes minoritárias.

Classe	Total	%
Risco Baixo	1082	80,03%
Risco Médio	227	16,79%
Risco Alto	43	3,18%
Total	**1352**	**100%**

Tabela 36: Distribuição dos dados em cada uma das classes do Nível 02

Essa estratégia permitiu avaliar não apenas o desempenho global do modelo, mas também sua capacidade de identificar corretamente indivíduos pertencentes às classes menos frequentes, aspecto particularmente importante em aplicações relacionadas à saúde.

7.3.1.3 Métricas de Avaliação e Análise dos Resultados

A avaliação dos modelos desenvolvidos foi realizada por meio de métricas globais e métricas específicas por classe, permitindo analisar tanto o desempenho geral dos classificadores quanto sua capacidade de identificar corretamente cada categoria diagnóstica ou de risco.

As métricas globais adotadas foram *Subset Accuracy*, *Micro-F1* e *Average Precision*. Essas métricas fornecem uma visão abrangente do desempenho do sistema, considerando simultaneamente a capacidade de classificação correta e a consistência das predições realizadas.

Adicionalmente, foram utilizadas métricas específicas por classe, incluindo Precisão *Recall* e *F1-Score*. Essas métricas permitem avaliar individualmente o comportamento do modelo em cada categoria, aspecto particularmente importante em problemas de classificação multiclasses com diferentes distribuições amostrais.

Para complementar a análise, foram empregadas matrizes de confusão, possibilitando identificar os tipos de erros cometidos pelos modelos e analisar suas potenciais implicações clínicas.

No Nível 01, destinado à classificação diagnóstica utilizando dados reais do HH, o modelo apresentou elevado desempenho global, como mostra a Tabela 37. Foram obtidos valores de *Hamming Loss* igual a 0,0179, *Example-Based Accuracy* de 0,9575, *Subset Accuracy* de 0,9515, *Micro-F1* de 0,9729 e *Average Precision* de 0,9887. Esses resultados evidenciam a capacidade do sistema em reproduzir de forma consistente os critérios diagnósticos utilizados para a classificação de indivíduos em não diabéticos, pré-diabéticos e diabéticos.

Métrica	**Valor**
Hamming Loss	0,0179
Example-Based Accuracy	0,9575
Subset Accuracy	0,9515
Micro-F1	0,9729
Average Precision	**0,9887**

Tabela 37: Métricas Globais para análise do Nível 01

A análise por classe, apresentada pela Tabela 38, mostrou desempenho equilibrado entre as três categorias diagnósticas. Os valores de F1-Score foram de 0,9911 para a classe Não Diabético, 0,9857 para a classe Pré-Diabético e 0,9333 para a classe Diabético. Embora a classe Diabético tenha apresentado desempenho ligeiramente inferior, os resultados obtidos indicam elevada capacidade discriminativa do modelo em todas as categorias avaliadas.

No Nível 02, destinado à avaliação de risco, foram conduzidos experimentos utilizando tanto a distribuição original dos dados quanto uma configuração de teste balanceada

Classe	Precisão	*Recall*	F1
Não-Diabéticos	0,9940	0,9882	0,9911
Pré-Diabéticos	0,9918	0,9798	0,9857
Diabéticos	**0,9545**	**0,9130**	**0,9333**

Tabela 38: Métricas por Classe para análise do Nível 01

para aumentar a representatividade das classes minoritárias. No teste original, o modelo obteve *Micro-F1* de 0,9972, *Average Precision* de 0,9966 e *Subset Accuracy* de 0,9972. Entretanto, devido ao forte desbalanceamento entre as classes de risco, tais resultados poderiam superestimar o desempenho do sistema.

Por essa razão, foi realizada uma avaliação complementar utilizando um conjunto de teste balanceado. Nessa condição, os valores observados foram *Micro-F1* de 0,8247, *Average Precision* de 0,9222 e *Subset Accuracy* de 0,7933. Embora inferiores aos resultados obtidos no teste original, esses valores fornecem uma estimativa mais realista da capacidade de generalização do modelo em cenários nos quais todas as classes apresentam representatividade semelhante. A Tabela 39 apresenta os dados apresentados para os dois testes realizados.

Métrica	Teste Original	Teste Balanceado
Micro-F1	0,9972	0,8247
Average Precision	0,9986	0,9222
Subset Accuracy	0,9972	0,7333
Classe risco alto no teste	**3 casos**	**15 casos**

Tabela 39: Comparação dos resultados obtidos no Nível 02 para o teste original e teste balanceado

A análise por classe no teste balanceado demonstrou, como mostra a Tabela 40, que o modelo apresentou *Recall* igual a 1,0000 para as classes de risco baixo e risco médio, indicando identificação completa dessas categorias. Para a classe de risco alto, considerada a mais crítica e menos representada na base de dados, foi obtido *Recall* de 0,6667 e F1-Score de 0,8000. Esses resultados evidenciam que o modelo mantém desempenho satisfatório mesmo em cenários mais desafiadores, embora a identificação de indivíduos pertencentes à classe de maior risco ainda represente um aspecto passível de aprimoramento em estudos futuros.

Classe	Precisão	*Recall*	F1
Risco Baixo	0,7500	1,0000	0,8571
Risco Médio	0,6818	1,0000	0,8108
Risco Alto	**1,0000**	**0,6667**	**0,8000**

Tabela 40: Resultados obtidos para a análise por classe realizada para o teste balanceado

Para o teste balanceado também foi construída a matriz de confusão para avaliar onde

o modelo está acertando e onde está errando. Como pode ser observado na Tabela 41, o modelo acerta nas classes de risco baixo e médio, enquanto apresentou erros na classe de risco alto, justamente a classe minoritária.

Real/Pred	**Baixo**	**Médio**	**Alto**
Baixo	15	0	0
Médio	0	15	0
Alto	0	7	8

Tabela 41: Matriz de Confusão para análise de desempenho do teste balanceado

8 CONCLUSÕES E TRABALHOS FUTUROS

O crescente impacto das doenças crônicas, em especial o DM2, na saúde pública, aliado às limitações dos métodos tradicionais de diagnóstico e avaliação de risco, frequentemente baseados em limiares fixos e incapazes de capturar adequadamente as incertezas inerentes aos dados clínicos, evidencia a necessidade de abordagens mais robustas e inteligentes. Soma-se a isso a crescente disponibilidade de informações provenientes de exames laboratoriais e dispositivos de monitoramento pessoal, o que demanda soluções capazes de integrar múltiplas fontes de dados de forma interpretável e confiável. Nesse contexto, os sistemas neuro-*fuzzy* destacam-se como uma alternativa promissora, ao combinarem a capacidade de aprendizado dos métodos baseados em dados com a transparência dos sistemas baseados em regras, permitindo não apenas melhorar a acurácia das predições, mas também oferecer suporte à tomada de decisão clínica de maneira compreensível e alinhada às diretrizes médicas.

O presente estudo foi estruturado em três frentes principais de desenvolvimento. A primeira refere-se à implementação do aplicativo; a segunda diz respeito à definição de uma metodologia para extração automática de variáveis a partir de exames laboratoriais em formato PDF; e a terceira envolve a proposição de uma metodologia para classificação e previsão do DM2.

Com o objetivo de apresentar de forma clara como cada um dos objetivos específicos foi atendido, optou-se por organizá-los em tópicos, acompanhados de suas respectivas análises:

- **Avaliar o estado da arte acerca da utilização da lógica *fuzzy* para classificação, a extração de informação de arquivos PDF (*Portable Document Format*) e da utilização do modelo ANFIS (*Adaptive Neuro-Fuzzy Inference System*) para classificação de doenças**;

  – O primeiro objetivo específico concentra-se na análise do estado da arte de temas fundamentais para a condução da pesquisa, a saber: a aplicação da lógica *fuzzy* em problemas de classificação, técnicas de extração de informação a partir de arquivos PDF e o uso de modelos baseados em ANFIS para classificação

de doenças. Para tanto, foram realizadas três revisões sistemáticas da literatura, das quais duas resultaram em artigos científicos publicados, contribuindo diretamente para a fundamentação teórica e validação da proposta desenvolvida.

- **Desenvolver um método baseado em aprendizado de máquina para extrair um conjunto de informações de saúde automaticamente a partir de documentos com resultados de exames médicos:**

  – Foi definida uma metodologia para extração automática de variáveis a partir de exames de sangue em arquivos PDF. Os resultados demonstraram desempenho expressivo, evidenciado por elevados valores de precisão, *recall* e *F1-score*. Verificou-se, ainda, que a abordagem é eficaz tanto para PDFs nativos quanto para documentos baseados em imagem. Contudo, observou-se que, neste último caso, a qualidade da imagem impacta diretamente o desempenho do sistema, podendo comprometer a acurácia em situações de baixa resolução ou presença de ruído visual.

- **Desenvolver modelos interpretáveis baseados em regras *fuzzy* para desfechos em saúde a partir de variáveis extraídas automaticamente:**

  – Foi proposto um Sistema de Inferência Neuro-*Fuzzy* Adaptativo baseado em TSK, estruturado em dois níveis, para classificação laboratorial de pacientes e estimativa do risco metabólico atual associado ao DM2. A abordagem contempla um modelo baseado em glicose e hemoglobina glicada, e um modelo de avaliação de risco metabólico que incorpora variáveis adicionais como IMC, sexo, raça e diagnóstico prévio.

  Os resultados evidenciam que a abordagem *fuzzy* é eficaz no tratamento de incertezas em dados clínicos, apresentando elevada precisão, revocação e acurácia. O modelo de risco também apresentou desempenho consistente, com *Example-Based Accuracy* de 0,9040 e *Average Precision* de 0,9540, demonstrando robustez frente a entradas heterogêneas.

  A integração entre os dois níveis reflete o fluxo clínico real, no qual diagnósticos prévios influenciam a avaliação de risco metabólico. Além disso, o uso de regras *fuzzy* garante transparência e alinhamento com padrões médicos, enquanto a representação contínua do risco metabólico atual se mostra particularmente relevante em aplicações em saúde.

  Em síntese, a metodologia proposta é robusta, interpretável e adequada para sistemas de apoio à decisão, evidenciando o potencial dos sistemas *fuzzy* TSK como alternativa aos métodos tradicionais baseados em limiares.

- **Desenvolver a plataforma propriamente dita**:

  - Foi desenvolvido um aplicativo integrador voltado à interação entre pacientes e profissionais de saúde. A plataforma permite o envio de exames, a extração automática de informações e a visualização dos dados em formato de linha do tempo, facilitando a análise longitudinal e o acompanhamento clínico. Além disso, possibilita a comunicação entre pacientes, médicos, enfermeiros e cuidadores, contribuindo para um cuidado mais eficiente e integrado.

Dessa forma, conclui-se que o trabalho atingiu de maneira satisfatória todos os objetivos propostos, apresentando resultados consistentes e evidenciando a viabilidade prática das metodologias desenvolvidas no contexto da saúde digital.

Além dos objetivos, foram estabelecidas duas hipóteses científicas e uma hipótese tecnológica:

- **Hipótese Científica**: Métodos de extração de informações de exames médicos guiados por aprendizado de máquina podem mitigar o problema da partida fria da rede e melhorar o apoio à classificação laboratorial e à tomada de decisão clínica quando comparados a métodos clássicos;

- **Hipótese Científica**: Modelos interpretáveis baseados em regras *fuzzy* para detectar anomalias em variáveis colhidas manualmente ou por meio de *wearables* (alimentação, sono, frequência cardíaca, glicemia, exercício físico etc.) permitem compreender os relacionamentos entre os preditores e as variáveis alvo relacionadas às patologias do paciente, em particular DM2.

- **Hipótese Tecnológica**: Uma plataforma, no formato de linha do tempo, que se constitua em um repositório de informações pregressas e atuais de saúde do usuário e que esteja conectado a uma rede social, formada por profissionais de saúde e responsáveis pelo seu cuidado, pode auxiliar na prevenção de eventos e desfechos através do monitoramento e controle da saúde dos usuários.

Os resultados obtidos demonstram forte aderência a essas hipóteses. A extração automática de dados aliada ao modelo neuro-*fuzzy* contribuiu para melhor desempenho na classificação laboratorial e na estratificação do risco metabólico atual. Os modelos baseados em regras demonstraram elevada interpretabilidade, permitindo compreender relações entre variáveis clínicas e risco de doença. Por fim, a plataforma desenvolvida evidencia a viabilidade de aplicação prática em sistemas de monitoramento em saúde.

Assim, conclui-se que o trabalho valida tanto as hipóteses científicas quanto a tecnológica, fornecendo base sólida para aplicações reais em saúde digital.

8.1 Limitações do Estudo

Apesar dos resultados promissores obtidos, algumas limitações devem ser consideradas na interpretação dos achados e no planejamento de trabalhos futuros.

Em primeiro lugar, o modelo desenvolvido no Nível 1 baseia-se exclusivamente em critérios laboratoriais, utilizando principalmente os valores de glicemia e hemoglobina glicada para classificação dos indivíduos. Dessa forma, o sistema não substitui o diagnóstico clínico confirmado, que requer avaliação médica abrangente, incluindo histórico clínico, sintomas, exames complementares e acompanhamento do paciente.

Além disso, o estudo não contemplou formas específicas de diabetes, como o Diabetes Autoimune Latente do Adulto (LADA) e o Diabetes Monogênico de Início na Juventude (MODY). Embora menos prevalentes que o Diabetes Mellitus Tipo 1 e Tipo 2, essas condições apresentam características clínicas e fisiopatológicas distintas, podendo demandar abordagens diagnósticas e terapêuticas específicas.

Outra limitação refere-se à natureza transversal da análise realizada no Nível 2. Como o modelo foi construído a partir de dados observados em um único momento, sua aplicação está voltada para a classificação do estado atual do paciente, não permitindo estimar o risco prospectivo de desenvolvimento da doença ao longo do tempo, conforme realizado por modelos epidemiológicos de predição de risco.

Também deve ser destacado que o conjunto de variáveis preditoras disponível foi relativamente restrito quando comparado aos fatores tradicionalmente utilizados na literatura clínica. Informações relevantes, como histórico familiar de diabetes, hábitos de vida, nível de atividade física, alimentação, circunferência abdominal e presença de comorbidades, não estavam disponíveis na base de dados utilizada. A ausência dessas variáveis limita a capacidade de comparação direta com modelos de risco amplamente consolidados na literatura.

Em relação às variáveis sociodemográficas, a representação da raça foi realizada por meio de uma categorização binária. Embora essa simplificação tenha sido necessária em função da estrutura dos dados disponíveis, ela não representa adequadamente a complexidade da composição étnico-racial da população brasileira, potencialmente reduzindo a capacidade do modelo de capturar diferenças associadas a esse fator.

Adicionalmente, não foi possível realizar uma comparação direta com escores clínicos de rastreamento amplamente utilizados, como FINDRISC, ADA Diabetes Risk Test ou AUSDRISK. A ausência de variáveis fundamentais exigidas por esses instrumentos inviabilizou sua aplicação ao conjunto de dados empregado, impossibilitando a utilização desses métodos como *baseline* comparativo.

Por fim, a plataforma desenvolvida foi avaliada sob a perspectiva técnica e computacional, incluindo testes funcionais, análise de desempenho e validação dos modelos propostos. Entretanto, não foi conduzido um estudo de uso envolvendo pacientes, médicos,

enfermeiros ou outros profissionais da saúde em ambiente real de utilização. Consequentemente, aspectos relacionados à usabilidade, aceitação da tecnologia, impacto no fluxo clínico e experiência dos usuários permanecem como oportunidades para investigações futuras.

Dessa forma, os resultados obtidos devem ser interpretados dentro do contexto das limitações apresentadas. Ao mesmo tempo, tais limitações apontam direções relevantes para pesquisas futuras, incluindo a incorporação de novas variáveis clínicas, a avaliação longitudinal dos pacientes, a investigação de subtipos específicos de diabetes e a validação da plataforma em cenários reais de uso.

8.2 Trabalhos Futuros

Os resultados obtidos nesta pesquisa evidenciam o potencial da plataforma desenvolvida para apoiar o gerenciamento de dados clínicos e a identificação de indivíduos com diabetes. Entretanto, diversas oportunidades de evolução podem ser exploradas em trabalhos futuros, visando ampliar tanto o escopo funcional da solução quanto sua aplicabilidade em contextos clínicos reais.

Uma primeira linha de investigação consiste na ampliação do conjunto de variáveis utilizadas pelos modelos de classificação, incorporando informações clínicas, antropométricas e comportamentais frequentemente empregadas na literatura especializada, tais como histórico familiar de diabetes, índice de massa corporal, circunferência abdominal, hábitos alimentares, nível de atividade física e presença de comorbidades. A inclusão dessas informações poderá contribuir para o aumento da capacidade preditiva dos modelos e possibilitar comparações mais abrangentes com instrumentos de rastreamento consolidados, como FINDRISC, ADA Diabetes Risk Test e AUSDRISK.

Outra perspectiva relevante envolve a realização de estudos longitudinais, permitindo acompanhar os pacientes ao longo do tempo e desenvolver modelos capazes de estimar o risco futuro de desenvolvimento da doença. Essa abordagem possibilitaria a transição de modelos predominantemente classificatórios para modelos preditivos de risco metabólico, ampliando sua utilidade em estratégias de prevenção e intervenção precoce.

Também se mostra promissora a investigação de formas específicas de diabetes, incluindo Diabetes Autoimune Latente do Adulto (LADA) e Diabetes Monogênico de Início na Juventude (MODY). A incorporação dessas condições poderá contribuir para o desenvolvimento de mecanismos de apoio ao diagnóstico diferencial, tornando a plataforma mais abrangente e alinhada à prática clínica especializada.

No contexto da inteligência artificial, futuras pesquisas poderão comparar a abordagem proposta com diferentes técnicas de modelagem e aprendizado de máquina, incluindo métodos estatísticos tradicionais, algoritmos baseados em árvores, modelos de *ensemble* e técnicas de aprendizado profundo. Essas comparações poderão fornecer evidências adi-

cionais sobre o desempenho, interpretabilidade e robustez da solução desenvolvida.

Do ponto de vista da plataforma computacional, uma evolução natural consiste na implementação de mecanismos de monitoramento contínuo e geração automática de alertas clínicos. A partir da análise dos resultados laboratoriais e demais informações disponíveis, o sistema poderá notificar usuários e profissionais de saúde sobre alterações relevantes, contribuindo para a identificação precoce de situações que demandem atenção clínica.

Outra direção importante refere-se à validação da plataforma em ambientes reais de utilização, envolvendo pacientes, médicos, enfermeiros e demais profissionais da saúde. Estudos de usabilidade, aceitação tecnológica e impacto clínico poderão fornecer evidências sobre a efetividade da solução no suporte à tomada de decisão e no acompanhamento dos pacientes.

Além disso, a plataforma poderá ser expandida para contemplar uma abordagem multidisciplinar, permitindo a participação integrada de profissionais de diferentes áreas da saúde, como nutricionistas, psicólogos, farmacêuticos, fisioterapeutas e educadores físicos. Essa expansão tem potencial para favorecer o cuidado integral dos pacientes e promover uma gestão mais abrangente das condições crônicas de saúde.

Por fim, a metodologia proposta poderá ser adaptada e aplicada ao monitoramento e classificação de outras doenças crônicas, como hipertensão arterial sistêmica, doenças cardiovasculares e distúrbios metabólicos. Essa ampliação contribuirá para transformar a plataforma em um ambiente unificado de apoio ao acompanhamento longitudinal da saúde, aumentando seu potencial de impacto tanto em contextos clínicos quanto em iniciativas de saúde pública.

REFERÊNCIAS

[1] KALAISELVI, R.; MEENA, K.; VANITHA, V. Liver disease prediction using machine learning algorithms. In: IEEE. *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*. [S.l.], 2021. p. 1–6.

[2] GAETE, R. A. C.; PINTO, I. C. Informatização do processo de enfermagem na atenção primária à saúde. In: SBC. *Anais Estendidos do XXI Simpósio Brasileiro de Computação Aplicada à Saúde*. [S.l.], 2021. p. 31–36.

[3] JIN, M. et al. A scoping review of clinical unstructured text information extraction. In: IEEE. *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*. [S.l.], 2020. p. 853–857.

[4] AWAN, A. A. *O que é tokenização?* 2024. Disponível em: <https://www.datacamp.com/pt/blog/what-is-tokenization>. Acesso em: 21 agosto 2024.

[5] JAUME, G.; EKENEL, H. K.; THIRAN, J.-P. Funsd: A dataset for form understanding in noisy scanned documents. In: IEEE. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. [S.l.], 2019. v. 2, p. 1–6.

[6] SANTOS, G. L. dos. *Uma abordagem para Form Understanding baseada em multitarefas aplicada a documentação técnica de engenharia*. 106 p. Dissertação (Mestrado) — Universidade Federal do Rio Grande, Rio Grande, 2022.

[7] LASSOUED, Y.; DELERIS, L. Thesaurus-based hierarchical semantic grouping of medical terms in information extraction. In: *Exploring Complexity in Health: An Interdisciplinary Systems Approach*. [S.l.]: IOS Press, 2016. p. 446–450.

[8] KIPF, T. N.; WELLING, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[9] PHAM, M. et al. Semantic labeling: a domain-independent approach. In: SPRINGER. *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part I 15*. [S.l.], 2016. p. 446–462.

[10] HAN, X.; SUN, L. An entity-topic model for entity linking. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. [S.l.: s.n.], 2012. p. 105–115.

[11] RODRIGUES, J. *Tudo o que você precisa saber sobre Redes Sociais*. 2023. Disponível em: <https://www.rdstation.com/blog/marketing/redes-sociais/>. Acesso em: 14 agosto 2024.

[12] DIANA, J. *Redes Sociais*. 2024. Disponível em: <https://www.todamateria.com.br/redes-sociais/>. Acesso em: 14 agosto 2024.

[13] SAúDE, M. da. *Diabetes (diabetes mellitus)*. 2024. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/d/diabetes>. Acesso em: 15 agosto 2024.

[14] DIABETES, S. B. de. *Diabetes*. 2024. Disponível em: <https://diabetes.org.br/>. Acesso em: 15 agosto 2024.

[15] American Diabetes Association Professional Practice Committee. 2. diagnosis and classification of diabetes: Standards of care in diabetes—2025. *Diabetes Care*, v. 48, n. Supplement$_1$, $p.S27 - -S49, 2025$.

[16] World Health Organization. *Global Report on Diabetes*. Geneva, Switzerland: World Health Organization, 2016. ISBN 9789241565257. Disponível em: <https://iris.who.int/handle/10665/204871>.

[17] International Diabetes Federation. *IDF Diabetes Atlas*. 10. ed. Brussels, Belgium: International Diabetes Federation, 2021. ISBN 978-2-930229-98-0. Disponível em: <https://diabetesatlas.org>.

[18] GASPARIN, A. *Exame de sangue: quais são os principais e sua importância*. 2023. Disponível em: <https://hilab.com.br/blog/exame-de-sangue/>. Acesso em: 15 agosto 2024.

[19] BEM, C. do. *Exames de sangue: quais são os mais comuns?* 2022. Disponível em: <https://www.clinicadobem.com.br/sem-categoria/exames-de-sangue-quais-sao-os-mais-comuns/>. Acesso em: 15 agosto 2024.

[20] MELO, K. *Para que serve o exame de sangue?* 2021. Disponível em: <https://www.acessosaudesp.com.br/single-post/para-que-serve-o-exame-de-sangue>. Acesso em: 15 agosto 2024.

[21] SAúDE, B. *Quais são os tipos de exame de sangue mais solicitados e como se preparar para a testagem?* 2023. Disponível em: <https://beepsaude.com.br/como-fazer-exame-de-sangue/>. Acesso em: 15 agosto 2024.

[22] ZADEH, L. A. Fuzzy sets. *Information and Control*, v. 8, n. 3, p. 338–353, 1965.

[23] KLIR, G. J.; YUAN, B. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. [S.l.]: Prentice Hall, 1995.

[24] ROSS, T. J. *Fuzzy Logic with Engineering Applications*. 3. ed. [S.l.]: Wiley, 2010.

[25] GRABISCH, M. et al. *Aggregation functions*. [S.l.]: Cambridge University Press, 2009.

[26] KLEMENT, E. P.; MESIAR, R.; PAP, E. *Triangular Norms*. Dordrecht: Kluwer Academic Publisher, 2000.

[27] JANG, J.-S. Anfis: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, IEEE, v. 23, n. 3, p. 665–685, 1993.

[28] TAKAGI, T.; SUGENO, M. Fuzzy identification of systems and its applications to modeling and control. *IEEE transactions on systems, man, and cybernetics*, IEEE, n. 1, p. 116–132, 1985.

[29] SUGENO, M.; KANG, G. Fuzzy modelling and control of multilayer incinerator. *Fuzzy sets and systems*, Elsevier, v. 18, n. 3, p. 329–345, 1986.

[30] SUGENO, M.; KANG, G. Structure identification of fuzzy model. *Fuzzy sets and systems*, Elsevier, v. 28, n. 1, p. 15–33, 1988.

[31] LOU, Q. et al. Multilabel takagi-sugeno-kang fuzzy system. *IEEE Transactions on Fuzzy Systems*, IEEE, v. 30, n. 9, p. 3410–3425, 2021.

[32] KEELE, S. et al. *Guidelines for performing systematic literature reviews in software engineering*. [S.l.], 2007, 65 p.

[33] NAYAK, R. et al. A semi-automatic data extraction system for heterogeneous data sources: a case study from cotton industry. In: SPRINGER. *Data Mining: 19th Australasian Conference on Data Mining, AusDM 2021, Brisbane, QLD, Australia, December 14-15, 2021, Proceedings 19*. [S.l.], 2021. p. 209–222.

[34] CHOWDHURY, A. G.; SCHUT, N.; ATZMUELLER, M. A hybrid information extraction approach using transfer learning on richly-structured documents. In: *LWDA*. [S.l.: s.n.], 2021. p. 13–25.

[35] AHUJA, A. et al. A new annotation method and dataset for layout analysis of long documents. In: *Companion Proceedings of the ACM Web Conference 2023*. [S.l.: s.n.], 2023. p. 834–842.

[36] FARAJI-BIREGANI, M.; NEMATBAKHSH, N. Diabetes prediction recommender system based on artificial neural networks and sine-cosine optimization algorithm. In: IEEE. *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*. [S.l.], 2019. p. 263–268.

[37] ALAMOUDI, A. et al. A rule-based information extraction approach for extracting metadata from pdf books. *ICIC Express Letters, Part B: Applications*, ICIC Express Letters Office, v. 12, n. 2, p. 121–132, 2021.

[38] PARIZI, R. M. et al. Cyberpdf: smart and secure coordinate-based automated health pdf data batch extraction. In: *Proceedings of the 2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies*. [S.l.: s.n.], 2018. p. 106–111.

[39] AL-ZAIDY, R. A.; GILES, C. L. Extracting semantic relations for scholarly knowledge base construction. In: IEEE. *2018 IEEE 12th international conference on semantic computing (ICSC)*. [S.l.], 2018. p. 56–63.

[40] ALGHAMDI, H. et al. Extracting toc and metadata from pdf books: A rule-based approach. *ICIC Express Letters*, v. 12, n. 2, p. 133–143, 2022.

[41] HA, H. T.; HORÁK, A. Information extraction from scanned invoice images using text analysis and layout features. *Signal Processing: Image Communication*, Elsevier, v. 102, p. 116601, 2022.

[42] OUSSAID, I. et al. Information extraction from visually rich documents with font style embeddings. In: IEEE. *2022 26th International Conference on Pattern Recognition (ICPR)*. [S.l.], 2022. p. 1657–1663.

[43] GUPTA, A. et al. Intelligent information retrieval: Handling variability in document structure. In: IEEE. *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*. [S.l.], 2022. p. 1635–1640.

[44] ZENG, X. et al. Method of automatically extracting squeeze casting process data from the text of periodical literature. In: IEEE. *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)*. [S.l.], 2021. p. 416–422.

[45] MEUSCHKE, N. et al. A benchmark of pdf information extraction tools using a multi-task and multi-domain evaluation framework for academic documents. In: SPRINGER. *International Conference on Information*. [S.l.], 2023. p. 383–405.

[46] CHUMWATANA, T.; RATTANA-UMNUAYCHAI, W. Using ocr framework and information extraction for thai documents digitization. In: IEEE. *2021 9th International Electrical Engineering Congress (iEECON)*. [S.l.], 2021. p. 440–443.

[47] GELMAN, B. et al. Toward a robust method for understanding the replicability of research. In: *SDU@ AAAI*. [S.l.: s.n.], 2021.

[48] BERARDI, M. et al. Text mining from free unstructured text: an experiment of time series retrieval for volcano monitoring. *Applied Sciences*, MDPI, v. 12, n. 7, p. 3503, 2022.

[49] ANGGAKUSUMA, J.; MAWARDI, V. C.; LAURO, M. D. Resume extraction with conditional random field method. *IOP Conference Series: Materials Science and Engineering*, v. 1007, n. 1, p. 012154, 2020.

[50] OPASJUMRUSKIT, K. et al. Ontohuman: Ontology-based information extraction tools with human-in-the-loop interaction. In: SPRINGER. *International Conference on Cooperative Design, Visualization and Engineering*. [S.l.], 2022. p. 68–74.

[51] HASHMI, A. M.; QAYYUM, F.; AFZAL, M. T. Insights to the state-of-the-art pdf extraction techniques. *IPSI Trans. Internet Res*, v. 16, n. 8, p. 1–8, 2020.

[52] ZHOU, N. et al. Optimal dynamic treatment regime estimation using information extraction from unstructured clinical text. *Biometrical Journal*, Wiley Online Library, v. 64, n. 4, p. 805–817, 2022.

[53] ZHU, W. et al. Knowledge extraction of long-term complications from clinical narratives of blood cancer patients with hct treatments. In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. [S.l.: s.n.], 2018. p. 525–526.

[54] GOEDHART, M. *Hoe kan het algoritme voor documentanalyse verbeterd worden zodat deze de algehele structuur van een document beter herkent?* Tese (Doutorado) — Saxion, 2023.

[55] VENKATESH, P. R. et al. Knowledge graphs using cloud services. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, v. 12, n. 10, p. 3979–3991, 2021.

[56] MAHANI, G. K.; PAJOOHAN, M.-R. Predicting lab values for gastrointestinal bleeding patients in the intensive care unit: A comparative study on the impact of comorbidities and medications. *Artificial Intelligence in Medicine*, Elsevier, v. 94, p. 79–87, 2019.

[57] WEDAGU, M. A. et al. Medicine recommendation system for diabetes using prior medical knowledge. In: *Proceedings of the 2020 4th International Conference on Vision, Image and Signal Processing*. [S.l.: s.n.], 2020. p. 1–5.

[58] DEIF, M.; HAMMAM, R.; SOLYMAN, A. Adaptive neuro-fuzzy inference system (anfis) for rapid diagnosis of covid-19 cases based on routine blood tests. *Int. J. Intell. Eng. Syst*, v. 14, n. 2, 2021.

[59] JAMUNA, D. K. M. K. S. Prediction of diabetes and clustering based on its levels using fuzzy c means algorithm. *International Journal of Scientific & Technology research*, v. 9, n. 2, 2020.

[60] SRIVASTAVA, S. et al. Prediction of diabetes using artificial neural network approach. In: SPRINGER. *Engineering Vibration, Communication and Information Processing: ICoEVCI 2018, India*. [S.l.], 2019. p. 679–687.

[61] OSMANOVIĆ, A. et al. Machine learning techniques for classification of breast cancer. In: SPRINGER. *World Congress on Medical Physics and Biomedical Engineering 2018: June 3-8, 2018, Prague, Czech Republic (Vol. 1)*. [S.l.], 2019. p. 197–200.

[62] MIRJALILI, S.; MIRJALILI, S. Evolutionary radial basis function networks. *Evolutionary Algorithms and Neural Networks: Theory and Applications*, Springer, p. 105–139, 2019.

[63] COMMUNITY, T. global diabetes. *Blood Sugar Level Ranges*. 2023. Urlhttps://www.diabetes.co.uk/.

[64] GOLDAR, S. Z. et al. An anfis-pso algorithm for predicting four grades of non-alcoholic fatty liver disease. In: IEEE. *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. [S.l.], 2020. p. 1–5.

[65] ALASAADY, M. T. et al. A proposed approach for diabetes diagnosis using neuro-fuzzy technique. *Bulletin of Electrical Engineering and Informatics*, v. 11, n. 6, p. 3590–3597, 2022.

[66] PRIYADARSHINI, L.; SHRINIVASAN, L. Design of an ANFIS based Decision Support System for Diabetes Diagnosis. In: *2020 International Conference on Com-*

munication and Signal Processing (ICCSP). Chennai, India: IEEE, 2020. p. 1486–1489.

[67] SINGH, B. K.; NIRALA, N. et al. Expert diagnostic system for detection of hypertension and diabetes mellitus using discrete wavelet decomposition of photoplethysmogram signal and machine learning technique. *Medicine in Novel Technology and Devices*, Elsevier, v. 19, p. 100251, 2023.

[68] HOSSEINI, H.; BARDSIRI, A. K. Improving diagnosis accuracy of diabetic disease using radial basis function network and fuzzy clustering. *Frontiers in Health Informatics*, v. 8, n. 1, p. 24, 2019.

[69] SHYNU, P. G. et al. Blockchain-Based Secure Healthcare Application for Diabetic-Cardio Disease Prediction in Fog Computing. *IEEE Access*, v. 9, p. 45706–45720, 2021.

[70] PADMARAJ, R.; SURESH, D.; RAVICHANDRAN, S. Neiwda with hybrid ensemble model for diagnosis of diabetes type ii. *Int J Adv Sci Technol*, v. 29, p. 5765–5778, 2020.

[71] SINGH, A. et al. ediapredict: an ensemble-based framework for diabetes prediction. *ACM Transactions on Multimidia Computing Communications and Applications*, ACM New York, NY, v. 17, n. 2s, p. 1–26, 2021.

[72] LAUJAN; URBAN, E. *Modelo de layout da Informação de Documentos*. 2023. Disponível em: <https://learn.microsoft.com/pt-br/azure/ai-services/document-intelligence/concept-layout?view=doc-intel-3.0.0>. Acesso em: 15 setembro 2023.

[73] LAUJAN; URBAN, E. *Usar modelos de Informação de Documentos*. 2023. Disponível em: <https://learn.microsoft.com/pt-br/azure/ai-services/document-intelligence/how-to-guides/use-sdk-rest-api?view=doc-intel-3.0.0&preserve-view=true%3Fpivots%3Dprogramming-language-python&tabs=windows&pivots=programming-language-python>. Acesso em: 21 setembro 2023.

[74] LAUJAN; URBAN, E. *Introdução à Informação de Documentos*. 2023. Disponível em: <https://learn.microsoft.com/pt-br/azure/ai-services/document-intelligence/quickstarts/get-started-sdks-rest-api?view=doc-intel-3.1.0&preserve-view=true&pivots=programming-language-python>. Acesso em: 15 setembro 2023.

[75] MANNING, C. D. *Introduction to information retrieval*. [S.l.]: Syngress Publishing,, 2008.

[76] JUNG, M. O. et al. Sint multiverse: Prototype simulator for generating synthetic health determinant data for explainable ai training. *IEEE Access*, p. 1–1, 2026.

[77] RODACKI, M. et al. *Diagnóstico de diabetes mellitus*. 2024. Disponível em: <https://diretriz.diabetes.org.br/diagnostico-de-diabetes-mellitus/>. Acesso em: 05 maio 2025.

[78] Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *The New England Journal of Medicine*, v. 346, n. 6, p. 393–403, 2002. Disponível em: <https://doi.org/10.1056/NEJMoa012512>.

[79] TUOMILEHTO, J. et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *The New England Journal of Medicine*, v. 344, n. 18, p. 1343–1350, 2001. Disponível em: <https://doi.org/10.1056/NEJM200105033441801>.

[80] FRASER, C. G. Biological variation: From principles to practice. *AACC Press*, AACC Press, 2001.

[81] JAFARPISHEH, N. et al. A deep neural network combined with radial basis function for abnormality classification. *Mobile Networks and Applications*, v. 26, p. 2318–2328, 2021.

[82] SAVITHA, V. et al. A distributed key authentication and OKM-ANFIS scheme based breast cancer prediction system in the IoT environment. *Journal of Ambient Intelligence and Humanized Computing*, v. 12, n. 2, p. 1757–1769, 2021.

[83] KHAN, M.; ALGARNI, F. A healthcare monitoring system for the diagnosis of heart disease in the iomt cloud environment using msso-anfis. *IEEE Access*, v. 8, p. 122259–122269, 2020.

[84] CASTELLAZZI, G. et al. A Machine Learning Approach for the Differential Diagnosis of Alzheimer and Vascular Dementia Fed by MRI Selected Features. *Frontiers in Neuroinformatics*, v. 14, p. 25, 2020. ISSN 1662-5196.

[85] ARAB, S.; REZAEE, K.; MOGHADDAM, G. A novel fuzzy expert system design to assist with peptic ulcer disease diagnosis. *Cogent Engineering*, v. 8, n. 1, jan. 2021. ISSN 2331-1916.

[86] PUNARSELVAM, E. A pragmatic approach of Parkinson disease detection using hybrid case-based reasoning neuro-fuzzy classification system over Mobile edge computing. *Journal of Intelligent and Fuzzy Systems*, v. 44, n. 5, p. 7653–7668, 2023.

[87] NILASHI, M. et al. A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. *Journal of Infection and Public Health*, v. 12, n. 1, p. 13–20, 2019. ISSN 1876-035X.

[88] PRIYA, S. M.; A., S. P. A Robust Cardiovascular Disease Predictor Based on Genetic Feature Selection and Ensemble Learning Classification. *International journal of electrical and computer engineering systems*, v. 14, n. 7, p. 799–809, set. 2023. ISSN 18477003, 18476996.

[89] ANTER, A. et al. A robust swarm intelligence-based feature selection model for neuro-fuzzy recognition of mild cognitive impairment from resting-state fMRI. *Information Sciences*, v. 503, p. 670–687, 2019.

[90] PADMAVATHY, T.; VIMALKUMAR, M.; BHARGAVA, D. Adaptive clustering based breast cancer detection with ANFIS classifier using mammographic images. *Cluster Computing*, v. 22, p. 13975–13984, 2019.

[91] JEYSHRI, J.; KOWSIGAN, M. Adaptive fuzzy-region growing fusion and improved CNN-ANFIS-based automated segmentation and classification of cervical cancer. *Concurrency and Computation: Practice and Experience*, v. 35, n. 3, 2023.

[92] DEVI, R. L.; KALAIVANI, V.; VELAMMAL, B. Adaptive neuro-fuzzy inference healthcare system for cardiac arrhythmia detection using heart rate variability features. In: . [S.l.: s.n.], 2021. v. 700.

[93] NALLASIVAM, M. P. et al. Advanced Glaucoma Detection by Integrating ANFIS with CNN Models for Enhanced Diagnosis. In: *2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT)*. Kollam, India: IEEE, 2024. p. 250–254.

[94] PILEHVARI, S.; MORGAN, Y.; PENG, W. An analytical review on the use of artificial intelligence and machine learning in diagnosis, prediction, and risk factor analysis of multiple sclerosis. *Multiple Sclerosis and Related Disorders*, v. 89, p. 105761, set. 2024.

[95] GOLDAR, S. Z. et al. An ANFIS-PSO Algorithm for Predicting Four Grades of Non-Alcoholic Fatty Liver Disease. In: *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. Ankara, Turkey: IEEE, 2020. p. 1–5.

[96] PANKAJAVALLI, P. B.; KARTHICK, G. S.; SAKTHIVEL, R. An Efficient Machine Learning Framework for Stress Prediction via Sensor Integrated Keyboard Data. *IEEE Access*, v. 9, p. 95023–95035, 2021.

[97] KAVITHA, D.; RENUMADHAVI, C. An efficient multilayer deep detection perceptron (mlddp) methodology for detecting testicular anomalies with or without congenital heart disease (tachd). *The Journal of Supercomputing*, Springer, v. 78, n. 3, p. 4057–4072, 2022.

[98] SHINEY, T. S. S.; JEROME, S. A. An Intelligent System to Enhance the Performance of Brain Tumor Diagnosis from MR Images. *Journal of Digital Imaging*, v. 36, n. 2, p. 510–525, nov. 2022.

[99] KUMAR, V. et al. An Intensive Review Analysis about Fuzzy Logic usage in Smart Medical Care System. In: *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. Greater Noida, India: IEEE, 2024. p. 918–923.

[100] SAIKIA, S. J.; NIRMALA, S. R. An NB-ANN based Fusion Approach for Disease Genes Prediction and LFKH-ANFIS Classifier for Eye Diseases Identification. *International Journal of Advanced Computer Science and Applications*, v. 12, n. 10, 2021.

[101] SURYANI, E. et al. Application of the Adaptive-Neuro Fuzzy Inference System (ANFIS) Method for Rapid Diagnosis COVID-19 Patient Cases. In: *2022 5th International Conference on Information and Communications Technology (ICOIACT)*. Yogyakarta, Indonesia: IEEE, 2022.

[102] NGUYEN, T.-L. et al. Attentive Hierarchical ANFIS with interpretability for cancer diagnostic. *Expert Systems with Applications*, v. 201, 2022.

[103] MANDAL, S.; ROY, A.; MONDAL, P. Automated detection of fibrillations and flutters based on fused feature set and ANFIS classifier. *Biomedical Signal Processing and Control*, v. 69, 2021.

[104] AL-NAAMI, B. et al. Automated Detection of Left Bundle Branch Block from ECG Signal Utilizing the Maximal Overlap Discrete Wavelet Transform with ANFIS. *Computers*, v. 11, n. 6, p. 93, jun. 2022.

[105] CASALINO, G. et al. Balancing Accuracy and Interpretability through Neuro-Fuzzy Models for Cardiovascular Risk Assessment. In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. Orlando, FL, USA: IEEE, 2021. p. 1–8.

[106] RUPESHY, P.; BEE, M. M.; SURESH, V. Bandlet transform based brain tumor detection and classification of Magnetic resonance image using Coactive Neuro Fuzzy Inference System in comparison with Adaptive Neuro Fuzzy Inference System classifier. In: *2022 14th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*. Karachi, Pakistan: IEEE, 2022. p. 1–5.

[107] RAHOUMA, K.; ALY, R.; HAMED, H. Brain Cancer Diagnosis and Prediction Based on Neural Gas Network and Adaptive Neuro Fuzzy. In: . [S.l.: s.n.], 2019. v. 163, p. 518–526.

[108] VANKDOTHU, R.; HAMEED, M. A. Brain tumor segmentation of MR images using SVM and fuzzy classifier in machine learning. *Measurement: Sensors*, v. 24, p. 100440, dez. 2022.

[109] THAWKAR, S. et al. Breast cancer: A hybrid method for feature selection and classification in digital mammography. *International Journal of Imaging Systems and Technology*, v. 33, p. 1696–1712, 2023.

[110] IBRAHIM, S.; KHALID, N. E. A.; MANAF, M. CAPSOCA: hybrid technique for nosologic segmentation of primary brain tumors. *Indonesian Journal of Electrical Engineering and Computer Science*, v. 16, n. 1, p. 267, out. 2019.

[111] RIVERA, J.; RODRIGUEZ, K.; YU, X.-H. Cardiovascular Conditions Classification Using Adaptive Neuro-Fuzzy Inference System. In: *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. New Orleans, LA, USA: IEEE, 2019. p. 1–6. ISBN 978-1-5386-1728-1. Disponível em: <https://ieeexplore.ieee.org/document/8858896/>.

[112] KR, V. et al. Classification of adaptive back propagation neural network along with fuzzy logic in chronic kidney disease. *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, v. 7, p. 100463, mar. 2024.

[113] AARTHI, R.; PRABHA, K. H. Classification of brain neoplasm from multimodality MRI with the aid of ANFIS classifier. *Multidimensional Systems and Signal Processing*, v. 32, p. 933–957, 2021.

[114] DEHGHANDAR, M.; REZVANI, S. Classification of COVID-19 Individuals Using Adaptive Neuro-Fuzzy Inference System. *Journal of Medical Signals and Sensors*, v. 12, n. 4, p. 334–340, 2022.

[115] INTAN, I.; SOETIKNO, Y. Classification of Thyroid Carcinoma using Sobel Edge Detection and Adaptive Neuro Fuzzy Inference System Approach. In: . [S.l.: s.n.], 2019.

[116] VAIRAVEL, C.; NITHYA, N. S. Data Analytics on Unpredictable Pregnancy Data Records Using Ensemble Neuro-Fuzzy Techniques. *Computer Systems Science and Engineering*, v. 46, n. 2, p. 2159–2175, 2023. ISSN 0267-6192.

[117] KUMAR, A. et al. Depress-DCNF: A deep convolutional neuro-fuzzy model for detection of depression episodes using IoMT. *Applied Soft Computing*, v. 122, 2022.

[118] RAJESHWARI, J.; SUGHASINY, M. Dermatology disease prediction based on firefly optimization of ANFIS classifier. *AIMS Electronics and Electrical Engineering*, v. 6, n. 1, p. 61–80, 2022.

[119] SRINATH, R.; GAYATHRI, R. Detection and classification of electroencephalogram signals for epilepsy disease using machine learning methods. *International Journal of Imaging Systems and Technology*, v. 31, n. 2, p. 729–740, 2021.

[120] RAMASAMY, R.; CHINNASAMY, C. Detection and segmentation of cancer regions in cervical images using fuzzy logic and adaptive neuro fuzzy inference system classification method. *International Journal of Imaging Systems and Technology*, v. 30, n. 2, p. 412–420, 2020.

[121] KALAM, R.; THOMAS, C.; RAHIMAN, M. Detection of brain tumor in MRI images using optimized ANFIS classifier. *International Journal of Uncertainty, Fuzziness and Knowldege-Based Systems*, v. 29, n. Suppl1, p. 1–29, 2021.

[122] DAS, S. et al. Detection of Chronic Kidney Disease Using Neuro-Fuzzy Rule-based Classifier. In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Las Vegas, NV, USA: IEEE, 2022. p. 2839–2845.

[123] SEIFALLAHI, M. et al. Detection of mild cognitive Impairment from gait using Adaptive Neuro-Fuzzy Inference system. v. 71, 2022.

[124] BUI, C. et al. Developing Neural-fuzzy-based Unscented Kalman Filter Algorithm for Atrial Fibrillation Onset Prediction. In: . [S.l.: s.n.], 2020. v. 69, p. 119–125.

[125] FATHI, S. et al. Development and use of a clinical decision support system for the diagnosis of social anxiety disorder. *Computer Methods and Programs in Biomedicine*, v. 190, 2020.

[126] MASOODI, Z.; MANOCHA, A. Dew Computing enabled ANFIS Model for Accurate Classification of Season-Stratified Acute Respiratory Infections in High Altitude Region. In: . [S.l.: s.n.], 2024. p. 1669–1674.

[127] MANOCHA, A.; BHATIA, M.; KUMAR, G. Dew computing-inspired health-meteorological factor analysis for early prediction of bronchial asthma. *Journal of Network and Computer Applications*, v. 179, 2021.

[128] CHIDAMBARAM, S. et al. Diagnosing Breast Cancer Based on the Adaptive Neuro-Fuzzy Inference System. *Computational and Mathematical Methods in Medicine*, v. 2022, 2022.

[129] BABU, M.; NANTHAKUMAR, G. Diagnosis of liver images lesions in mr images using improved segmentation and classification task. *International Journal of Recent Technology and Engineering*, v. 8, n. 1 SpecialIssue4, p. 748–752, 2019.

[130] NABOVATI, E. et al. Diagnosis of Mechanical Low Back PainUsing a Fuzzy Logic-Based Approach. *International Journal of Intelligent Systems and Applications in Engineering*, v. 9, n. 3, p. 116–120, set. 2021.

[131] TZALAVRA, A. G. et al. Dynamic contrast enhanced-magnetic resonance imaging radiomics combined with a hybrid adaptive neuro-fuzzy inference system-particle swarm optimization approach for breast tumour classification. *Expert Systems*, Wiley Online Library, v. 39, n. 4, p. e12895, 2022.

[132] CHITHRA, B.; NEDUNCHEZHIAN, R. Dynamic neutrosophic cognitive map with improved cuckoo search algorithm (dncm-icsa) and ensemble classifier for rheumatoid arthritis (ra) disease. *Journal of King Saud University-Computer and Information Sciences*, Elsevier, v. 34, n. 6, p. 3236–3246, 2022.

[133] TAYLAN, O. et al. Early prediction in classification of cardiovascular diseases with machine learning, neuro-fuzzy and statistical methods. *Biology*, MDPI, v. 12, n. 1, p. 117, 2023.

[134] CÁRDENAS, O. A. et al. Ecg arrhythmia classification based on fuzzy cognitive maps. In: IEEE. *2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*. [S.l.], 2019. p. 1–4.

[135] ALROWAIS, F. et al. Enhanced pelican optimization algorithm with deep learning-driven mitotic nuclei classification on breast histopathology images. *Biomimetics*, MDPI, v. 8, n. 7, p. 538, 2023.

[136] SONG, J. et al. Enhancing cervical precancerous lesion detection using african vulture optimization algorithm with deep learning model. *Biomedical Signal Processing and Control*, Elsevier, v. 97, p. 106665, 2024.

[137] AL-QUDSY, Z. N.; KANONA, R. M.; ALYAWER, S. A. Expert system for diagnosis fever with rash diseases for pediatric age group using adaptive neuro-fuzzy inference system. In: IEEE. *2022 3rd Information Technology To Enhance e-learning and Other Application (IT-ELA)*. [S.l.], 2022. p. 67–72.

[138] GOLGOUNEH, A.; TARVIRDIZADEH, B. Fabrication of a portable device for stress monitoring using wearable sensors and soft computing algorithms. *Neural Computing and Applications*, Springer, v. 32, n. 11, p. 7515–7537, 2020.

[139] AHMED, L. J. et al. Fog-assisted real-time coronary heart disease risk detection in iot-based healthcare system. In: SPRINGER. *3rd EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*. [S.l.], 2022. p. 117–128.

[140] ARJI, G. et al. Fuzzy logic approach for infectious disease diagnosis: A methodical evaluation, literature and classification. *Biocybernetics and biomedical engineering*, Elsevier, v. 39, n. 4, p. 937–955, 2019.

[141] MISHRA, P.; BHOI, N. Genomic signal processing of microarrays for cancer gene expression and identification using cluster-fuzzy adaptive networking. *Soft Computing*, Springer, v. 24, n. 24, p. 18447–18462, 2020.

[142] KEIKHOSROKIANI, P. et al. Heartbeat sound classification using a hybrid adaptive neuro-fuzzy inferences system (anfis) and artificial bee colony. *Digital health*, SAGE Publications Sage UK: London, England, v. 9, p. 20552076221150741, 2023.

[143] SOLIKHAH, M.; NURYANI, N.; NUGROHO, A. S. Identification of atrial fibrillation using descriptive statistic features and adaptive neuro-fuzzy inference system. In: IOP PUBLISHING. *Journal of Physics: Conference Series*. [S.l.], 2019. v. 1153, n. 1, p. 012043.

[144] QIDWAI, U. et al. imigs: An innovative ai based prediction system for selecting the best patient-specific glaucoma treatment. *MethodsX*, Elsevier, v. 10, p. 102209, 2023.

[145] SINGH, C.; GANGWAR, M.; KUMAR, U. Improving accuracy of integrated neuro-fuzzy classifier with fcm-based clustering for diagnosis of psychiatric disorder. *International Journal on Recent and Innovation Trends in Computing and Communication*, 2023.

[146] SINGH, D. et al. Intelligent medical diagnostic system for hepatitis b. *Computers, Materials & Continua*, v. 73, n. 3, 2022.

[147] KARSIDANI, S. D. et al. Intelligent prediction of major adverse cardiovascular events (macce) following percutaneous coronary intervention using anfis-pso model. *BMC Cardiovascular Disorders*, Springer, v. 22, n. 1, p. 389, 2022.

[148] LU, Z. et al. Jointly composite feature learning and autism spectrum disorder classification using deep multi-output takagi-sugeno-kang fuzzy inference systems. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, IEEE, v. 20, n. 1, p. 476–488, 2022.

[149] PASHA, A. et al. Leveraging anfis with adam and pso optimizers for parkinson's disease. *Heliyon*, Elsevier, v. 10, n. 9, 2024.

[150] NUHIĆ, J.; KEVRIĆ, J. Lung cancer typology classification based on biochemical markers using machine learning techniques. In: IEEE. *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*. [S.l.], 2020. p. 292–297.

[151] JULIUS, M. S. et al. Machine learning framework to predict patient non-adherence to medication using non-clinical data: a prognosis approach. In: *Proceedings of the 9th International Conference on Computer and Communications Management*. [S.l.: s.n.], 2021. p. 98–103.

[152] RAYEN, S. J.; SUBHASHINI, R. Mammogram image retrieval using ipso optimized anfis classifier. *Int. J. Innov. Technol. Explor. Eng*, v. 8, n. 9, p. 799–804, 2019.

[153] SUPRIYA, M.; DEEPA, A.; MYTHILI, C. Mamographic image for breast cancer detection and identification of stages of cancer using mffc and optimized anfis. *Journal of Ambient Intelligence and Humanized Computing*, Springer, v. 12, n. 9, p. 8731–8745, 2021.

[154] RASHED, B. M.; POPESCU, N. medical image-based diagnosis using a hybrid adaptive neuro-fuzzy inferences system (anfis) optimized by ga with a deep network model for features extraction. *Mathematics*, MDPI, v. 12, n. 5, p. 633, 2024.

[155] REZAEE, K. et al. Multi-mass breast cancer classification based on hybrid descriptors and memetic meta-heuristic learning. *SN Applied Sciences*, Springer, v. 2, n. 7, p. 1297, 2020.

[156] ALGHAMDI, M. I. Neutrosophic set with adaptive neuro-fuzzy inference system for liver tumor segmentation and classification model. *International Journal of Neutrosophic Science (IJNS)*, v. 18, n. 2, 2022.

[157] PUTRA, R. E. et al. Non-proliferative diabetic retinopathy classification based on hard exudates using combination of frcnn, morphology, and anfis. In: IEEE. *2020 Third international conference on vocational education and electrical engineering (ICVEE)*. [S.l.], 2020. p. 1–6.

[158] LEFTEH, A. et al. Optimization of modified adaptive neuro-fuzzy inference system (manfis) with artificial bee colony (abc) algorithm for classification of bone cancer. In: IEEE. *2022 second international conference on distributed computing and high performance computing (DCHPC)*. [S.l.], 2022. p. 78–81.

[159] HAZNEDAR, B.; ARSLAN, M. T.; KALINLI, A. Optimizing anfis using simulated annealing algorithm for classification of microarray gene expression cancer data. *Medical & Biological Engineering & Computing*, Springer, v. 59, n. 3, p. 497–509, 2021.

[160] NAGALAKSHMI, K.; SURIYA, S. Performance analysis of breast cancer detection method using anfis classification approach. *Computer Systems Science & Engineering*, v. 44, n. 1, 2023.

[161] KAUR, S.; CHAHAL, K. K. Prediction of chikungunya disease using pso-based adaptive neuro-fuzzy inference system model. *International Journal of Computers and Applications*, Taylor & Francis, v. 44, n. 7, p. 641–649, 2022.

[162] RAIESDANA, S. Automated detection of multiple sclerosis lesions using texture-based features and a hybrid classifier. *Caspian Journal of Neurological Sciences*, Caspian Journal of Neurological Sciences, v. 6, n. 1, p. 16–30, 2020.

[163] MADAAN, V.; KAUR, R.; AGRAWAL, P. Rheumatoid arthritis anticipation using adaptive neuro fuzzy inference system. In: IEEE. *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*. [S.l.], 2019. p. 340–346.

[164] SALAHUDDIN, T. et al. Smart neuropathy detection using machine intelligence: Filling the void between clinical practice and early diagnosis. In: IEEE. *2019 Third World Conference on Smart Trends in Systems Security and Sustainablity (WorldS4)*. [S.l.], 2019. p. 141–146.

[165] ABDULLAH, S. B. et al. Spiral optimization algorithm with neurocomputing and fuzzy logic based human gait analysis for osteoarthritis prediction. In: IEEE. *2023 6th International Conference on Engineering Technology and its Applications (IICETA)*. [S.l.], 2023. p. 835–841.

[166] KUSUMADEWI, S.; ROSITA, L.; WAHYUNI, E. G. Stability of classification performance on an adaptive neuro fuzzy inference system for disease complication prediction. *Int J Artif Intell ISSN*, v. 2252, n. 8938, p. 8938, 2023.

[167] BECERRA, M. A. et al. Stochastic-and neuro-fuzzy-analysis-based characterization and classification of 4-channel phonocardiograms for cardiac murmur detection. *Revista Ibérica de Sistemas e Tecnologias de Informação*, Associação Ibérica de Sistemas e Tecnologias de Informacao, n. E33, p. 65–78, 2020.

[168] SIAHAAN, W. F.; SITOMPUL, O.; SITUMORANG, Z. The accuracies of anfis and genetic algorithm with tournament selection on classifying hepatitis data. In: IOP PUBLISHING. *Journal of Physics: Conference Series*. [S.l.], 2020. v. 1566, n. 1, p. 012121.

[169] MANIKANDAN, S.; DEEPIKA, V.; PHILIP, N. Variational mode decomposition based retinal area detection and merging of superpixels in slo image. *Current Signal Transduction Therapy*, Bentham Science Publishers direct, v. 16, n. 2, p. 178–190, 2021.

[170] SUNDARAVADIVEL, T.; MAHALAKSHMI, V. Weighted butterfly optimization algorithm with intuitionistic fuzzy gaussian function based adaptive-neuro fuzzy inference system for covid-19 prediction. *Materials Today: Proceedings*, Elsevier, v. 56, p. 3317–3324, 2022.

[171] HIDAYAH, R. et al. Wheezing feature classification using anfis for asthma diagnosis. In: IEEE. *2024 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*. [S.l.], 2024. p. 1–6.

A TABELAS COM AS INFORMAÇÕES BÁSICAS DOS ARTIGOS SELECIONADOS NA REVISÃO ANFIS

As Tabelas 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52 e 53 apresentam as informações básicas dos artigos selecionados, como título, autores, ano de publicação e total de citações no *google scholar*.

Tabela 42: Informações básicas dos artigos selecionados

Referência	Título	Autores	Ano	Total de Citações
[81]	*A Deep Neural Network Combined with Radial Basis Function for Abnormality Classification*	Noushin Jafarpisheh, Effat J. Zaferani, Mohammad Teshnehlab, Hadis Karimipour, Reza M. Parizi and Gautam Srivastava	2021	8
[82]	*A distributed key authentication and OKM-ANFIS scheme based breast cancer prediction system in the IoT environment*	V. Savitha, N. Karthikeyan, S. Karthik and R. Sabitha	2021	18
[83]	*A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS*	Mohammad Ayoub Khan and Fahad Algarn	2020	272
[84]	*A Machine Learning Approach for the Differential Diagnosis of Alzheimer and Vascular Dementia Fed by MRI Selected Features*	Gloria Castellazzi; Maria Giovanna Cuzzoni; Matteo Cotta Ramusino; Daniele Martinelli; Federica Denaro; Antonio Ricciardi; Paolo Vitali; Nicoletta Anzalone; Sara Bernini; Fulvia Palesi; Elena Sinforiani; Alfredo Costa; Giuseppe Micieli; Egidio D'Angelo; Giovanni Magenes; Claudia A. M. Gandini Wheeler-Kingshott	2020	113
[85]	*A novel fuzzy expert system design to assist with peptic ulcer disease diagnosis*	Saeedreza Arab, Kianaz Rezaee and Ghazaleh Moghaddam	2021	13
[86]	*A pragmatic approach of Parkinson disease detection using hybrid case-based reasoning neuro-fuzzy classification system over Mobile edge computing*	Punarselvam, E.	2023	5

Fonte: Autora.

Tabela 43: Informações básicas dos artigos selecionados (continuação)

Referência	Título	Autores	Ano	Total de Citações
[87]	*A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique*	Mehrbakhsh Nilashi, Hossein Ahmadi, Leila Shahmoradi, Othman Ibrahim, Elnaz Akbari	2019	153
[65]	*A proposed approach for diabetes diagnosis using neuro-fuzzy technique*	MT Alasaady, TNM Aris, NM Sharef, H Hamdan	2022	12
[88]	*A Robust Cardiovascular Disease Predictor Based on Genetic Feature Selection and Ensemble Learning Classification*	PA Sadiyamole, S Manju Priya	2023	1
[89]	*A robust swarm intelligence-based feature selection model for neuro-fuzzy recognition of mild cognitive impairment from resting-state fMRI*	Anter Ahmed M., Wei Yichen, Su Jiahui, Yuan Yueming, Lei Beiying, Duan Gaoxiong, Mai Wei, Nong Xiucheng, Yu Bihan, Li Chong, Fu Zening, Zhao Lihua, Deng Demao, Zhang Zhiguo	2019	44
[90]	*Adaptive clustering based breast cancer detection with ANFIS classifier using mammographic images*	T. V. Padmavathy, M. N. Vimalkumar and D. S. Bhargava	2019	4
[91]	*Adaptive fuzzy-region growing fusion and improved CNN-ANFIS-based automated segmentation and classification of cervical cancer*	J. Jeyshri, M. Kowsigan	2023	2
[92]	*Adaptive Neuro-Fuzzy Inference Healthcare System for Cardiac Arrhythmia Detection Using Heart Rate Variability Features*	R. Lakshmi Devi, V. Kalaivani and B. L. Velammal	2021	4
[93]	*Advanced Glaucoma Detection by Integrating ANFIS with CNN Models for Enhanced Diagnosis*	Manikanda Prabu Nallasivam; Rajesh K P; Sharmila M; Sengaba Sundari D; Shivani E; Surendar R	2024	0

Fonte: Autora.

Tabela 44: Informações básicas dos artigos selecionados (continuação)

Referência	Título	Autores	Ano	Total de Citações
[94]	*An analytical review on the use of artificial intelligence and machine learning in diagnosis, prediction, and risk factor analysis of multiple sclerosis*	Shima Pilehvari, Yasser Morgan, Wei Peng	2024	2
[95]	*An ANFIS-PSO Algorithm for Predicting Four Grades of Non-Alcoholic Fatty Liver Disease*	Sakineh Zeynali Goldar; Amir Rikhtegar Ghiasi; Mohammad Ali Badamchizadeh; Manouchehr khoshbaten	2020	10
[96]	*An Efficient Machine Learning Framework for Stress Prediction via Sensor Integrated Keyboard Data*	P. B. Pankajavalli; G. S. Karthick; R. Sakthivel	2021	23
[97]	*An efficient multilayer deep detection perceptron (MLDDP) methodology for detecting testicular anomalies with or without congenital heart disease (TACHD)*	D. Kavitha and C. H. Renumadhavi	2022	3
[98]	*An Intelligent System to Enhance the Performance of Brain Tumor Diagnosis from MR Images*	Tuan-Linh Nguyen, Swathi Kavuri, Soo-Yeon Park, Minho Lee	2022	10
[99]	*An Intensive Review Analysis about Fuzzy Logic usage in Smart Medical Care System*	Vikram Kumar; Dimple Saini; Shailesh Rastogi; Ramakrishnan Raman; Apurv Verma; R. Meenaks	2024	0
[100]	*An NB-ANN based Fusion Approach for Disease Genes Prediction and LFKH-ANFIS Classifier for Eye Diseases Identification*	Saikia, Samar Jyoti; Nirmala, S R.	2021	0
[101]	*Application of the Adaptive-Neuro Fuzzy Inference System (ANFIS) Method for Rapid Diagnosis COVID-19 Patient Cases*	Esti Suryani; Handy Adi Putra; Wiharto; Umi Salamah; Wisnu Widiarto	2022	0

Fonte: Autora.

Tabela 45: Informações básicas dos artigos selecionados (continuação)

Referência	Título	Autores	Ano	Total de Citações
[102]	*Attentive Hierarchical ANFIS with interpretability for cancer diagnostic*	Tuan-Linh Nguyen, Swathi Kavuri, Soo-Yeon Park, Minho Lee	2022	22
[103]	*Automated detection of fibrillations and flutters based on fused feature set and ANFIS classifier*	Saurav Mandal, Anisha Halder Roy, Pulak Mondal	2021	4
[104]	*Automated Detection of Left Bundle Branch Block from ECG Signal Utilizing the Maximal Overlap Discrete Wavelet Transform with ANFIS*	Bassam Al-Naami, Hossam Fraihat, Hamza Abu Owida, Khalid Al-Hamad, Roberto De Fazio and Paolo Visconti	2022	17
[105]	*Balancing Accuracy and Interpretability through Neuro-Fuzzy Models for Cardiovascular Risk Assessment*	Gabriella Casalino; Giovanna Castellano; Uzay Kaymak; Gianluca Zaza	2021	22
[106]	*Bandlet transform based brain tumor detection and classification of Magnetic resonance image using Coactive Neuro Fuzzy Inference System in comparison with Adaptive Neuro Fuzzy Inference System classifier*	P. Rupeshy; M.K. Mariam Bee; V. Suresh	2022	0
[69]	*Blockchain-Based Secure Healthcare Application for Diabetic-Cardio Disease Prediction in Fog Computing*	P. G. Shynu; Varun G. Menon; R. Lakshmana Kumar; Seifedine Kadry; Yunyoung Nam	2021	116
[107]	*Brain Cancer Diagnosis and Prediction Based on Neural Gas Network and Adaptive Neuro Fuzzy*	Kamel H. Rahouma; Rabab Hamed M. Aly, Hesham F.A. Hamed	2019	9
[108]	*Brain tumor segmentation of MR images using SVM and fuzzy classifier in machine learning*	Ramdas Vankdothu, Mohd Abdul Hameed	2022	65

Fonte: Autora.

Tabela 46: Informações básicas dos artigos selecionados (continuação)

Referência	Título	Autores	Ano	Total de Citações
[109]	*Breast cancer: A hybrid method for feature selection and classification in digital mammography*	Shankar Thawkar, Vijay Katta, Ajay Raj Parashar, Law Kumar Singh, Munish Khanna	2023	7
[110]	*CAPSOCA: hybrid technique for nosologic segmentation of primary brain tumors*	Shafaf Ibrahim, Noor Elaiza Abd Khalid, Mazani Manaf	2019	6
[111]	*Cardiovascular Conditions Classification Using Adaptive Neuro-Fuzzy Inference System*	Jason Rivera; Kelsey Rodriguez; Xiao-Hua Yu	2019	6
[112]	*Classification of adaptive back propagation neural network along with fuzzy logic in chronic kidney disease*	Vineetha KR, M.S. Maharajan, Bhagyashree K, N. Sivakumar	2024	4
[113]	*Classification of brain neoplasm from multi-modality MRI with the aid of ANFIS classifier*	R. Aarthi and K. Helen Prabha	2021	3
[114]	*Classification of COVID-19 Individuals Using Adaptive Neuro-Fuzzy Inference System*	Celestine Iwendi, Kainaat Mahboob, Zarnab Khalid, Abdul Rehman Javed, Muhammad Rizwan and Uttam Ghosh	2022	97
[115]	*Classification of Thyroid Carcinoma using Sobel Edge Detection and Adaptive Neuro Fuzzy Inference System Approach*	I Intan, YJW Soetikno	2019	0
[116]	*Data Analytics on Unpredictable Pregnancy Data Records Using Ensemble Neuro-Fuzzy Techniques*	C Vairavel, NS Nithya	2023	1

Fonte: Autora.

Tabela 47: Informações básicas dos artigos selecionados (continuação)

Referência	Título	Autores	Ano	Total de Citações
[117]	*Depress-DCNF: A deep convolutional neuro-fuzzy model for detection of depression episodes using IoMT*	A Kumar, SR Sangwan, A Arora, VG Menon	2022	26
[118]	*Dermatology disease prediction based on firefly optimization of ANFIS classifier*	J Rajeshwari, M Sughasiny	2022	8
[66]	*Design of an ANFIS based Decision Support System for Diabetes Diagnosis*	L Priyadarshini; Lakshmi Shrinivasan	2020	16
[119]	*Detection and classification of electroencephalogram signals for epilepsy disease using machine learning methods*	Rajagopalan Srinath, Rajagopal Gayathri	2021	31
[120]	*Detection and segmentation of cancer regions in cervical images using fuzzy logic and adaptive neuro fuzzy inference system classification method*	Raghupathy Ramasamy, Chitra Chinnasamyo	2020	4
[121]	*Detection of brain tumor in MRI images using optimized ANFIS classifier*	Rehna Kalam, Ciza Thomas, and M. Abdul Rahiman	2021	3
[122]	*Detection of Chronic Kidney Disease Using Neuro-Fuzzy Rule-based Classifier*	Supantha Das; Arnab Hazra; Soumen Kumar Pati; Soumadip Ghosh; Saurav Mallik; Suharta Banerjee	2022	0
[123]	*Detection of mild cognitive Impairment from gait using Adaptive Neuro-Fuzzy Inference system*	Mahmoud Seifallahi a c, Afsoon Hassani Mehraban b, Fatemeh Khamseh c, Masoumeh Salehi c, Gholamreza Hajati c, Mehrdad Khansari c, Mina Ahmadzadeh c, Mehrzad Babaie	2022	6

Fonte: Autora.

Tabela 48: Informações básicas dos artigos selecionados (continuação)

Referência	Título	Autores	Ano	Total de Citações
[124]	*Developing Neural-fuzzy-based Unscented Kalman Filter Algorithm for Atrial Fibrillation Onset Prediction*	Cuc T. Bui, Phat K. Huynh, Hao T. Phan, Trung Q. Le and Vo Van Toi	2020	1
[125]	*Development and use of a clinical decision support system for the diagnosis of social anxiety disorder*	Sina Fathi, Maryam Ahmadi, Behrouz Birashk, Afsaneh Dehnad	2020	24
[126]	*Dew Computing enabled ANFIS Model for Accurate Classification of Season-Stratified Acute Respiratory Infections in High Altitude Region*	Zubair Sayeed Masoodi; Ankush Manocha	2024	0
[127]	*Dew computing-inspired health-meteorological factor analysis for early prediction of bronchial asthma*	Ankush Manocha, Munish Bhatia, Gulshan Kumar	2021	22
[128]	*Diagnosing Breast Cancer Based on the Adaptive Neuro-Fuzzy Inference System*	S. Chidambaram, S. Sankar Ganesh, Alagar Karthick, Prabhu Jayagopal, Bhuvaneswari Balachander and S. Manoharan	2022	17
[129]	*Diagnosis of liver images lesions in mr images using improved segmentation and classification task*	Babu, M.	2019	1
[130]	*Diagnosis of Mechanical Low Back PainUsing a Fuzzy Logic-Based Approach*	Nanthakumar, G.	2021	1
[131]	*Dynamic contrast enhanced-magnetic resonance imaging radiomics combined with a hybrid adaptive neuro-fuzzy inference system-particle swarm optimization approach for breast tumour classification*	Alexia G. Tzalavra, Ioannis Andreadis, Kalliopi V. Dalakleidi, Fotios Constantinidis, Evangelia I. Zacharaki, Konstantina S. Nikita	2021	8
[132]	*Dynamic neutrosophic cognitive map with improved cuckoo search algorithm (DNCM-ICSA) and ensemble classifier for rheumatoid arthritis (RA) disease*	B. Chithra, R. Nedunchezhian	2022	19

Fonte: Autora.

Tabela 49: Informações básicas dos artigos selecionados (continuação)

Referência	Título	Autores	Ano	Total de Citações
[133]	*Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods*	Osman Taylan , Abdulaziz S. Alkabaa , Hanan S. Alqabbaa, Esra Pamukçu, Víctor Leiva	2023	50
[134]	*ECG Arrhythmia Classification based on Fuzzy Cognitive Maps*	Oliverio Arellano Cárdenas, Luis Martín Flores Nava, Felipe Gómez Castañeda, José Antonio Moreno Cadenas	2019	3
[71]	*eDiaPredict: An Ensemble-based Framework for Diabetes Prediction*	Ashima Singh, Arwinder Dhillon, Neeraj Kumar, M. Shamim Hossain, Ghulam Muhammad, Manoj Kumar	2021	56
[135]	*Enhanced Pelican Optimization Algorithm with Deep Learning-Driven Mitotic Nuclei Classification on Breast Histopathology Images*	Fadwa Alrowais , Faiz Abdullah Alotaibi , Abdulkhaleq Q. A. Hassan , Radwa Marzouk , Mrim M. Alnfiai , Ahmed Sayed	2023	1
[136]	*Enhancing cervical precancerous lesion detection using African Vulture Optimization Algorithm with Deep Learning model*	Jiayu Song, Le Wang, Jiazhuo Yan, Yue Feng, Yunyan Zhang	2024	1
[67]	*Expert diagnostic system for detection of hypertension and diabetes mellitus using discrete wavelet decomposition of photoplethysmogram signal and machine learning technique*	Muzaffar khan, Bikesh Kumar Singh, Neelamshobha Nirala	2023	5
[137]	*Expert System for Diagnosis Fever with Rash Diseases for Pediatric Age Group using Adaptive Neuro-Fuzzy Inference System*	Zainab N. Al-Qudsy, Rusul M. Kanona, Shaymaa Azzam Alyawer	2022	1
[138]	*Fabrication of a portable device for stress monitoring using wearable sensors and soft computing algorithms*	Alireza Golgouneh, Bahram Tarvirdizadeh	2019	70
[139]	*Fog-Assisted Real-Time Coronary Heart Disease Risk Detection in IoT-Based Healthcare System*	L.Jubair Ahmed, Gokulavasan Brindavanam, Anishfathima Basheer	2022	7

Fonte: Autora.

Tabela 50: Informações básicas dos artigos selecionados (continuação)

Referência	Título	Autores	Ano	Total de Citações
[140]	*Fuzzy logic approach for infectious disease diagnosis: A methodical evaluation, literature and classification*	Goli Arji, Hossein Ahmadi, Mehrbakhsh Nilashi, Tarik A. Rashid, Omed Hassan Ahmed, Nahla Aljojo, Azida Zainol	2019	118
[141]	*Genomic signal processing of microarrays for cancer gene expression and identification using cluster-fuzzy adaptive networking*	Purnendu Mishra, Nilamani Bhoi	2020	2
[142]	*Heartbeat sound classification using a hybrid adaptive neuro-fuzzy inferences system (ANFIS) and artificial bee colony*	Pantea Keikhosrokiani , A. Bhanupriya Naidu A/P Anathan, Suzi Iryanti Fadila, Selvakumar Manickam2, Zuoyong Li	2023	14
[143]	*Identification of atrial fibrillation using descriptive statistic features and adaptive Neuro-Fuzzy inference system*	Maratus Solikhah, Nuryani Nuryani, Anto Satriyo Nugroho	2019	4
[144]	*iMIGS: An innovative AI based prediction system for selecting the best patient-specific glaucoma treatment*	Uvais Qidwai, Umair Qidwai, Thurka Sivapalan, Gokulan Ratnarajan	2023	3
[145]	*Improving Accuracy of Integrated Neuro-Fuzzy Classifier with FCM based Clustering for Diagnosis of Psychiatric Disorder*	Chandrabhan Singh, Dr. Mohit Gangwar, Dr. Upendra Kumar	2023	0
[68]	*Improving Diagnosis Accuracy of Diabetic Disease Using Radial Basics Function Network and Fuzzy Clustering*	Hadiseh Hosseini, Amid Khatibi Bardsiri	2019	3
[146]	*Intelligent Medical Diagnostic System for Hepatitis B*	Dalwinder Singh, Deepak Prashar, Jimmy Singla, Arfat Ahmad Khan, Mohammed Al-Sarem, Neesrin Ali Kurdi	2022	18
[147]	*Intelligent prediction of major adverse cardiovascular events (MACCE) following percutaneous coronary intervention using ANFIS-PSO model*	Sahar Dehdar Karsidani, Maryam Farhadian , Hossein Mahjub, Azadeh Mozayanimonfared	2022	5

Fonte: Autora.

Tabela 51: Informações básicas dos artigos selecionados (continuação)

Referência	Título	Autores	Ano	Total de Citações
[148]	*Jointly Composite Feature Learning and Autism Spectrum Disorder Classification Using Deep Multi-Output Takagi-Sugeno-Kang Fuzzy Inference Systems*	Zhaowu Lu, Jun Wang , Rui Mao , Minhua Lu , Jun Shi	2023	10
[149]	*Leveraging ANFIS with Adam and PSO optimizers for Parkinson's disease*	Akram Pasha, Syed Thouheed Ahmed, Ranjith Kumar Painam, Sandeep Kumar Mathivanan, Karthikeyan P, Saurav Mallik, Hong Qin	2024	4
[1]	*Liver Disease Prediction Using Machine Learning Algorithms*	R. Kalaiselvi, K. Meena, V. Vanitha	2021	18
[150]	*Lung cancer typology classification based on biochemical markers using machine learning techniques*	Jasna Nuhic, Jasmin Kevric	2020	5
[151]	*Machine Learning Framework to Predict Patient Non-Adherence to Medication using Non-Clinical Data: A Prognosis Approach*	Michael Sunday Julius, Uzoma Rita Alo, Fergus Uchenna Onu, Chinyere Ihuoma Akobundu	2021	2
[152]	*Mammogram Image Retrieval using IPSO Optimized Anfis Classifier*	Sonia Jenifer Rayen, R.Subhashini	2019	2
[153]	*Mamographic image for breast cancer detection and identification of stages of cancer using MFFC and optimized ANFIS*	M. Supriya, A. J. Deepa, C. Mythili	2020	4
[154]	*Medical Image-Based Diagnosis Using a Hybrid Adaptive Neuro-Fuzzy Inferences System (ANFIS) Optimized by GA with a Deep Network Model for Features Extraction*	Baidaa Mutasher Rashed, Nirvana Popescu	2024	6
[155]	*Multi-mass breast cancer classification based on hybrid descriptors and memetic meta-heuristic learning*	Khosro Rezaee, Adele Rezaee, Negar Shaikhi, Javad Haddadnia	2020	12

Fonte: Autora.

Tabela 52: Informações básicas dos artigos selecionados (continuação)

Referência	Título	Autores	Ano	Total de Citações
[70]	*NEIWDA with Hybrid Ensemble Model for Diagnosis of Diabetes Type-II*	R. Padmaraj, D. Suresh, Suban Ravichandran	2020	1
[156]	*Neutrosophic set with Adaptive Neuro-Fuzzy Inference System for Liver Tumor Segmentation and Classification Model*	Mohammed I. Alghamdi	2022	7
[157]	*Non-Proliferative Diabetic Retinopathy Classification Based on Hard Exudates Using Combination of FRCNN, Morphology, and ANFIS*	Ricky Eka Putra, Handayani Tjandrasa, Nanik Suciati, Ardian Yusuf Wicaksono	2020	32
[158]	*Optimization of Modified Adaptive Neuro-Fuzzy Inference System (MANFIS) with Artificial Bee Colony (ABC) Algorithm for Classification of Bone Cancer*	Ali Lefteh, Monireh Houshmand, Mahsa Khorrampanah, Ghassan Fadhil Smaisim	2022	35
[159]	*Optimizing ANFIS using simulated annealing algorithm for classification of microarray gene expression cancer data*	Bulent Haznedar, Mustafa Turan Arslan, Adem Kalinli	2021	40
[160]	*Performance Analysis of Breast Cancer Detection Method Using ANFIS Classification Approach*	K. Nagalakshmi, Dr. S. Suriya	2023	3
[161]	*Prediction of Chikungunya disease using PSO-based adaptive neuro-fuzzy inference system model*	Sandeep Kaur, Kuljit Kaur Chahal	2021	8
[162]	*Research Paper: Automated Detection of Multiple Sclerosis Lesions Using Texture-based Features and a Hybrid Classifier*	Somayeh Raiesdana	2020	1
[163]	*Rheumatoid Arthritis anticipation using Adaptive Neuro Fuzzy Inference System*	Vishu Madaan, Rupinder Kaur, Prateek Agrawal	2019	5

Fonte: Autora.

Tabela 53: Informações básicas dos artigos selecionados (continuação)

Referência	Título	Autores	Ano	Total de Citações
[164]	*Smart neuropathy detection using machine intelligence: Filling the void between clinical practice and early diagnosis*	Salahuddin T. , Al-Maadeed S.A, Petropoulos I.N. , Malik R.A, Ilyas S.K , Qidwai U	2019	7
[165]	*Spiral Optimization Algorithm with Neurocomputing and Fuzzy Logic Based Human Gait Analysis for Osteoarthritis Prediction*	Salima Baji Abdullah, Hanaa Ali Abed, Abbas Hameed Abdul Hussein, Sahar R. Abdul Kadeem, Safa Majed, Hassan M. Al-Jawahry	2023	0
[166]	*Stability of classification performance on an adaptive neuro fuzzy inference system for disease complication prediction*	Sri Kusumadewi, Linda Rosita, Elyza Gustri Wahyuni	2022	4
[167]	*Stochastic- and Neuro-Fuzzy-Analysis-based characterization and classification of 4-Channel Phonocardiograms for Cardiac Murmur Detection*	Miguel A. Becerra, Edilson DelgadoTrejos, Cristian Mejía-Arboleda, Diego H. Peluffo-Ordóñez, Ana C. Umaquinga-Criollo	2020	0
[168]	*The Accuracies of ANFIS and Genetic Algorithm with Tournament Selection on Classifying Hepatitis Data*	W FA Siahaan, O S Sitompul, Z Situmorang	2020	5
[169]	*Variational Mode Decomposition Based Retinal Area Detection and Merging of Superpixels in SLO Image*	Suchetha Manikandan, V. Deepika, Nada Philip	2021	0
[170]	*Weighted butterfly optimization algorithm with intuitionistic fuzzy Gaussian function based adaptive-neuro fuzzy inference system for covid-19 prediction*	T. Sundaravadivel, V. Mahalakshmi	2021	7
[171]	*Wheezing Feature Classification using ANFIS for Asthma Diagnosis*	Rohmah Hidayah, Ardyono Priyadi, Eko Mulyanto Yuniarno, Mauridhi Hery Purnomo	2024	0

Fonte: Autora.

B TABELAS COM OS RESULTADOS NA ÍNTEGRA PARA CADA UM DOS 210 EXAMES

As Tabelas 54, 55, 56, 57, 58, 59, 60, 61 e 62 representam os resultados obtidos na íntegra a partir das análises de todos os 210 exames, ou seja, são avaliadas as métricas de acordo com a quantidade de variáveis extraídas de forma correta em cada um dos exames.

Tabela 54: Resultados obtidos considerando todos os exames de sangue

Exame	**Total de variáveis de interesse presente no exame**	**Total de variáveis de interesse extraídas**	**Total de variáveis de interesse presentes no exame NÃO extraídas**	**Total de variáveis de interesse presente no exame extraídas**	**Total de variáveis de interesse presente no exames extraídas corretamente**	**Total de variáveis de interesse presente no exame extraídas de maneira errada**	**Total de variáveis de interesse extraídas NÃO presentes no exame**	***Precision***	***Recall***	***F1 Score***
1	18	18	0	18	15	3	0	83%	83%	83%
2	21	21	0	21	17	4	0	81%	81%	81%
3	23	23	0	23	20	3	0	87%	87%	87%
4	24	24	0	24	22	2	0	92%	92%	92%
5	21	21	0	21	20	1	0	95%	95%	95%
6	19	19	0	19	17	2	0	89%	89%	89%
7	23	21	2	21	20	1	0	95%	87%	91%
8	30	25	5	25	22	3	0	88%	73%	80%
9	26	25	1	25	23	2	0	92%	88%	90%
10	19	16	3	16	15	1	0	94%	79%	86%
11	15	15	0	15	14	1	0	93%	93%	93%
12	21	21	0	21	21	0	0	100%	100%	100%
13	24	24	0	24	24	0	0	100%	100%	100%
14	25	25	0	25	25	0	0	100%	100%	100%
15	13	14	0	13	13	0	1	100%	100%	100%
16	22	22	0	22	22	0	0	100%	100%	100%
17	15	15	0	15	15	0	0	100%	100%	100%
18	26	26	0	26	24	2	0	92%	92%	92%
19	23	23	0	23	23	0	0	100%	100%	100%
20	23	23	0	23	23	0	0	100%	100%	100%
21	22	22	0	22	21	1	0	95%	95%	95%
22	16	16	0	16	16	0	0	100%	100%	100%
23	27	27	0	27	27	0	0	100%	100%	100%
24	0	0	0	0	0	0	0			
25	24	20	4	20	17	3	0	85%	71%	77%

Fonte: Autora.

Tabela 55: Resultados obtidos considerando todos os exames de sangue (continuação)

Exame	**Total de variáveis de interesse presente no exame**	**Total de variáveis de interesse extraídas**	**Total de variáveis de interesse presentes no exame NÃO extraídas**	**Total de variáveis de interesse presente no exame extraídas**	**Total de variáveis de interesse presente no exames extraídas corretamente**	**Total de variáveis de interesse presente no exame extraídas de maneira errada**	**Total de variáveis de interesse extraídas NÃO presentes no exame**	***Precision***	***Recall***	***F1 Score***
26	25	24	1	24	24	0	0	100%	96%	98%
27	22	22	1	21	21	0	1	100%	95%	98%
28	25	24	1	24	22	2	0	92%	88%	90%
29	15	15	0	15	15	0	0	100%	100%	100%
30	16	16	0	16	13	3	0	81%	81%	81%
31	17	17	0	17	15	2	0	88%	88%	88%
32	20	19	1	19	14	5	0	74%	70%	72%
33	14	14	0	14	12	2	0	86%	86%	86%
34	14	14	0	14	11	3	0	79%	79%	79%
35	15	15	0	15	12	3	0	80%	80%	80%
36	16	16	0	16	14	2	0	88%	88%	88%
37	14	14	0	14	12	2	0	86%	86%	86%
38	21	21	0	21	21	0	0	100%	100%	100%
39	21	21	0	21	21	0	0	100%	100%	100%
40	22	21	1	21	21	0	0	100%	95%	98%
41	18	18	0	18	18	0	0	100%	100%	100%
42	17	15	2	15	14	1	0	93%	82%	88%
43	14	14	0	14	14	0	0	100%	100%	100%
44	15	15	0	15	15	0	0	100%	100%	100%
45	25	24	1	24	13	11	0	54%	52%	53%
46	15	15	0	15	14	1	0	93%	93%	93%
47	16	16	0	16	14	2	0	88%	88%	88%
48	8	8	1	7	0	7	1	0%	0%	
49	17	17	0	17	14	3	0	82%	82%	82%
50	17	17	0	17	14	3	0	82%	82%	82%

Fonte: Autora.

Tabela 56: Resultados obtidos considerando todos os exames de sangue (continuação)

Exame	Total de variáveis de interesse presente no exame	Total de variáveis de interesse extraídas	Total de variáveis de interesse presentes no exame NÃO extraídas	Total de variáveis de interesse presente no exame extraídas	Total de variáveis de interesse presente no exames extraídas corretamente	Total de variáveis de interesse presente no exame extraídas de maneira errada	Total de variáveis de interesse extraídas NÃO presentes no exame	*Precision*	*Recall*	*F1 Score*
51	21	21	0	21	14	7	0	67%	67%	67%
52	15	16	0	15	14	1	1	93%	93%	93%
53	26	25	1	25	14	11	0	56%	54%	55%
54	18	18	0	18	17	1	0	94%	94%	94%
55	25	25	0	25	22	3	0	88%	88%	88%
56	20	17	3	17	16	1	0	94%	80%	86%
57	1	1	0	1	1	0	0	100%	100%	100%
58	25	24	1	24	21	3	0	88%	84%	86%
59	19	19	0	19	18	1	0	95%	95%	95%
60	32	32	0	32	30	2	0	94%	94%	94%
61	24	24	0	24	22	2	0	92%	92%	92%
62	27	26	1	26	21	5	0	81%	78%	79%
63	16	14	2	14	14	0	0	100%	88%	93%
64	15	15	0	15	15	0	0	100%	100%	100%
65	9	12	0	9	9	0	3	100%	100%	100%
66	23	23	0	23	23	0	0	100%	100%	100%
67	24	24	0	24	22	2	0	92%	92%	92%
68	17	17	0	17	17	0	0	100%	100%	100%
69	22	22	0	22	21	1	0	95%	95%	95%
70	22	22	0	22	14	8	0	64%	64%	64%
71	30	30	1	29	27	2	1	93%	90%	92%
72	18	18	0	18	18	0	0	100%	100%	100%
73	19	19	0	19	17	2	0	89%	89%	89%
74	25	25	0	25	23	2	0	92%	92%	92%
75	1	4	0	1	1	0	3	100%	100%	100%

Fonte: Autora.

Tabela 57: Resultados obtidos considerando todos os exames de sangue (continuação)

Exame	**Total de variáveis de interesse presente no exame**	**Total de variáveis de interesse extraídas**	**Total de variáveis de interesse presentes no exame NÃO extraídas**	**Total de variáveis de interesse presente no exame extraídas**	**Total de variáveis de interesse presente no exames extraídas corretamente**	**Total de variáveis de interesse presente no exame extraídas de maneira errada**	**Total de variáveis de interesse extraídas NÃO presentes no exame**	***Precision***	***Recall***	***F1 Score***
76	0	0	0	0	0	0	0			
77	0	0	0	0	0	0	0			
78	0	0	0	0	0	0	0			
79	19	19	0	19	19	0	0	100%	100%	100%
80	0	0	0	0	0	0	0			
81	1	1	0	1	1	0	0	100%	100%	100%
82	23	23	0	23	23	0	0	100%	100%	100%
83	24	24	0	24	24	0	0	100%	100%	100%
84	24	24	0	24	24	0	0	100%	100%	100%
85	24	24	0	24	24	0	0	100%	100%	100%
86	17	15	2	15	14	1	0	93%	82%	88%
87	17	17	0	17	16	1	0	94%	94%	94%
88	34	34	0	34	29	5	0	85%	85%	85%
89	3	5	0	3	3	0	2	100%	100%	100%
90	25	24	1	24	20	4	0	83%	80%	82%
91	24	24	0	24	24	0	0	100%	100%	100%
92	26	26	0	26	25	1	0	96%	96%	96%
93	0	0	0	0	0	0	0			
94	25	25	0	25	23	2	0	92%	92%	92%
95	25	25	0	25	24	1	0	96%	96%	96%
96	25	25	0	25	23	2	0	92%	92%	92%
97	12	13	0	12	11	1	1	92%	92%	92%
98	24	24	0	24	22	2	0	92%	92%	92%
99	8	9	0	8	6	2	1	75%	75%	75%
100	27	26	1	26	24	2	0	92%	89%	91%

Fonte: Autora.

Tabela 58: Resultados obtidos considerando todos os exames de sangue (continuação)

Exame	Total de variáveis de interesse presente no exame	Total de variáveis de interesse extraídas	Total de variáveis de interesse presentes no exame NÃO extraídas	Total de variáveis de interesse presente no exame extraídas	Total de variáveis de interesse presente no exames extraídas corretamente	Total de variáveis de interesse presente no exame extraídas de maneira errada	Total de variáveis de interesse extraídas NÃO presentes no exame	*Precision*	*Recall*	*F1 Score*
101	26	24	2	24	21	3	0	88%	81%	84%
102	28	28	0	28	27	1	0	96%	96%	96%
103	25	25	0	25	25	0	0	100%	100%	100%
104	30	30	0	30	30	0	0	100%	100%	100%
105	9	9	0	9	7	2	0	78%	78%	78%
106	14	14	0	14	14	0	0	100%	100%	100%
107	17	16	1	16	13	3	0	81%	76%	79%
108	16	16	0	16	13	3	0	81%	81%	81%
109	24	24	0	24	24	0	0	100%	100%	100%
110	1	2	0	1	0	1	1	0%	0%	
111	20	16	4	16	14	2	0	88%	70%	78%
112	24	24	0	24	20	4	0	83%	83%	83%
113	25	25	0	25	24	1	0	96%	96%	96%
114	26	26	0	26	25	1	0	96%	96%	96%
115	1	1	0	1	1	0	0	100%	100%	100%
116	22	22	0	22	20	2	0	91%	91%	91%
117	24	24	0	24	23	1	0	96%	96%	96%
118	0	1	0	0	0	0	1			
119	24	24	0	24	19	5	0	79%	79%	79%
120	28	26	2	26	24	2	0	92%	86%	89%
121	21	21	0	21	17	4	0	81%	81%	81%
122	26	23	3	23	20	3	0	87%	77%	82%
123	21	21	0	21	18	3	0	86%	86%	86%
124	25	23	2	23	21	2	0	91%	84%	88%
125	1	2	0	1	1	0	1	100%	100%	100%

Fonte: Autora.

Tabela 59: Resultados obtidos considerando todos os exames de sangue (continuação)

Exame	**Total de variáveis de interesse presente no exame**	**Total de variáveis de interesse extraídas**	**Total de variáveis de interesse presentes no exame NÃO extraídas**	**Total de variáveis de interesse presente no exame extraídas**	**Total de variáveis de interesse presente no exames extraídas corretamente**	**Total de variáveis de interesse presente no exame extraídas de maneira errada**	**Total de variáveis de interesse extraídas NÃO presentes no exame**	***Precision***	***Recall***	***F1 Score***
126	29	26	3	26	25	1	0	96%	86%	91%
127	27	27	0	27	27	0	0	100%	100%	100%
128	23	23	0	23	23	0	0	100%	100%	100%
129	28	28	0	28	26	2	0	93%	93%	93%
130	17	16	1	16	16	0	0	100%	94%	97%
131	15	15	0	15	15	0	0	100%	100%	100%
132	29	29	0	29	28	1	0	97%	97%	97%
133	3	4	0	3	2	1	1	67%	67%	67%
134	4	5	0	4	3	1	1	75%	75%	75%
135	0	0	0	0	0	0	0			
136	22	21	1	21	14	7	0	67%	64%	65%
137	16	16	0	16	15	1	0	94%	94%	94%
138	9	9	1	8	5	3	1	63%	56%	59%
139	22	22	0	22	20	2	0	91%	91%	91%
140	26	26	0	26	25	1	0	96%	96%	96%
141	28	27	1	27	15	12	0	56%	54%	55%
142	31	31	0	31	30	1	0	97%	97%	97%
143	28	26	2	26	23	3	0	88%	82%	85%
144	8	7	1	7	6	1	0	86%	75%	80%
145	20	20	0	20	20	0	0	100%	100%	100%
146	32	32	0	32	29	3	0	91%	91%	91%
147	28	26	2	26	23	3	0	88%	82%	85%
148	28	28	0	28	25	3	0	89%	89%	89%
149	14	14	0	14	12	2	0	86%	86%	86%
150	26	25	1	25	19	6	0	76%	73%	75%

Fonte: Autora.

Tabela 60: Resultados obtidos considerando todos os exames de sangue (continuação)

Exame	Total de variáveis de interesse presente no exame	Total de variáveis de interesse extraídas	Total de variáveis de interesse presentes no exame NÃO extraídas	Total de variáveis de interesse presente no exame extraídas	Total de variáveis de interesse presente no exames extraídas corretamente	Total de variáveis de interesse presente no exame extraídas de maneira errada	Total de variáveis de interesse extraídas NÃO presentes no exame	*Precision*	*Recall*	*F1 Score*
151	30	29	1	29	23	6	0	79%	77%	78%
152	18	17	2	16	16	0	1	100%	89%	94%
153	30	30	0	30	29	1	0	97%	97%	97%
154	27	27	0	27	27	0	0	100%	100%	100%
155	27	27	0	27	23	4	0	85%	85%	85%
156	28	25	3	25	19	6	0	76%	68%	72%
157	7	7	0	7	7	0	0	100%	100%	100%
158	27	27	0	27	25	2	0	93%	93%	93%
159	27	27	0	27	25	2	0	93%	93%	93%
160	2	4	0	2	2	0	2	100%	100%	100%
161	18	18	0	18	16	2	0	89%	89%	89%
162	26	23	3	23	16	7	0	70%	62%	65%
163	21	21	0	21	16	5	0	76%	76%	76%
164	20	20	0	20	20	0	0	100%	100%	100%
165	30	27	3	27	25	2	0	93%	83%	88%
166	11	12	0	11	9	2	1	82%	82%	82%
167	18	18	0	18	18	0	0	100%	100%	100%
168	30	28	2	28	27	1	0	96%	90%	93%
169	17	15	2	15	15	0	0	100%	88%	94%
170	20	20	0	20	17	3	0	85%	85%	85%
171	20	20	0	20	18	2	0	90%	90%	90%
172	23	23	0	23	18	5	0	78%	78%	78%
173	20	20	0	20	16	4	0	80%	80%	80%
174	2	2	0	2	1	1	0	50%	50%	50%
175	2	2	0	2	1	1	0	50%	50%	50%

Fonte: Autora.

Tabela 61: Resultados obtidos considerando todos os exames de sangue (continuação)

Exame	**Total de variáveis de interesse presente no exame**	**Total de variáveis de interesse extraídas**	**Total de variáveis de interesse presentes no exame NÃO extraídas**	**Total de variáveis de interesse presente no exame extraídas**	**Total de variáveis de interesse presente no exames extraídas corretamente**	**Total de variáveis de interesse presente no exame extraídas de maneira errada**	**Total de variáveis de interesse extraídas NÃO presentes no exame**	***Precision***	***Recall***	***F1 Score***
176	20	20	0	20	18	2	0	90%	90%	90%
177	20	20	0	20	19	1	0	95%	95%	95%
178	24	24	0	24	22	2	0	92%	92%	92%
179	19	19	0	19	19	0	0	100%	100%	100%
180	23	23	0	23	22	1	0	96%	96%	96%
181	25	25	0	25	21	4	0	84%	84%	84%
182	31	29	2	29	26	3	0	90%	84%	87%
183	29	28	1	28	24	4	0	86%	83%	84%
184	32	32	0	32	27	5	0	84%	84%	84%
185	24	20	4	20	17	3	0	85%	71%	77%
186	0	0	0	0	0	0	0			
187	31	27	4	27	20	7	0	74%	65%	69%
188	24	24	0	24	21	3	0	88%	88%	88%
189	29	29	0	29	27	3	0	93%	93%	93%
190	29	29	0	29	25	4	0	86%	86%	86%
191	0	0	0	0	0	0	0			
192	16	16	0	16	13	3	0	81%	81%	81%
193	29	29	0	29	28	1	0	97%	97%	97%
194	2	3	0	2	0	2	1	0%	0%	
195	27	27	0	27	24	3	0	89%	89%	89%
196	27	27	0	27	21	6	0	78%	78%	78%
197	27	27	0	27	21	6	0	78%	78%	78%
198	28	28	0	28	22	6	0	79%	79%	79%
199	27	27	0	27	24	3	0	89%	89%	89%
200	18	18	0	18	13	5	0	72%	72%	72%

Fonte: Autora.

Tabela 62: Resultados obtidos considerando todos os exames de sangue (continuação)

Exame	**Total de variáveis de interesse presente no exame**	**Total de variáveis de interesse extraídas**	**Total de variáveis de interesse presentes no exame NÃO extraídas**	**Total de variáveis de interesse presente no exame extraídas**	**Total de variáveis de interesse presente no exames extraídas corretamente**	**Total de variáveis de interesse presente no exame extraídas de maneira errada**	**Total de variáveis de interesse extraídas NÃO presentes no exame**	***Precision***	***Recall***	***F1 Score***
201	0	0	0	0	0	0	0			
202	15	15	0	15	15	0	0	100%	100%	100%
203	15	15	0	15	12	3	0	80%	80%	80%
204	15	15	0	15	15	0	0	100%	100%	100%
205	15	15	0	15	15	0	0	100%	100%	100%
206	22	22	0	22	22	0	0	100%	100%	100%
207	0	0	0	0	0	0	0			
208	24	24	0	24	24	0	0	100%	100%	100%
209	7	7	0	7	6	1	0	86%	86%	86%
210	25	25	0	25	23	2	0	92%	92%	92%
MÉDIA								**89%**	**87%**	**89%**
DESVIO PADRÃO								**15%**	**16%**	**11%**

Fonte: Autora.