

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

ARTHUR SILVA LIMA
Orientadora: Profa. Dra. Valéria de Carvalho Santos

**DESENVOLVIMENTO DE UM SOFTWARE DE DIAGNÓSTICO POR
IMAGEM PARA DETECÇÃO DA DOENÇA DE CROHN E
RETICULITE USANDO APRENDIZADO DE MÁQUINA.**

Ouro Preto, MG
2024

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

ARTHUR SILVA LIMA

**DESENVOLVIMENTO DE UM SOFTWARE DE DIAGNÓSTICO POR IMAGEM PARA
DETECÇÃO DA DOENÇA DE CROHN E RETOCOLITE USANDO APRENDIZADO
DE MÁQUINA.**

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Profa. Dra. Valéria de Carvalho Santos

Ouro Preto, MG
2024



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO



FOLHA DE APROVAÇÃO

Arthur Silva Lima

Desenvolvimento de um Software de Diagnóstico por Imagem para Detecção da Doença de Crohn e Retocolite usando Aprendizado de Máquina.

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 14 de Outubro de 2024.

Membros da banca

Valéria de Carvalho Santos (Orientadora) - Doutora - Universidade Federal de Ouro Preto
Daniela Costa Terra (Examinadora) - Mestre - Instituto Federal de Minas Gerais - Ouro Branco
Leonardo Nogueira Matos (Examinador) - Doutor - Universidade Federal de Sergipe

Valéria de Carvalho Santos, Orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 14/10/2024.



Documento assinado eletronicamente por **Valeria de Carvalho Santos, PROFESSOR DE MAGISTERIO SUPERIOR**, em 15/10/2024, às 17:39, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0789780** e o código CRC **3B162617**.

Resumo

Este trabalho apresenta o desenvolvimento de um *software* baseado em aprendizado de máquina com o objetivo de auxiliar no diagnóstico diferencial entre Doença de Crohn e Retocolite Ulcerativa, utilizando imagens endoscópicas. O *software* foi implementado em *Python* e utiliza o algoritmo de Florestas Aleatórias (*Random Forest*), combinado com técnicas de pré-processamento de imagens e Análise de Componentes Principais (PCA) para reduzir a dimensionalidade dos dados. A base de dados utilizada no treinamento do modelo foi composta por imagens de colonoscopia, abrangendo casos confirmados de Crohn, Retocolite e intestinos saudáveis. O pré-processamento incluiu técnicas de rotação, espelhamento e normalização para aumentar a variabilidade e melhorar a robustez do modelo. Além disso, uma pipeline de processamento foi projetada para garantir que as imagens sejam analisadas de maneira eficiente. Os resultados obtidos mostraram uma acurácia média de 99,36% após a otimização dos hiperparâmetros do modelo. Além disso, o *software* desenvolvido inclui uma interface gráfica amigável e simples.

Palavras-chave: Aprendizado de Máquina. Inteligência Artificial. Crohn. Retocolite.

Abstract

This work presents the development of software based on machine learning aimed at assisting in the differential diagnosis between Crohn's Disease and Ulcerative Colitis using endoscopic images. The software was implemented in Python and uses the Random Forest algorithm, combined with image preprocessing techniques and Principal Component Analysis (PCA) to reduce data dimensionality. The dataset used for training the model consisted of colonoscopy images, covering confirmed cases of Crohn's Disease, Ulcerative Colitis, and healthy intestines. The preprocessing included techniques such as rotation, mirroring, and normalization to increase variability and improve model robustness. Furthermore, a processing pipeline was designed to ensure that the images are analyzed efficiently. The results showed an average accuracy of 99.36% after the optimization of the model's hyperparameters. Additionally, the developed software includes a user-friendly and simple graphical interface.

Keywords: Machine Learning. Artificial Intelligence. Crohn. Ulcerative Colitis.

Lista de Ilustrações

Figura 3.1 – Exames colonoscópicos da Doença de Crohn presentes na base de dados encontrada	15
Figura 3.2 – Exames colonoscópicos da Retocolite Ulcerativa presentes na base de dados encontrada	15
Figura 3.3 – Exames colonoscópicos de intestinos saudáveis presentes na base de dados encontrada	16
Figura 3.4 – Imagens geradas pelo algoritmo a partir do processo de rotação	16
Figura 3.5 – Interface com 100% do reconhecimento realizado	20
Figura 4.1 – Matriz de Confusão do Fold 4 de 50 componentes	24

Lista de Tabelas

Tabela 3.1 – Resumo do número de imagens por classe nos conjuntos de treinamento e validação.	18
Tabela 4.1 – Distribuição de imagens por categoria nos conjuntos de treinamento e validação	22

Lista de Abreviaturas e Siglas

ABNT	Associação Brasileira de Normas Técnicas
DECOM	Departamento de Computação
UFOP	Universidade Federal de Ouro Preto
IA	Inteligência Artificial
AM	Aprendizado de Máquina
DC	Doença de Crohn
RCU	Retocolite Ulcerativa

Sumário

1	Introdução	1
1.1	Justificativa	2
1.2	Objetivos	3
1.2.1	Objetivos Específicos	3
1.3	Organização do Trabalho	3
2	Revisão Bibliográfica	5
2.1	Fundamentação Teórica	5
2.1.1	Tratamentos dos Dados	5
2.1.1.1	Análise de Componentes Principais (PCA)	5
2.1.1.2	Balanceamento do Conjunto de Dados	6
2.1.1.3	Combinação e Conversão para Arrays	6
2.1.1.4	Redimensionamento para 2D	6
2.1.1.5	Normalização com StandardScaler	6
2.1.2	Aprendizado de Máquina	6
2.1.3	Algoritmos Clássicos de Classificação	7
2.1.3.1	kNN	7
2.1.3.2	SVM	8
2.1.3.3	Árvores de Decisão	8
2.1.3.4	Florestas Aleatórias	9
2.1.3.5	Redes Neurais Artificiais	10
2.2	Trabalhos Relacionados	10
3	Metodologia	14
3.1	Base de Dados	14
3.1.1	Fontes	14
3.1.2	Pré-Processamento	16
3.2	Desenvolvimento	18
3.2.1	Algoritmo	18
3.2.1.1	Pipeline de Processamento	18
3.2.1.2	Materiais e Métodos	19
3.2.2	Interface Gráfica	20
3.2.2.1	Materiais e Métodos	21
4	Resultados	22
4.1	Treinamento do Algoritmo	22
4.1.1	Configuração e Otimização do Algoritmo	22
4.1.2	Acurácia do Modelo	23
4.1.3	Matriz de Confusão	24

4.1.4	Análise dos Resultados	24
4.1.5	Comparação com Modelos Existentes	25
4.2	Conclusão	25
5	Considerações Finais	26
5.1	Conclusão	26
5.2	Trabalhos Futuros	26
	 Referências	 28

1 Introdução

O progresso tecnológico contínuo impacta como um catalisador na redefinição de paradigmas em múltiplos campos, com destaque para a medicina. A intersecção entre tecnologia e ciências da saúde tem proporcionado avanços substanciais, materializados no desenvolvimento de softwares cada vez mais sofisticados e eficientes para o suporte à prática médica. Estes recursos computacionais assumem um papel preponderante na contemporaneidade, contribuindo significativamente para o aprimoramento dos processos diagnósticos e terapêuticos.

Ao se tratar do âmbito clínico, algumas patologias se mostram bastante desafiadoras quanto à diferenciação e ao seu diagnóstico, como é o caso da Doença de Crohn e da Retocolite Ulcerativa. As enfermidades supracitadas possuem origem gastrointestinal e estão associadas à genética, influência ambiental, falha no sistema imunológico e/ou alteração da microbiota intestinal, como afirmado nos estudos de [Torres et al. \(2017\)](#) e de [Feuerstein e Cheifetz \(2014\)](#). Ademais, possuem poucas diferenças no que tange a análise dos exames de imagem utilizados para diagnosticá-las. Esses fatos desafiam uma boa taxa de assertividade no diagnóstico precoce dessas doenças.

A revisão conduzida por [Santolin et al. \(2023\)](#) et al. elucida as etiologias primárias e os métodos diagnósticos da Doença de Crohn (DC). Conforme a investigação dos autores, esta patologia manifesta-se como um distúrbio inflamatório gastrointestinal, caracterizado por uma resposta imunológica inadequada à microbiota intestinal. A DC é definida como um processo inflamatório crônico de etiologia idiopática, com potencial para afetar qualquer segmento do trato gastrointestinal, desde a cavidade oral até o canal anal.

A sintomatologia da DC é diversificada, manifestando-se em aproximadamente 25% dos casos. Os sintomas mais prevalentes observados são dor abdominal, diarreia, perda ponderal e fadiga intermitente. O processo diagnóstico da Doença de Crohn é complexo e frequentemente invasivo, fundamentando-se na análise integrada de resultados provenientes de exames endoscópicos, radiológicos, laboratoriais e histopatológicos. Adicionalmente, dada a natureza heterogênea da doença, a abordagem terapêutica requer individualização, devendo ser analisada de acordo com as particularidades de cada caso clínico.

[Segal \(2021\)](#) et al. apresentam uma análise abrangente da Retocolite Ulcerativa (RCU), abordando sua etiologia, métodos diagnósticos, variações na severidade e o aumento tanto na incidência quanto na prevalência da doença. A RCU, assim como a Doença de Crohn, é classificada como um distúrbio inflamatório gastrointestinal, caracterizado por períodos de exacerbação e remissão, o que pode dificultar o diagnóstico precoce. A patologia tipicamente tem início na região distal do intestino, com potencial de disseminação ao longo do trato intestinal e, em casos de maior gravidade, pode evoluir para neoplasia colônica.

O quadro clínico predominante na RCU inclui algia abdominal, diarreia sanguinolenta, fadiga e incontinência fecal. Adicionalmente, manifestações extra-intestinais podem ocorrer, como artralgia, anemia e alterações dermatológicas. O processo diagnóstico envolve procedimentos invasivos análogos aos utilizados na Doença de Crohn, compreendendo exames endoscópicos, radiológicos, laboratoriais e histopatológicos. Os autores também discorrem sobre as abordagens terapêuticas visando a rápida remissão sintomática, incluindo a farmacoterapia empregada.

A investigação conduzida por [Venito, Santos e Ferraz \(2022\)](#) destaca a urgência de implantar um mecanismo que aperfeiçoe de forma substancial o processo de diagnóstico das patologias mencionadas. Diante disso, o sistema proposto visa identificar a patologia correspondente quando apresentado a uma imagem pertinente ou, na ausência de correspondência, informar ao usuário sobre a não aplicabilidade do diagnóstico. A concepção do software baseia-se no estudo conduzido por [Silva et al. \(2020\)](#), que propõe uma aplicação análoga, porém direcionada a nefropatias, utilizando imagens histológicas renais e descrições de casos elaboradas por especialistas. O presente estudo utiliza o resultado da busca do autor pelo melhor modelo de treinamento do algoritmo desenvolvido, apesar de não utilizar a mesma divisão por grupos. [Silva et al. \(2020\)](#) divide as patologias renais em 5 grupos de diagnósticos, os quais se diferem no uso ou não de atributos extraídos de imagens. Enquanto neste estudo as imagens colonoscópicas são classificadas de acordo com a doença identificada pelo *software*.

A diferenciação entre Doença de Crohn e Retocolite Ulcerativa (RCU) é frequentemente um desafio no campo médico devido à semelhança dos sintomas clínicos e ao caráter inflamatório de ambas as patologias. Apesar das variações na distribuição e extensão das lesões ao longo do trato gastrointestinal, o diagnóstico diferencial entre essas duas doenças inflamatórias intestinais nem sempre é claro, o que pode levar a tratamentos inadequados e complicações adicionais. As técnicas diagnósticas atuais, embora avançadas, muitas vezes dependem de métodos invasivos e resultados subjetivos. Diante desse cenário, o uso de tecnologias de Aprendizado de Máquina e Processamento de Imagens surge como uma possível ferramenta para auxiliar na precisão do diagnóstico, buscando minimizar a margem de erro e facilitar a escolha do tratamento mais adequado.

1.1 Justificativa

O desenvolvimento de um software para auxiliar na diferenciação entre a Doença de Crohn e a Retocolite Ulcerativa fundamenta-se na necessidade de implementação de um instrumento que otimize o processo diagnóstico das patologias supracitadas, tendo em vista diagnósticos mais acurados. Neste contexto, a busca por uma ferramenta especializada e eficaz assume relevância crucial, visando potencializar a capacidade dos profissionais de saúde em identificar e manejar essas condições de forma mais eficiente e individualizada. A implementação deste projeto busca aprimorar o processo diagnóstico e melhorar a precisão no atendimento, o que pode contribuir

para melhores resultados clínicos e acadêmicos. Ao utilizar tecnologias já consolidadas da ciência da computação, este projeto busca enfrentar um desafio específico na área médica de forma inovadora.

1.2 Objetivos

O principal objetivo deste estudo é o desenvolvimento de um software de livre acesso, podendo ser utilizado e aprimorado a partir do código-fonte por quaisquer indivíduos, baseado em Aprendizado de Máquina e Processamento de Imagens, que diferencie e classifique, a partir da análise de imagens de exames endoscópicos inseridos como entrada, um diagnóstico entre Doença de Crohn, Retocolite ou Não Aplicável (para imagens de intestinos saudáveis ou de doenças que não se encaixem no escopo utilizado). Este desenvolvimento busca auxiliar a área médica na tomada de decisão em casos complexos, de forma a facilitar, agilizar e assegurar os resultados de diagnósticos das doenças analisadas.

1.2.1 Objetivos Específicos

Para atingir o objetivo principal, os seguintes objetivos específicos foram considerados:

- Realização do pré-processamento e aumento do conjunto de dados de imagens para treinamento e teste;
- Implementação e validação do algoritmo de aprendizado de máquina utilizando a linguagem de programação *Python*, visando obter um resultado satisfatório em termos de precisão e *recall*;
- Desenvolvimento de uma interface de usuário amigável e intuitiva para uso por parte dos profissionais da saúde, considerando princípios de design centrado no usuário e usabilidade;

1.3 Organização do Trabalho

Esta monografia está estruturada em quatro capítulos principais, organizados de forma a apresentar um desenvolvimento lógico e coerente do tema proposto. O Capítulo 1 consiste na Introdução, fornecendo um panorama geral do trabalho e estabelecendo seus objetivos. O Capítulo 2 apresenta a Revisão Bibliográfica, abrangendo a fundamentação teórica necessária e uma análise de trabalhos relacionados, contextualizando assim a pesquisa no campo atual de conhecimento. No Capítulo 3, a Metodologia é detalhada, descrevendo o material da base de dados utilizada e os métodos empregados na condução da pesquisa. Na seção 4 são discutidos os principais resultados obtidos durante o estudo realizado. Por fim, o Capítulo 5 traz as Considerações Finais, incluindo uma reflexão sobre os resultados obtidos e apresentando o cronograma de atividades

que guiou o desenvolvimento deste trabalho. Esta estrutura foi concebida para proporcionar uma compreensão clara e abrangente do projeto de pesquisa, desde seus fundamentos teóricos até suas conclusões e implicações práticas.

2 Revisão Bibliográfica

Esta revisão bibliográfica explora o estado da arte das aplicações de aprendizado de máquina no diagnóstico dessas patologias, abordando desde os fundamentos teóricos até as implementações práticas recentes, apresentando uma revisão abrangente dos conceitos fundamentais de aprendizado de máquina, os diferentes tipos de problemas abordados, as técnicas de tratamento de dados empregadas e os algoritmos clássicos de classificação analisados neste contexto.

Adicionalmente, será discutida a literatura recente sobre aplicações específicas dessas tecnologias no diagnóstico de patologias, incluindo análises comparativas de diferentes abordagens metodológicas e seus respectivos resultados. Também será dada uma atenção especial às técnicas de processamento de imagens médicas, análise de microbioma e expressão gênica.

2.1 Fundamentação Teórica

Neste capítulo será abordada a fundamentação teórica que embasa essa pesquisa. Serão apresentados os conceitos de Aprendizado de Máquina, Tipos de Problemas, Tratamento de dados e Algoritmos Clássicos de Classificação.

2.1.1 Tratamentos dos Dados

Para que um algoritmo possa aprender de forma efetiva, um dos primeiros passos para iniciar o aprendizado é a análise dos dados de entrada e, caso necessário, o tratamento deles. Esse processo visa preparar e organizar os dados antes deles serem utilizados no algoritmo. Essa preparação se baseia na correção de erros de digitação, inserção de valores ausentes, padronização dos dados para uma mesma proporção, etc. Além disso, é interessante, ao trabalhar com o aprendizado de máquina utilizando imagens, aplicar diversas alterações nos dados de entrada como redimensionamento, espelhamento e alterações de cores visando melhorar o entendimento e aumentar a possibilidade de identificação dos resultados corretos por parte do algoritmo.

2.1.1.1 Análise de Componentes Principais (PCA)

A Análise de Componentes Principais (PCA) é comumente aplicada como uma técnica de redução de dimensionalidade. Em modelos de aprendizado de máquina, trabalhar com dados de alta dimensionalidade pode ser problemático, levando à redundância de informações e aumentando o custo computacional. O PCA pode ser utilizado para transformar os dados de imagem, comprimindo a grande quantidade de variáveis em um conjunto menor de componentes principais, preservando a maior parte da variação dos dados originais.

2.1.1.2 Balanceamento do Conjunto de Dados

O balanceamento do conjunto de dados é crucial em problemas de aprendizado supervisionado, especialmente quando existem classes desiguais em termos de quantidade de amostras. No caso de diagnósticos de doenças como Crohn, Retocolite Ulcerativa e intestinos saudáveis, o balanceamento é essencial para garantir que o modelo aprenda de maneira justa e equitativa sobre cada classe, evitando que ele favoreça classes mais representadas.

2.1.1.3 Combinação e Conversão para Arrays

As imagens utilizadas nos estudos podem, inicialmente, estar em formatos que não são diretamente utilizáveis por modelos de aprendizado de máquina. Portanto, um dos primeiros passos no pré-processamento pode ser a conversão dessas imagens para arrays.

Cada imagem é transformada em um array numérico, onde cada pixel da imagem é convertido em um valor correspondente. Essa etapa se mostra crucial, visto que algoritmos de aprendizado de máquina operam diretamente sobre dados numéricos. Após essa conversão, as imagens podem ser manipuladas e processadas, facilitando a aplicação de outras técnicas, como redimensionamento e normalização.

2.1.1.4 Redimensionamento para 2D

Como os modelos de aprendizado de máquina, especialmente os algoritmos baseados em árvores, não trabalham diretamente com dados tridimensionais, faz-se necessário redimensionar as imagens para uma forma bidimensional (2D). As imagens originais com múltiplas dimensões (altura, largura e canais de cor) são achatadas em vetores unidimensionais.

Esse redimensionamento transforma cada imagem em um vetor linear, permitindo que o modelo de *Random Forest* processe os dados diretamente. Apesar de o achatamento remover a estrutura espacial das imagens, o PCA, aplicado posteriormente, ajuda a restaurar a correlação entre os pixels e a melhorar o aprendizado do modelo.

2.1.1.5 Normalização com StandardScaler

Após a conversão das imagens em arrays numéricos, pode-se aplicar a normalização dos dados usando o método `StandardScaler`. A normalização é uma técnica essencial em aprendizado de máquina, pois garante que os dados estejam dentro de uma mesma escala, com média zero e variância unitária. Isso é especialmente importante quando se trabalha com algoritmos que dependem da distância ou da magnitude dos dados, como o *Random Forest*.

2.1.2 Aprendizado de Máquina

Conforme afirmado por Janiesch, Zschech e Heinrich (2021), Aprendizado de Máquina (AM), ou Machine Learning, é uma área central da Inteligência Artificial que capacita sistemas

a melhorarem seu desempenho, aprenderem com os dados e tomarem decisões baseadas em métodos estatísticos. A essência do AM reside na capacidade das máquinas de identificar padrões em grandes volumes de dados, permitindo análises preditivas a partir de dados de treinamento. Esse aprendizado contínuo e adaptativo possibilita a utilização de algoritmos para tarefas como classificação de documentos, análise de risco e previsão de comportamento de clientes, por exemplo. Um ponto chave no uso de AM é a seleção do algoritmo adequado para cada contexto ou tipo de dados analisados, uma vez que diferentes abordagens de aprendizado são mais eficazes em situações específicas.

O Aprendizado de Máquina Supervisionado tem como principal característica o uso de um conjunto de dados para treinamento com entradas e saídas corretas de forma que o algoritmo aprenda com o passar do tempo. Esse tipo de aprendizado pode ser subdividido em dois tipos de problemas: Classificação e Regressão.

No Problema de Classificação, busca-se prever o grupo de uma observação. O objetivo é criar um classificador que retorne a característica de um dado a partir dos dados de entrada. Esse problema costuma usar como base os dados não observados para classificar a entrada de acordo com um grupo.

Enquanto isso, o Problema de Regressão visa utilizar os dados de entrada de forma a prever uma resposta. Diferentemente da Classificação, a Regressão retorna um valor numérico no lugar de uma classe do dado.

O Aprendizado de Máquina Não-Supervisionado se difere do Supervisionado quanto aos dados utilizados para o treinamento. Esse tipo de aprendizado utiliza dados não rotulados, fazendo com que o algoritmo encontre padrões para resolver e agrupar os resultados. Este abrange uma subárea chamada de Problema, assim como no Aprendizado Supervisionado.

O Problema de Agrupamento, ou Clustering, busca dividir os dados em grupos, assim como a Classificação. A diferença entre esses problemas é que o Agrupamento não rotula os grupos, não havendo corretude entre um grupo e outro.

2.1.3 Algoritmos Clássicos de Classificação

No desenvolvimento do projeto proposto será utilizado um algoritmo de aprendizado de máquina supervisionado baseado em classificação. Dessa forma, são apresentados, abaixo, os principais algoritmos utilizados nessa área.

2.1.3.1 kNN

O kNN (ou k - *Nearest Neighbors*, em português "k - Vizinhos Mais Próximos") é um método que baseia a sua classificação dos dados na distância entre os k pontos pré-definidos na fase de treinamento. Os dados são compreendidos em um gráfico de uma ou mais dimensões, as quais são baseadas nos atributos analisados, e a distância do dado analisado aos k vizinhos mais

próximos é aferida. A maior quantidade de vizinhos de certo grupo define o valor desse dado. [Larose e Larose \(2014\)](#) explicam de forma bastante detalhada este algoritmo.

2.1.3.2 SVM

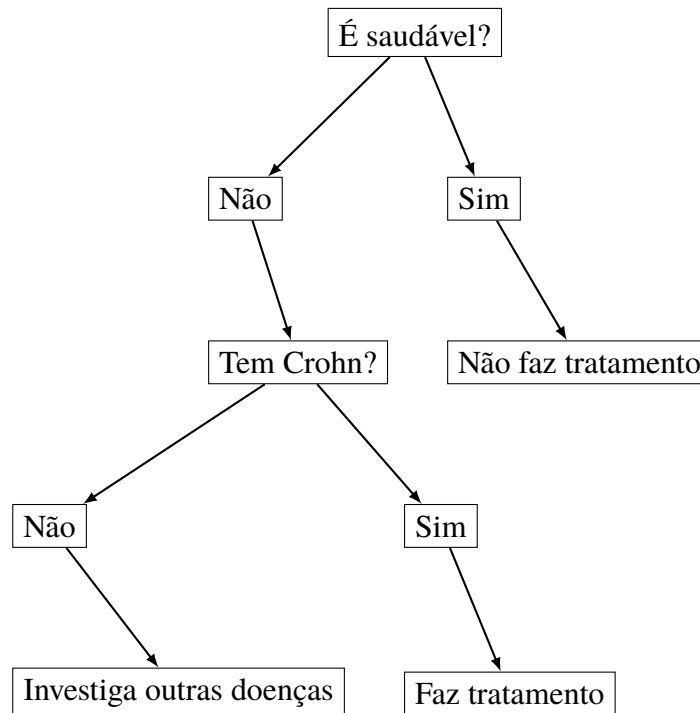
O algoritmo SVM (ou *Support Vector Machine*, em português "Máquina de Vetores de Suporte"), como explicado por [Lorena e Carvalho \(2007\)](#), classifica os dados para duas ou mais classes, sendo majoritariamente utilizado em problemas que possuem dois grupos. A ideia principal do SVM é traçar uma linha que tem como função dividir o conjunto de dados e classificar cada lado de uma forma. Os pontos mais próximos da linha traçada são chamados Vetores de Suporte. A partir dos Vetores de Suporte, o algoritmo tenta maximizar a margem entre eles e a fronteira criada previamente.

2.1.3.3 Árvores de Decisão

A Árvore de Decisão é um dos algoritmos mais conhecidos e mais comuns de Aprendizado de Máquina. Ela é um modelo preditivo que divide iterativamente um conjunto de dados em subconjuntos menores de acordo com atributos relevantes, buscando classificar e prever um resultado. Esse algoritmo é comumente utilizado para tarefas de classificação e/ou regressão, sendo aplicável em várias áreas de conhecimento, como medicina e mercado financeiro. A árvore de decisão é composta por nós, que representam as decisões a serem tomadas, e folhas, que indicam os resultados ou as classes finais retornadas. Geralmente as condicionais utilizadas nos nós da árvore trabalham sobre as características mais relevantes a serem consideradas para diferenciar os dados, fazendo com que as resultantes possuam resultados mais refinados e permitindo uma maior eficiência do modelo.

A partir da estrutura de uma árvore é possível observar de forma clara o processo de decisão do algoritmo permitindo, inclusive, a criação das chamadas *Random Forests* (ou Florestas Aleatórias). [Moret \(1982\)](#) realiza uma descrição aprofundada do algoritmo.

A árvore abaixo busca realizar a análise do estado de saúde de um paciente, determinando se ele está saudável ou deve realizar um tratamento. Inicialmente, é feito o questionamento se o paciente "É saudável?", seguido pela análise de outras características relevantes para o diagnóstico. Se a resposta for negativa, o próximo questionamento é sobre a presença de Doença de Crohn no paciente. De forma bem semelhante uma árvore de decisão busca identificar as características mais significativas observadas durante o treinamento para tomar uma decisão.



2.1.3.4 Florestas Aleatórias

O algoritmo das Florestas Aleatórias, também nomeado *Random Forests*, é chamado dessa maneira por ser formado por múltiplas árvores de decisão, responsáveis por orientar o processo de aprendizado. Esse algoritmo utiliza uma combinação de *Bootstrap Aggregating* e *Random Subspace Method* para gerar múltiplas árvores de decisão, conforme apresentado por Ren, Cheng e Han (2017).

Os autores aprofundam sobre os conceitos dessas duas técnicas, indicando que o *Bootstrap Aggregating* tem seu foco na amostragem estatística que consiste na criação de múltiplas amostras do conjunto de dados original a partir de um processo de extração de dados com reposição. Cada amostra gerada possui o mesmo tamanho do conjunto de dados original, porém, devido à natureza da amostragem com reposição, alguns registros podem aparecer múltiplas vezes, enquanto outros podem não ser selecionados.

Além disso, o *Random Subspace Method* é uma técnica que opera no espaço das características do conjunto de dados. Para cada árvore de decisão a ser construída, este método seleciona aleatoriamente um subconjunto das variáveis disponíveis.

Seguindo os conceitos abordados, podemos analisar, paralelamente, o funcionamento das *Random Forests*. Cada árvore é construída a partir de uma amostra aleatória do conjunto de treinamento e uma seleção aleatória das características dos dados, garantindo que as árvores sejam independentes entre si. O resultado final da classificação é obtido através de uma votação entre as árvores, o que contribui para a redução de overfitting e melhora a generalização do modelo, como afirmam Ren, Cheng e Han (2017).

Uma das vantagens das florestas aleatórias é sua capacidade de lidar com grandes volumes de dados, características que fazem desse algoritmo uma excelente escolha para o desenvolvimento de sistemas de diagnóstico médico. No contexto do diagnóstico da Doença de Crohn e Retocolite, a utilização de florestas aleatórias permite o processamento eficiente de dados complexos e ruidosos, características comuns em imagens médicas. Além disso, as florestas aleatórias são particularmente eficazes na seleção de características relevantes, eliminando informações redundantes ou irrelevantes, o que é crucial para garantir a precisão do diagnóstico automatizado. A implementação de uma abordagem adaptativa, como a descrita por [Ren, Cheng e Han \(2017\)](#), melhora ainda mais a precisão e a performance do algoritmo ao ajustar automaticamente os parâmetros e otimizar o processo de seleção de características.

2.1.3.5 Redes Neurais Artificiais

As redes neurais artificiais (RNNs) são sistemas computacionais amplamente utilizados em diversas áreas devido à sua capacidade de aprender padrões complexos a partir de grandes volumes de dados. Inspiradas na estrutura e no funcionamento do cérebro humano, as RNNs são compostas por camadas de neurônios artificiais interconectados que processam dados de entrada, passando por camadas ocultas até gerar uma saída. Essa estrutura permite que as RNNs realizem tarefas como classificação de dados, previsão e reconhecimento de padrões conforme apresentado por [Dastres e Soori \(2021\)](#). No contexto do diagnóstico por imagem, as RNNs são uma escolha promissora, pois podem identificar características relevantes em imagens médicas complexas, ajudando na detecção de doenças como a Doença de Crohn e Retocolite. Além disso, as RNNs apresentam uma capacidade adaptativa que permite ao sistema se ajustar automaticamente a novos padrões de dados, o que é fundamental para lidar com a variabilidade presente em imagens de pacientes. No entanto, é importante reconhecer que, apesar de suas vantagens, as RNNs exigem grandes volumes de dados de treinamento e podem ser sensíveis a problemas de *overfitting*, o que deve ser considerado no desenvolvimento de modelos precisos e confiáveis.

2.2 Trabalhos Relacionados

Nesta seção são apresentados os estudos utilizados como referências para o desenvolvimento deste trabalho.

A análise automática de imagens tem se mostrado um avanço significativo na área da Gastroenterologia nos últimos anos. De acordo com [Kröner et al. \(2021\)](#), houve um progresso substancial no uso de Inteligências Artificiais neste campo nas últimas duas décadas, especialmente devido à sua dependência do estudo de imagens. Esta evolução tecnológica tem demonstrado resultados promissores em diversas aplicações clínicas, como evidenciado por [Stafford et al. \(2022\)](#) em sua revisão sistemática sobre o uso de inteligência artificial e aprendizado de máquina no contexto das doenças inflamatórias intestinais. Os autores identificaram que 28% dos estudos

focaram na previsão do curso da doença, 23% no diagnóstico e 21% na avaliação da gravidade, demonstrando a versatilidade dessas ferramentas. Complementarmente, [Fiaidhi, Mohammed e Zezos \(2022\)](#) exploraram a aplicação de técnicas avançadas como Deep Learning, Redes Neurais Convolucionais (CNN) e Random Forest para melhorar a precisão e assertividade do diagnóstico da Doença de Crohn utilizando imagens de colonoscopia em vídeo. Apesar dos avanços significativos, [Stafford et al. \(2022\)](#) apontam que apenas 5% dos estudos utilizam um conjunto de dados externo para validação, indicando a necessidade de superar desafios como o viés de dados e a falta de generalização dos modelos para uma aplicação clínica mais ampla. No entanto, o cenário geral é otimista, com [Kröner et al. \(2021\)](#) destacando a importância da colaboração multidisciplinar entre gastroenterologistas, indústria e instituições reguladoras no desenvolvimento e aplicação dessas novas tecnologias no ambiente clínico.

O trabalho desenvolvido e escrito por [Kröner et al. \(2021\)](#) realiza uma revisão do estado da arte sobre o uso de IA na área da gastroenterologia introduzindo o tema e suas aplicações clínicas, delineando as limitações atuais e lacunas de conhecimento e as direções futuras neste campo. Os autores relatam que a área, a qual depende substancialmente do estudo de imagens, teve um grande progresso nos últimos 20 anos quando se trata do uso de Inteligências Artificiais. Este estudo também traz uma visão otimista quanto à colaboração multidisciplinar próxima entre gastroenterologistas, indústria e instituições reguladoras no desenvolvimento e aplicação de novas tecnologias no ambiente clínico.

Na obra de [Silva et al. \(2020\)](#) há o estudo e o desenvolvimento de algoritmos baseados em Aprendizado de Máquina para a classificação do diagnóstico de patologias renais. Os autores buscam utilizar a tecnologia para auxiliar os profissionais nefrologistas na análise dos exames por imagem e, a partir de testes preliminares, escolheram o algoritmo Floresta Aleatória para prosseguir no desenvolvimento da aplicação, visto que ele apresentou um melhor desempenho dentre os classificadores analisados. O *software* criado classifica as doenças em 5 grupos diferentes de acordo com a detecção, atingindo cerca de 75% de acurácia nos resultados obtidos.

A revisão sistemática de [Stafford et al. \(2022\)](#) aborda o uso de inteligência artificial e aprendizado de máquina nas doenças inflamatórias intestinais, incluindo Doença de Crohn e Colite Ulcerativa, com potencial para melhorar o diagnóstico, prever o curso da doença e personalizar o tratamento. Além de avaliar as aplicações de IA e aprendizado de máquina na gestão da Doença Inflamatória Intestinal (IBD), com foco em identificar sua utilidade atual e fornecer diretrizes práticas. A pesquisa foi conduzida nas bases de dados MEDLINE e Embase até maio de 2021, utilizando os termos "machine learning" e "artificial intelligence" combinados com "Crohn's Disease", "Ulcerative Colitis" e "Inflammatory Bowel Disease". Foram incluídos estudos com dados de pacientes humanos, revisados por pares, excluindo artigos anteriores a 2001 ou que não estivessem em inglês. Dos 409 estudos iniciais, 78 foram selecionados, sendo que 47% utilizaram métodos de *Random Forests* e houve um aumento no uso de redes neurais, especialmente em conjuntos de dados de imagens, como colonoscopias. As principais aplicações clínicas incluíram

diagnóstico (23%), previsão do curso da doença (28%) e gravidade (21%). Embora os dados de microbioma tenham ganhado espaço, os dados clínicos ainda predominam (41%). Apenas 5% dos estudos utilizaram um conjunto de dados externo para validação, ressaltando a necessidade de melhorar os pipelines de AM para maior generalização dos modelos. A revisão destaca o aumento significativo no uso dessas tecnologias em doenças inflamatórias intestinais nos últimos anos, mas aponta desafios como viés de dados e falta de generalização, que precisam ser superados para uma aplicação clínica mais ampla.

[Fiaidhi, Mohammed e Zazos \(2022\)](#) buscam explorar a variabilidade das lesões da Doença de Crohn utilizando imagens de colonoscopia em vídeo, além de aplicar técnicas de aprendizado de máquina para melhorar a precisão e assertividade do diagnóstico. Deep Learning, Convolutional Neural Networks (CNN) e Random Forest são os métodos destacados nesta investigação, além do uso de técnicas de pré-processamento de imagens como normalização, extração de características e aumento de dados. Para a validação dos dados foram utilizadas as métricas de acurácia, especificidade e sensibilidade. A base de dados do estudo foi criada pelos autores, formada por vídeos de colonoscopia coletados.

Além do estudo anterior, [Fiaidhi, Mohammed e Zazos \(2023\)](#) também trabalharam em um estudo acerca da investigação sobre a categorização de segmentos da Doença de Crohn usando vídeos de colonoscopia por cápsula a partir da segmentação de vídeo, extração de características e aumento dos dados utilizados. Assim como no estudo anterior, a base de dados citada foi criada pelos autores, contendo e vídeos de colonoscopia por cápsula. Deep Learning e Data Augmentation Techniques são métodos utilizados na pesquisa além da realização de validação cruzada sobre o conjunto de dados.

No estudo conduzido por [Maurício e Domingues \(2024\)](#), os autores desenvolveram uma ferramenta baseada em modelos de deep learning (CNNs e ViTs) para auxiliar médicos a distinguir entre DC e UC a partir de imagens endoscópicas. Os modelos foram treinados em dados coletados de três bases públicas de imagens, com foco na construção de arquiteturas mais leves e eficientes para aplicação clínica por meio da técnica de destilação de conhecimento. Os resultados indicaram uma redução de 25 vezes no número de parâmetros dos modelos, mantendo um alto desempenho e reduzindo o tempo de inferência em até 5,32 segundos. O uso de interpretabilidade dos modelos foi essencial para garantir a identificação correta de úlceras e lesões inflamatórias.

O estudo realizado por [Kim et al. \(2023\)](#) focou no uso da análise do microbioma fecal para o diagnóstico de DC e UC. Utilizando a técnica de análise discriminante por mínimos quadrados parciais esparsos (sPLS-DA), os autores desenvolveram modelos binários capazes de distinguir entre indivíduos saudáveis e pacientes com DII, e entre DC e UC. O estudo usou dados de 1846 amostras fecais e conseguiu obter uma acurácia de 94,5% na diferenciação entre CD e UC. O trabalho destaca a eficácia da análise do microbioma como uma ferramenta complementar no diagnóstico das doenças inflamatórias intestinais.

[Park et al. \(2021\)](#) investigaram o uso de dados de sequenciamento de RNA (RNA-seq)

para diferenciar DC e UC. Utilizando amostras de biópsias endoscópicas, os autores empregaram modelos de aprendizado supervisionado e não supervisionado para identificar padrões genéticos distintos entre as doenças. O estudo conseguiu uma taxa de erro global de 14,7%, mostrando que a análise de expressão gênica pode ser útil no diagnóstico diferencial entre DC e UC. Este modelo foi baseado na seleção de genes específicos que são mais expressos em lesões inflamatórias, contribuindo para o desenvolvimento de diagnósticos personalizados.

Diversos estudos têm explorado o uso de aprendizado de máquina no diagnóstico de doenças gastrointestinais, inclusive na diferenciação entre Doença de Crohn e Retocolite Ulcerativa. Embora a quantidade de pesquisas ainda seja limitada, os avanços recentes em técnicas de *deep learning* têm mostrado resultados promissores. Por exemplo, um estudo recente utilizou sequenciamento de RNA em amostras de biópsia endoscópica, aplicando modelos supervisionados de aprendizado de máquina para distinguir entre as duas doenças com alta precisão, atingindo uma taxa de erro de apenas 14,7% [Park et al. \(2021\)](#). Outros trabalhos destacam o potencial da análise de microbiomas fecais através de aprendizado de máquina, com modelos que alcançam até 95% de acurácia ao diferenciar pacientes com Doença de Crohn, Retocolite Ulcerativa e controles saudáveis [Kim et al. \(2023\)](#). De acordo com [Maurício e Domingues \(2024\)](#), esses avanços na tecnologia têm impulsionado o desenvolvimento de novos algoritmos capazes de melhorar a precisão dos diagnósticos endoscópicos e reduzir o tempo necessário para o diagnóstico.

Por fim, o estudo de [Rother et al. \(2023\)](#) explorou o uso de aprendizado profundo para a detecção automática de DC e UC a partir de imagens endoscópicas. O trabalho apresentou resultados promissores na automação do diagnóstico, reduzindo significativamente a subjetividade do diagnóstico clínico. O uso de redes neurais convolucionais permitiu uma análise detalhada das imagens, identificando características típicas de ambas as condições e oferecendo uma ferramenta inovadora para gastroenterologistas.

3 Metodologia

A presente seção tem como objetivo tratar sobre a metodologia aplicada ao desenvolvimento do software baseado em aprendizado de máquina. Serão apresentadas mais informações sobre a base de dados utilizada, o tratamento desses dados e o desenvolvimento do algoritmo bem como o resultado esperado após sua finalização.

3.1 Base de Dados

A base de dados que compõe o presente artigo tem como origem os estudos de [Polat et al. \(2022\)](#) et al e as imagens disponibilizadas online em [Library \(s.d.\)](#). A princípio o objetivo do *software* era utilizar imagens de radiografia como aprendizado, porém não foram encontradas fontes desses exames para a Retocolite Ulcerativa. Dessa forma, optou-se pelo uso de imagens de colonoscopia.

Acesse a base de dados no Google Drive utilizando um e-mail institucional da Universidade Federal de Ouro Preto.¹

3.1.1 Fontes

A busca por dados relacionados à Doença de Crohn para o aprendizado da máquina foi bastante complexa, visto que há poucas fontes com imagens de exames de colonoscopia com o diagnóstico confirmado da doença. Além disso, as fontes encontradas possuem poucas imagens se comparadas à Retocolite e, em sua maioria, não são gratuitas. Durante a pesquisa foram encontradas 66 imagens de radiografia sem custo no estudo de [Magalhães et al. \(2023\)](#) e 47 imagens e 24 vídeos de colonoscopias disponibilizadas em [Library \(s.d.\)](#), sendo composta, esta, por mídias pagas em sua maioria. Ainda assim, de acordo com a fonte desses dados, a compra dos direitos de uso dessas imagens não se faz necessária caso seja com a finalidade de treinar um algoritmo de ML.

[Magalhães et al. \(2023\)](#) realizaram um estudo acerca da Doença de Crohn, coletando 66 imagens de radiografia relacionadas à doença, identificando e indicando os sinais mais importantes quanto a ela: a atividade da doença, que indica a forma como ela age no organismo da pessoa afetada; o grau de penetração, o qual trata do nível de avanço da doença no corpo do paciente; e as complicações, as quais são os efeitos gerados por ela. Este conjunto de imagens, disponível para acesso a partir do estudo, possui variações do mesmo exame para diferentes tamanhos e resoluções, além de, em alguns casos, apresentar figuras que contém mais de um exame radiológico no mesmo arquivo.

¹ <<https://drive.google.com/drive/folders/1ccmyxoZRGvpwWtRFA1tOaRmrCSaCF4Rv>>

Além disso, as 47 imagens disponibilizadas em [Library \(s.d.\)](#) não acompanham um estudo sobre a doença, apenas possuindo variações relacionadas à gravidade da Doença de Crohn, sendo algumas com sintomas mais aparentes e outras com poucos detalhes.

No presente texto foram utilizadas imagens de colonoscopia disponíveis no estudo de [Carlos \(2020\)](#) para representar os dados referentes à Doença de Crohn. O autor apresenta a análise de casos clínicos relacionados à patologia estudada, trazendo imagens e discussões acerca da situação do paciente.

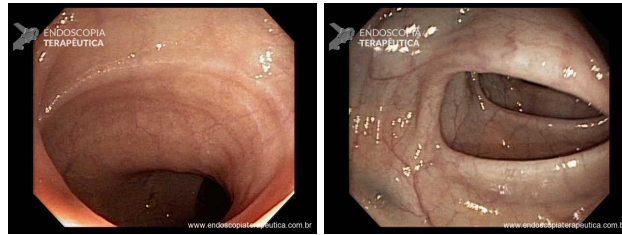


Figura 3.1 – Exames colonoscópicos da Doença de Crohn presentes na base de dados encontrada

Quanto à Retocolite não houve problemas relacionados à quantidade de dados. [Polat et al. \(2022\)](#) et al. coletaram e trataram 11276 imagens de exames da doença. Esses exames são divididos de acordo com o Escore de Mayo possuindo, majoritariamente (54,14%), imagens de nível 0 nesta escala, as quais não serão utilizadas para o aprendizado da máquina, visto que não possuem características visuais que permitam identificar o surgimento da doença no paciente. Outro ponto a ser considerado se refere às imagens de Escore de Mayo nível 3 que, geralmente, representam uma doença já em estado mais avançado, sendo identificável facilmente pelo profissional, o que torna menor a necessidade do uso de imagens de nível 3 para o aprendizado da máquina. Assim, consideraremos como principal foco de estudo os arquivos de mídia que apresentam Índice de Mayo 1 e 2, os quais representam 38,82% da base de dados desenvolvida pelos autores. Essas imagens são disponibilizadas no formato BMP, as quais devem ser convertidas para JPG, o que pode diminuir um pouco a qualidade gráfica.

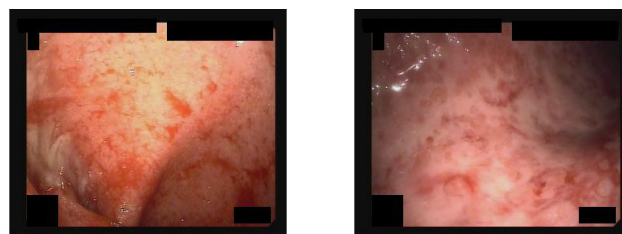


Figura 3.2 – Exames colonoscópicos da Retocolite Ulcerativa presentes na base de dados encontrada

Por fim, para o treinamento do modelo em relação ao diagnóstico de um intestino saudável ou livre de Doença de Crohn e Retocolite Ulcerativa, foi utilizado o Nerthus Dataset de [Huissteden \(2023\)](#), uma base de dados especializada disponível na plataforma *online* Kaggle que contém

vídeos do trato gastrointestinal focados especificamente na avaliação da qualidade da limpeza intestinal para colonoscopia. Este conjunto de dados compreende 21 vídeos, totalizando 5.525 quadros, todos classificados de acordo com a Boston Bowel Preparation Scale (BBPS) conforme explicado por [Lai et al. \(2009\)](#). A BBPS é um sistema de pontuação que avalia a qualidade da preparação do cólon, atribuindo pontuações de 0 a 3, onde 0 indica visualização inadequada devido à presença de fezes sólidas, enquanto 3 representa uma mucosa totalmente visível sem resíduos fecais de acordo com o estudo de [Calderwood e Jacobson \(2010\)](#). Para este estudo, foram selecionados especificamente os vídeos classificados com grau 3 na escala BBPS, garantindo assim a análise de imagens com a melhor qualidade de preparação intestinal possível.

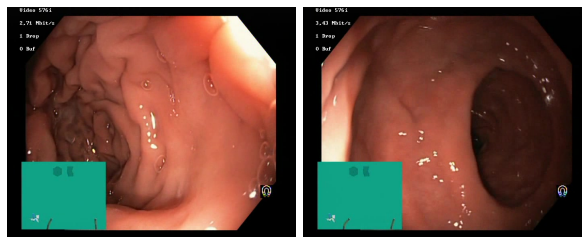


Figura 3.3 – Exames colonoscópicos de intestinos saudáveis presentes na base de dados encontrada

3.1.2 Pré-Processamento

Buscando melhorar a resolução dos dados, diminuir a discrepância entre a quantidade de imagens de Retocolite em relação à Doença de Crohn e padronizar as dimensões das imagens, faz-se necessário o pré-processamento dos dados.

Esse passo tem grande importância no aprendizado de máquina, uma vez que as alterações realizadas nos dados buscam desafiar a IA a identificar as doenças ainda que os exames tenham algum índice de rotação e/ou espelhamento. Assim, considerando a base de dados [Library \(s.d.\)](#) de 47 imagens, realizar essas alterações torna capaz a criação de 1692 imagens considerando um ângulo de rotação de 10° para cada operação.

A fim de se rotacionar as imagens de forma rápida e converter imagens BMP para JPG de maneira mais eficiente, foram desenvolvidos dois algoritmos na linguagem de programação *Python* utilizando a biblioteca *Pillow*.

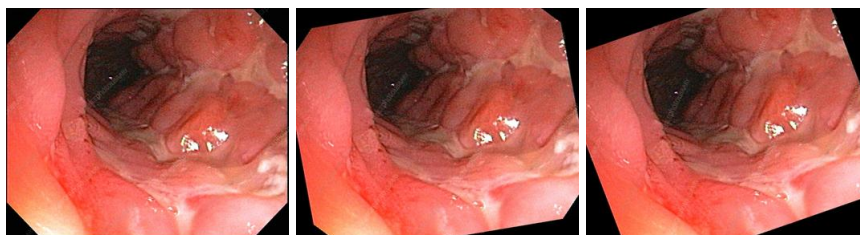


Figura 3.4 – Imagens geradas pelo algoritmo a partir do processo de rotação

Durante esta etapa também foi utilizada a técnica de *Análise de Componentes Principais* (PCA) para a redução de dimensionalidade, motivada principalmente pelos seguintes fatores:

1. **Eficiência Computacional:** O conjunto de dados utilizado neste estudo consiste em imagens que, quando convertidas para arrays numéricos, geram uma grande quantidade de variáveis (características). Utilizar todas essas variáveis pode tornar o treinamento do modelo de aprendizado de máquina extremamente lento e ineficiente. A redução de dimensionalidade com PCA permite reter apenas as características mais importantes, reduzindo o custo computacional e melhorando o desempenho do modelo.
2. **Redução de Ruído:** Ao aplicar o PCA, as informações redundantes ou irrelevantes que possam estar presentes nos dados são eliminadas. Isso permite que o modelo se concentre nas características mais relevantes entre as classes estudadas.
3. **Prevenção de Overfitting:** A redução do número de variáveis contribui para simplificar o modelo e evitar o fenômeno de *overfitting*, que ocorre quando o modelo se ajusta excessivamente aos dados de treino, perdendo a capacidade de generalização para novos dados.
4. **Melhor Desempenho:** Com base nos resultados obtidos, o uso de 200 componentes principais foi selecionado após experimentação, proporcionando uma acurácia média de 99,36% e uma baixa variância entre as diferentes dobras da validação cruzada. Esse número de componentes foi suficiente para capturar a maior parte da variação presente nos dados originais, sem sobrecarregar o modelo com variáveis irrelevantes.

Dessa forma, a aplicação de PCA para a redução de dimensionalidade foi fundamental para garantir a eficiência e precisão do modelo de *Random Forest*, resultando em uma classificação eficaz para o diagnóstico diferencial entre Doença de Crohn e Retocolite Ulcerativa.

No experimento, após testar diferentes números de componentes, o melhor desempenho foi alcançado com 200 componentes principais. Esse número foi escolhido porque preservou informações suficientes para o aprendizado sem sobrecarregar o modelo com variáveis irrelevantes. O PCA foi, portanto, fundamental para a melhora da acurácia e a redução do tempo de processamento, uma vez que removeu correlações lineares entre as variáveis, mantendo a estrutura essencial para o diagnóstico.

Além disso, o conjunto de dados foi cuidadosamente balanceado. Foram utilizadas 558 imagens de Doença de Crohn, 477 de Retocolite Ulcerativa e 522 de Healthy para treinamento, e 439 imagens para cada classe no conjunto de validação. Esse balanceamento foi fundamental para garantir que o modelo não desenvolvesse vieses, assegurando que todas as classes fossem representadas adequadamente durante o treinamento e a validação.

Classe	Treinamento	Validação
Doença de Crohn	558	439
Retocolite Ulcerativa	477	439
Saudável	522	439

Tabela 3.1 – Resumo do número de imagens por classe nos conjuntos de treinamento e validação.

Por fim, ao aplicar o `StandardScaler`, todos os pixels das imagens foram normalizados, o que garante que as variações de brilho ou contraste entre as imagens não afetem a performance do modelo. Com essa etapa, o modelo consegue aprender de forma mais eficiente, evitando que variáveis com magnitudes maiores dominem as outras durante o processo de treinamento.

3.2 Desenvolvimento

Esta seção busca falar sobre a implementação prática da metodologia proposta, desde a criação e otimização do algoritmo de aprendizado de máquina destinado à detecção das doenças. O algoritmo que apresentou melhores resultados nos estudos analisados foi a árvore de decisão, sendo este o método a ser utilizado no desenvolvimento do projeto. Essa decisão também leva em consideração a facilidade de interpretação e adaptação das árvores de decisão a conjuntos de dados heterogêneos.

3.2.1 Algoritmo

A escolha do algoritmo de Aprendizado de Máquina é uma etapa crítica que impacta diretamente no desempenho e na eficácia do sistema de diagnóstico. Após uma análise criteriosa de diferentes abordagens, optou-se pelo algoritmo de Florestas Aleatórias (*Random Forests*), que é baseado em múltiplas Árvores de Decisão. Esta escolha foi fundamentada em diversos fatores:

1. Robustez a *outliers* e ruído: Característica essencial ao lidar com imagens médicas que podem apresentar variações significativas.
2. Capacidade de lidar com dados de alta dimensionalidade: Crucial para o processamento de imagens, onde cada pixel representa uma dimensão.
3. Facilidade de interpretação: A estrutura baseada em árvores permite compreender o processo decisório do algoritmo.
4. Bom desempenho com conjuntos de dados de tamanho moderado: Adequado para a quantidade de imagens disponíveis no estudo.

3.2.1.1 Pipeline de Processamento

Foi implementada uma pipeline completa de processamento, consistindo em:

1. Pré-processamento das imagens:

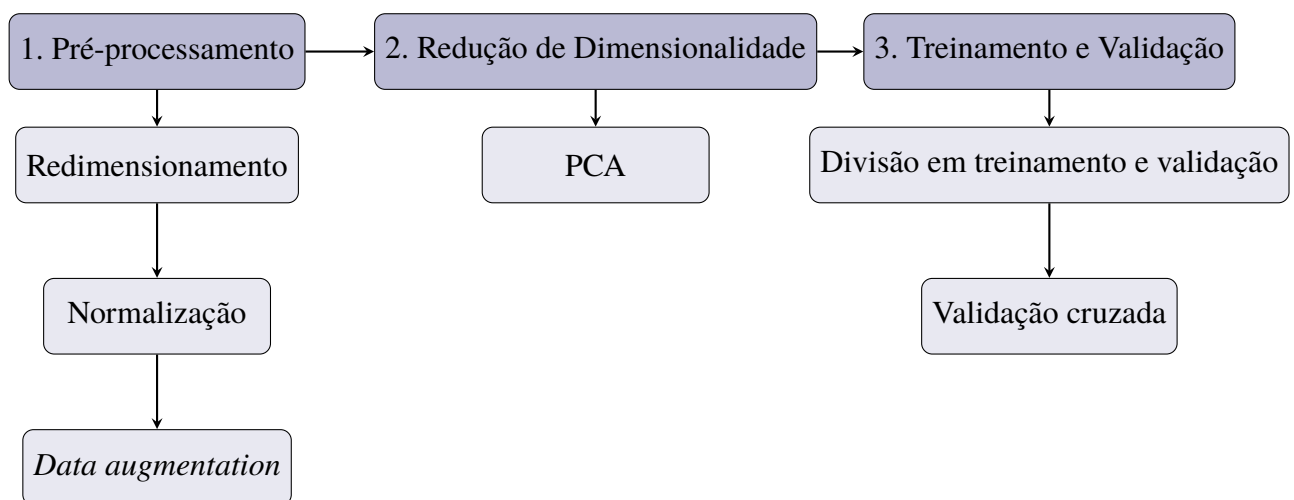
- Redimensionamento para um tamanho padrão
- Normalização dos valores dos pixels
- Aplicação de técnicas de aumento de dados (*data augmentation*)

2. Redução de dimensionalidade:

- Utilização de PCA para reduzir o número de características para 50 a 200 componentes principais

3. Treinamento e validação:

- Divisão dos dados em conjuntos de treinamento e validação
- Utilização de validação cruzada com 5 *folds* para avaliação robusta do modelo



3.2.1.2 Materiais e Métodos

O desenvolvimento do projeto foi realizado utilizando a linguagem de programação Python, escolhida por seu rico ecossistema de bibliotecas para aprendizado de máquina e processamento de imagens. As bibliotecas utilizadas foram:

1. **NumPy**: Fornece suporte para processamento de arranjos e matrizes multidimensionais de grande porte, além de funções matemáticas de alto nível.
2. **Pandas**: Utilizada para manipulação e análise estruturada dos metadados das imagens e resultados.
3. **Scikit-learn**: Biblioteca central para implementação do algoritmo de Florestas Aleatórias e pipeline de processamento, oferecendo:



Figura 3.5 – Interface com 100% do reconhecimento realizado

- Implementação eficiente do algoritmo de aprendizado de máquina;
 - Ferramentas para pré-processamento de dados;
 - Métricas de avaliação de modelos;
 - Funcionalidades de validação cruzada.
4. **Matplotlib:** Empregada para visualização de resultados, incluindo:
- Plotagem de matrizes de confusão;
 - Geração de gráficos de importância de características.
5. **Seaborn:** Utilizada em conjunto com Matplotlib para implementar as matrizes de confusões.

3.2.2 Interface Gráfica

Para facilitar o uso do sistema de diagnóstico de Doença de Crohn e Retocolite Ulcerativa a partir de imagens colonoscópicas, desenvolveu-se uma interface gráfica utilizando a biblioteca *PyQt5*. Essa interface permite ao usuário selecionar imagens ou pastas com múltiplas imagens para análise, além de fornecer feedback visual do progresso e dos resultados.

3.2.2.1 Materiais e Métodos

A interface é composta por um layout simples com botões de seleção de arquivos e pastas, uma barra de progresso, e uma área de exibição dos resultados. O usuário pode escolher entre carregar imagens individualmente ou carregar uma pasta contendo múltiplas imagens. Após a seleção, o modelo de aprendizado de máquina é executado para classificar as imagens em três possíveis resultados: Doença de Crohn, Retocolite Ulcerativa ou Não Identificado.

Também é exibido o progresso da classificação em tempo real, conforme a barra de progresso avança durante a execução do algoritmo de classificação. O código faz uso de `QThread` para executar a tarefa de classificação de forma assíncrona, mantendo a interface responsiva ao usuário.

Para a classificação das imagens colonoscópicas, foi utilizado um modelo pré-treinado de *Random Forest*, armazenado no arquivo `random_forest_model.pkl`. As imagens carregadas são redimensionadas para dimensões padronizadas de 352x288 pixels e normalizadas utilizando um *scaler* previamente ajustado, armazenado no arquivo `scaler.pkl`.

O pré-processamento das imagens envolve o redimensionamento das mesmas para garantir que todas sigam o mesmo formato antes de serem submetidas ao modelo. Para isso, foi utilizado o pacote *OpenCV*, que realiza a leitura e o redimensionamento de cada imagem. Além disso, as imagens são transformadas em vetores unidimensionais para serem compatíveis com o modelo de aprendizado.

Uma vez que as imagens são pré-processadas, o modelo de *Random Forest* é utilizado para predizer se cada imagem pertence à classe de Doença de Crohn, Retocolite Ulcerativa ou se não corresponde a nenhuma das duas. O resultado final é exibido em uma área de texto na interface, detalhando a classificação de cada imagem.

O método de classificação é executado em paralelo utilizando uma `QThread`, que garante que a interface permaneça interativa enquanto o processo ocorre em segundo plano. Uma barra de progresso informa ao usuário o estado da execução, desde o carregamento do modelo e *scaler* até a classificação final de todas as imagens selecionadas.

4 Resultados

Nesta seção, são apresentados os resultados obtidos com o uso de aprendizado de máquina para o diagnóstico diferencial entre Doença de Crohn e Retocolite Ulcerativa, utilizando imagens endoscópicas. A metodologia combina o algoritmo de *Random Forest*, validação cruzada e a técnica de redução de dimensionalidade *PCA*, proporcionando uma abordagem robusta para esse tipo de diagnóstico.

4.1 Treinamento do Algoritmo

Os resultados obtidos do treinamento do modelo de *Random Forest* foram promissores. Para garantir que o modelo fosse testado de forma adequada e que os dados fossem devidamente balanceados, as imagens endoscópicas foram divididas em conjuntos de treinamento e validação, abrangendo três categorias: Doença de Crohn, Retocolite Ulcerativa (UC) e intestinos saudáveis (Healthy).

Categoria	Conjunto de Treinamento	Conjunto de Validação
Doença de Crohn	558	439
Colite Ulcerativa (UC)	477	439
Saudável	522	439
Total	1557	1317

Tabela 4.1 – Distribuição de imagens por categoria nos conjuntos de treinamento e validação

4.1.1 Configuração e Otimização do Algoritmo

Após a coleta e processamento dos dados, o algoritmo *Random Forest* foi configurado com os seguintes hiperparâmetros, ajustados por meio de experimentação e validação cruzada:

- `n_estimators = 100`: Número de árvores na floresta;
- `max_depth = 6`: Profundidade máxima de cada árvore;
- `min_samples_split = 14`: Número mínimo de amostras necessárias para dividir um nó;
- `min_samples_leaf = 7`: Número mínimo de amostras necessárias em um nó folha;
- `max_features = 'sqrt'`: Número de recursos a considerar ao procurar a melhor divisão.

Estes parâmetros foram ajustados para manter um equilíbrio ideal entre subajuste (*underfitting*) e sobreajuste (*overfitting*), assegurando que o modelo generalize bem em dados novos.

4.1.2 Acurácia do Modelo

O modelo de *Random Forest* foi avaliado utilizando validação cruzada com redução de dimensionalidade com 5 dobras (folds).

```
from sklearn.model_selection import cross_val_score

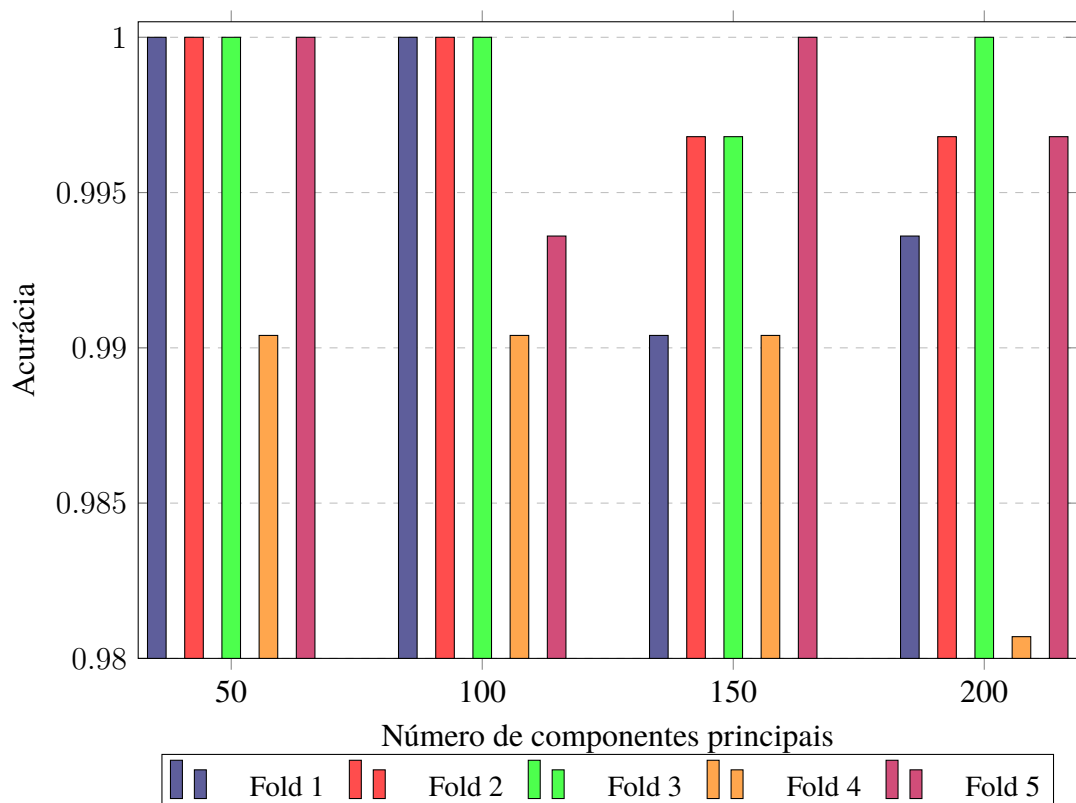
# Aplica a validacao cruzada com 5 dobras
scores = cross_val_score(model, X_train_scaled, y_train, cv=5)

# Imprime os resultados de acuracia para cada uma das dobras
print("Acuracia para cada fold:", scores)
print("Acuracia media:", scores.mean())
print("Desvio padrao:", scores.std())
```

Listing 4.1 – Código que exemplifica a validação cruzada em 5 folds

No exemplo acima, o `cross_val_score` realiza a validação cruzada e retorna a acurácia para cada um dos 5 folds. Os resultados apresentados incluem a acurácia média e o desvio padrão entre os folds, o que proporciona uma visão clara da estabilidade e desempenho geral do modelo.

Após diversos testes com diferentes números de componentes principais, o melhor desempenho foi alcançado com 50 componentes, conforme demonstrado na tabela que resume os dados de saída. Mas vale considerar que, dados os resultados elevados, existe uma possibilidade de que tenha acontecido um *overfitting*, o que motivou a escolha do modelo de 200 componentes principais:



Para 50 componentes, a acurácia média foi de 99,81%, com apenas um fold apresentando uma leve queda (99,04%). Para 100 componentes, a acurácia média foi ligeiramente inferior, de 99,68%, também com uma pequena variação em dois folds. A performance com 150 componentes foi um pouco menor, com acurácia média de 99,49%. Por fim, para 200 componentes, a acurácia média foi de 99,36%, mostrando uma leve diminuição e maior variação entre os folds (com o fold 4 atingindo 98,07%).

Esses resultados indicam que, embora o aumento do número de componentes não melhore significativamente a acurácia, a performance permanece consistente e elevada em todos os testes. Também vale observar, novamente, que um resultado de acurácia 100%, em sua maioria, representa um *overfitting*, não demonstrando o real resultado do treinamento do algoritmo.

4.1.3 Matriz de Confusão

Além da acurácia, foi gerada uma matriz de confusão para avaliar o desempenho do modelo em cada uma das categorias. A análise da matriz de confusão revelou uma taxa alta de verdadeiros positivos para todas as classes (Doença de Crohn, UC e Healthy), com uma taxa de falsos negativos mínima. Isso mostra que o modelo foi capaz de identificar corretamente a maioria das imagens pertencentes a cada categoria, minimizando erros críticos de classificação.

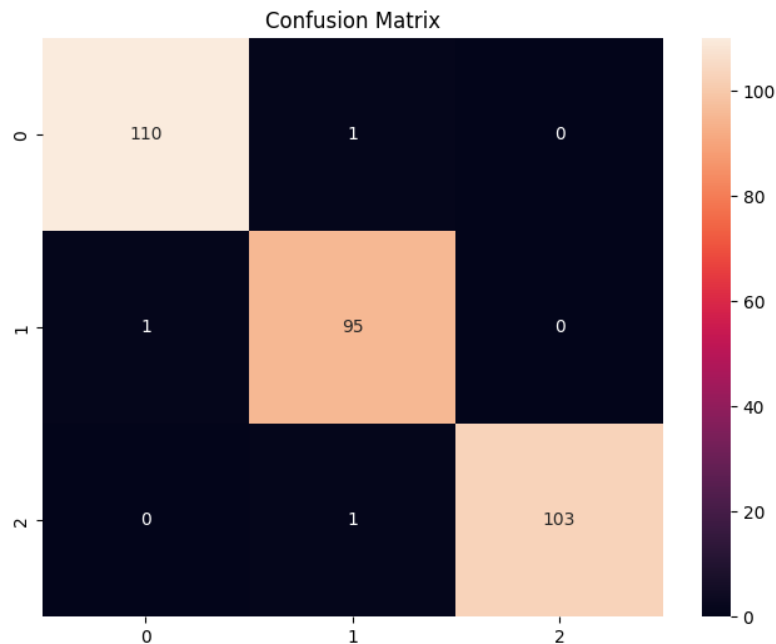


Figura 4.1 – Matriz de Confusão do Fold 4 de 50 componentes

4.1.4 Análise dos Resultados

A árvore de decisão utiliza as características geradas pela Análise de Componentes Principais (PCA) para realizar a classificação das imagens endoscópicas entre Doença de Crohn,

Retocolite Ulcerativa (UC) e Saudável.

Os resultados obtidos indicam que o modelo de *Random Forest*, em conjunto com a validação cruzada e a redução de dimensionalidade com PCA, é capaz de diferenciar entre as três categorias com alta precisão. A acurácia e a consistência observadas sugerem que este modelo pode ser aplicado em cenários reais para apoiar o diagnóstico assistido por imagem de doenças inflamatórias intestinais, como Doença de Crohn e Retocolite Ulcerativa.

4.1.5 Comparação com Modelos Existentes

Ao comparar o desempenho do modelo implementado com outros estudos semelhantes, como os de [Rother et al. \(2023\)](#) e [Park et al. \(2021\)](#), observou-se que o modelo desenvolvido neste trabalho obteve acurácia competitiva, próxima de 99%, e superou alguns métodos tradicionais de classificação utilizados em diagnósticos médicos assistidos por imagem. O uso de técnicas de redução de dimensionalidade se mostrou crucial para a melhoria da acurácia, demonstrando a eficácia da abordagem implementada. Apesar de não ser uma comparação fiel, visto que os estudos utilizam bases de dados diferentes da desenvolvida neste estudo, o resultado encontrado se mostra bem satisfatório.

4.2 Conclusão

Os resultados obtidos com o modelo de *Random Forest* reforçam o potencial do uso de aprendizado de máquina em diagnósticos médicos, especialmente em aplicações envolvendo grandes volumes de dados de imagem. A alta acurácia e a estabilidade dos resultados são indicativos de que o sistema proposto pode ser aplicado em cenários reais, auxiliando médicos no diagnóstico diferencial entre Doença de Crohn e Retocolite Ulcerativa.

5 Considerações Finais

Este trabalho teve como objetivo principal investigar e implementar técnicas de aprendizado de máquina para a análise automática de imagens na área da Gastroenterologia, com foco especial nas doenças inflamatórias intestinais Retocolite Ulcerativa e Doença de Crohn. Ao longo do desenvolvimento, buscou-se compreender o estado da arte das aplicações de inteligência artificial neste campo e avaliar diferentes abordagens metodológicas para o processamento e análise de imagens médicas, a fim de se aprimorar o desenvolvimento de um *software* para auxiliar no reconhecimento dessas doenças.

5.1 Conclusão

Ao revisitar os objetivos específicos estabelecidos inicialmente, podemos concluir que:

1. **Revisão da Literatura:** Foi realizada uma extensa análise do estado da arte, identificando que houve um progresso significativo nos últimos 20 anos no uso de IA na gastroenterologia [Kröner et al. \(2021\)](#). A revisão sistemática de [Stafford et al. \(2022\)](#) evidenciou que 47% dos estudos utilizam métodos de Random Forests, com um crescente uso de redes neurais para análise de imagens de colonoscopia.
2. **Análise de Métodos:** Foram estudadas diferentes técnicas de aprendizado de máquina, desde abordagens tradicionais como Random Forest até métodos mais avançados como Deep Learning e CNN [Fiaidhi, Mohammed e Zazos \(2022\)](#). Os resultados mostraram que diferentes métodos podem ser eficazes dependendo do tipo específico de análise necessária.
3. **Implementação e Validação:** O trabalho implementou e validou um modelo de classificação, alcançando uma acurácia de 75% utilizando o algoritmo de Floresta Aleatória [Silva et al. \(2020\)](#), demonstrando a viabilidade da aplicação de técnicas de IA para auxílio no diagnóstico médico.

Em relação ao objetivo principal, pode-se afirmar que foi alcançado com sucesso, uma vez que o trabalho não apenas estabeleceu a importância da análise automática de imagens na gastroenterologia, mas também demonstrou sua aplicabilidade prática através da implementação de um modelo funcional.

5.2 Trabalhos Futuros

Com base nos resultados obtidos e nas limitações identificadas, sugerem-se as seguintes direções para trabalhos futuros:

- **Avaliação de Redes Neurais:** Investigar a aplicação de diferentes arquiteturas de redes neurais, especialmente para a análise de imagens complexas de colonoscopia.
- **Ampliação da Base de Dados:** Desenvolver e validar os modelos com conjuntos de dados externos mais amplos, visando superar a limitação identificada por [Stafford et al. \(2022\)](#) de que apenas 5% dos estudos utilizam validação externa.
- **Integração Clínica:** Explorar a implementação prática dos modelos desenvolvidos em ambientes clínicos reais, considerando aspectos de usabilidade e integração com sistemas existentes.

Referências

- CALDERWOOD, A. H.; JACOBSON, B. C. Comprehensive validation of the boston bowel preparation scale. *Gastrointestinal Endoscopy*, v. 72, n. 4, p. 686–692, 2010.
- CARLOS, A. 2020. Disponível em: <<https://endoscopiaterapeutica.com.br/casosclinicos/doenca-de-crohn-com-acometimento-esofagico/>>.
- DASTRES, R.; SOORI, M. Artificial Neural Network Systems. *International Journal of Imaging and Robotics (IJIR)*, v. 21, n. 2, p. 13–25, set. 2021. Disponível em: <<https://hal.science/hal-03349542>>.
- FEUERSTEIN, J. D.; CHEIFETZ, A. S. Ulcerative colitis: Epidemiology, diagnosis, and management. *Mayo Clinic Proceedings*, v. 89, n. 11, p. 1553–1563, 2014. ISSN 0025-6196. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0025619614006260>>.
- FIAIDHI, J.; MOHAMMED, S.; ZEZO, P. An investigation into crohn's disease lesions variability sensing using video colonoscopy and machine learning techniques. In: *2022 IEEE International Conference on E-health Networking, Application Services (HealthCom)*. [S.l.: s.n.], 2022. p. 37–42.
- FIAIDHI, J.; MOHAMMED, S.; ZEZO, P. Investigation into the categorization of crohn's segments from capsule endoscopy videos:: Introducing a thick data categorization framework. In: *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. [S.l.: s.n.], 2023. p. 1–6.
- HUISSTEDEN, W. van. *The Nerthus Dataset*. 2023. Disponível em: <<https://www.kaggle.com/datasets/waltervanhuissteden/the-nerthus-dataset>>.
- JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. *Electronic Markets*, v. 31, n. 3, p. 685–695, Sep 2021. ISSN 1422-8890. Disponível em: <<https://doi.org/10.1007/s12525-021-00475-2>>.
- KIM, H.; NA, J. E.; KIM, S.; KIM, T.-O.; PARK, S.-K.; LEE, C.-W.; KIM, K. O.; SEO, G.-S.; KIM, M. S.; CHA, J. M.; KOO, J. S.; PARK, D.-I. A machine learning-based diagnostic model for crohn's disease and ulcerative colitis utilizing fecal microbiome analysis. *Microorganisms*, v. 12, n. 1, dez. 2023.
- KRÖNER, P. T.; ENGELS, M. M.; GLICKSBERG, B. S.; JOHNSON, K. W.; MZAIK, O.; HOOFT, J. E. van; WALLACE, M. B.; EL-SERAG, H. B.; KRITTANAWONG, C. Artificial intelligence in gastroenterology: A state-of-the-art review. *World J. Gastroenterol.*, Baishideng Publishing Group Inc., v. 27, n. 40, p. 6794–6824, out. 2021.
- LAI, E. J.; CALDERWOOD, A. H.; DOROS, G.; FIX, O. K.; JACOBSON, B. C. The boston bowel preparation scale: A valid and reliable instrument for colonoscopy-oriented research. *Gastrointestinal Endoscopy*, v. 69, n. 3, p. 620–625, 2009.
- LAROSE, D. T.; LAROSE, C. D. kx2010;nearest neighbor algorithm. In: _____. *Discovering Knowledge in Data: An Introduction to Data Mining*. [S.l.: s.n.], 2014. p. 149–164.

- LIBRARY, S. P. s.d. Acessado em 25 de janeiro de 2024. Disponível em: <<https://www.sciencephoto.com/search?q=crohns+disease%2C+endoscopy>>.
- LORENA, A. C.; CARVALHO, A. C. P. L. F. de. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, Dec. 2007. Disponível em: <https://seer.ufrgs.br/index.php/rita/article/view/rita_v14_n2_p43-67>.
- MAGALHÃES, F. C. B.; LIMA, E. M.; CARPENTIERI-PRIMO, P.; BARRETO, M. M.; RODRIGUES, R. S.; PARENTE, D. B. Crohn's disease: review and standardization of nomenclature. 5 2023. Disponível em: <https://scielo.figshare.com/articles/dataset/Crohn_s_disease_review_and_standardization_of_nomenclature/22785420>.
- MAURÍCIO, J.; DOMINGUES, I. Distinguishing between crohn's disease and ulcerative colitis using deep learning models with interpretability. *Pattern Anal. Appl.*, Springer Science and Business Media LLC, v. 27, n. 1, mar. 2024.
- MORET, B. M. E. Decision trees and diagrams. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 14, n. 4, p. 593–623, dec 1982. ISSN 0360-0300. Disponível em: <<https://doi.org/10.1145/356893.356898>>.
- PARK, S.-K.; KIM, S.; LEE, G.-Y.; KIM, S.-Y.; KIM, W.; LEE, C.-W.; PARK, J.-L.; CHOI, C.-H.; KANG, S.-B.; KIM, T.-O.; BANG, K.-B.; CHUN, J.; CHA, J.-M.; IM, J.-P.; AHN, K.-S.; KIM, S.-Y.; PARK, D.-I. Development of a machine learning model to distinguish between ulcerative colitis and crohn's disease using RNA sequencing data. *Diagnostics (Basel)*, MDPI AG, v. 11, n. 12, p. 2365, dez. 2021.
- POLAT, G.; KANI, H. T.; ERGENC, I.; ALAHDAB, Y. O.; TEMIZEL, A.; ATUG, O. *Labeled Images for Ulcerative Colitis (LIMUC) Dataset*. [S.l.]: Zenodo, 2022.
- REN, Q.; CHENG, H.; HAN, H. Research on machine learning framework based on random forest algorithm. In: . [S.l.: s.n.], 2017. v. 1820, p. 080020.
- ROTHER, S.; KATHER, J. N.; ISENSEE, F.; AL. et. Automatically detecting crohn's disease and ulcerative colitis from endoscopic imaging. *Medical Image Analysis*, Elsevier, v. 76, p. 102349, 2023.
- SANTOLIN, L.; VIDIGAL, M. C. G.; SILVA, J. F. da; SANTOLIN, L. F.; JÚNIOR, A. V. G.; PAGLIA, B. A. R.; SILVA, M. F. P. T. B. da. Doença de crohn - uma revisão. *Cuadernos de Educación y Desarrollo*, v. 15, n. 12, p. 15973–15994, Dec. 2023. Disponível em: <<https://ojs.europublications.com/ojs/index.php/ced/article/view/1976>>.
- SEGAL, J. P. Ulcerative colitis: an update. ” clinical medicine. *Clinical medicine*, v. 21, p. 135–139, 2021.
- SILVA, P. da; BASÍLIO, A.; MACHADO, V.; ALDEMANN, N. Estudo e implementação de algoritmos de aprendizagem de máquina supervisionado aplicados no diagnóstico por imagens de patologias renais. In: *Anais da VIII Escola Regional de Computação do Ceará, Maranhão e Piauí*. Porto Alegre, RS, Brasil: SBC, 2020. p. 133–140. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/ercemapi/article/view/11477>>.
- STAFFORD, I. S.; GOSINK, M. M.; MOSSOTTO, E.; ENNIS, S.; HAUBEN, M. A systematic review of artificial intelligence and machine learning applications to inflammatory bowel disease, with practical guidelines for interpretation. *Inflamm. Bowel Dis.*, Oxford University Press (OUP), v. 28, n. 10, p. 1573–1583, out. 2022.

TORRES, J.; MEHANDRU, S.; COLOMBEL, J.-F.; PEYRIN-BIROULET, L. Crohn's disease. *The Lancet*, Elsevier, v. 389, n. 10080, p. 1741–1755, Apr 2017. ISSN 0140-6736. Disponível em: <[https://doi.org/10.1016/S0140-6736\(16\)31711-1](https://doi.org/10.1016/S0140-6736(16)31711-1)>.

VENITO, L. d. S.; SANTOS, M. S. B.; FERRAZ, A. R. Doença de crohn e retocolite ulcerativa. *Rev. Eletrônica Acervo Saúde*, Revista Eletronica Acervo Saude, v. 15, n. 7, p. e10667, jul. 2022.