PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS NÚCLEO DE EDUCAÇÃO A DISTÂNCIA

Categorização e Classificação de cânceres de mama utilizando Rede Neural Convolucional (CNN)

São João de Meriti

Rodrigo de Brito de Oliveira Cabral

Classificação de cânceres de mama utilizando Rede Neural Convolucional (CNN)

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

São João de Meriti
2022

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

SIB PUC MINAS

AGRADECIMENTOS

Agradeço primeiramente a Deus pela saúde e oportunidade de concluir esse trabalho em tempos pandêmicos.

Aos meus pais, "Carlos Cesar Monteiro Cabral"e "Edinalva de Brito de Oliveira Cabral", por me acompanharem em muitos momentos importantes da minha vida e sempre incentivando o meu crescimento intelectual.

Aos amigos que me ajudaram nesse momento tão complicado que o mundo vive.

Aos professores e a PUC Minas por todo o conhecimento partilhado durante todo o curso.

Ao meu antigo mentor e minha atual mentora profissionais: "Rômulo Canella"e "Andreza Gonçalves", respectivamente.Por sempre me apoiarem enquanto a realização do curso.

Por último e não menos importante, gostaria de agradecer ao meu futuro filho, Nicolas Ravi, que está com previsão de nascimento em março de 2022. Pois, sua futura chegada me deu animo para ser cada vez melhor

LISTA DE FIGURAS

FIGURA 1 – Amostra de Histopatologia, na escala 40x 16
FIGURA 2 – Amostra de Histopatologia 17
FIGURA 3 – Amostra de Histopatologia nas amplificações (a)4x, (b)10x, (c)20x e (d)40x 17
FIGURA 4 – Screenshot da tela principal do Pycharm 18
FIGURA 5 – Código para mostrar nomes dos arquivos (Jupyter notebook) 19
FIGURA 6 – Amostra com nomes de imagens (Jupyter notebook) 19
FIGURA 7 – Função para mudar o nome das imagens (Jupyter notebook) 20
FIGURA 8 – Amostra com novos nomes, ampliação 40x 20
FIGURA 9 – Amostra com novos nomes, ampliação 100x 20
FIGURA 10 – Amostra com novos nomes, ampliação 200x 21
FIGURA 11 – Amostra com novos tamanhos das imagens 21
FIGURA 12 – Amostra da base dados Breast invasive ductal carcinomas (Jupyter notebook) 23
FIGURA 13 – Código de importação dos módulos (Google Colab) 24
FIGURA 14 – Código de criação do *"DataFrame"* (Google Colab) 25
FIGURA 15 – Amostra do *"DataFrame"* criado (Google Colab) 25
FIGURA 16 – Valores únicos de cada coluna (Google Colab) 25
FIGURA 17 – Representação gráfica de uma rede neural fictícia 32
FIGURA 18 – Representação gráfica de um neurônio fictício 33
FIGURA 19 – Representação gráfica de perceptron 34
FIGURA 20 – Representação gráfica de perceptron com entrada e saída 35
FIGURA 21 – Modelo de Perceptron não linear 35
FIGURA 22 – Diagrama de uma MLPs 36
FIGURA 23 – Diagrama de uma "Rede neural feedforward"e uma "Rede neural recorrente" 37
FIGURA 24 – Arquitetura de uma Rede neural convolucional 40
FIGURA 25 – Imagem representada em forma de matriz. 41
FIGURA 26 – Representação da camada de convolução 41

FIGURA 27 – Max pooling retângulo 2x2 42
FIGURA 28 – Average pooling retângulo 2x2 42
FIGURA 29 – Importações Biblioteca (Google Colab) 44
FIGURA 30 – Declaração das constantes (Google Colab) 44
FIGURA 31 – Segmentação Dataframe 400x (Google Colab) 44
FIGURA 32 – Segmentação Dataframe 200x (Google Colab) 45
FIGURA 33 – Diagrama de fluxo criação do modelo categoria 46
FIGURA 34 – Diagrama de fluxo criação do modelo classe 46
FIGURA 35 – Modelo Canvas 47
FIGURA 36 – Criação e carregamento de pesos (Google Colab) 48
FIGURA 37 – Código para teste do dataset (Google Colab) 48
FIGURA 38 – Criação e carregamento de pesos (Google Colab) 48
FIGURA 39 – Amostra com novos nomes, ampliação 400x 68
FIGURA 40 – Amostra com nomes de imagens (Jupyter notebook) 68
FIGURA 41 – Amostra da base dados Breast invasive ductal carcinomas (Jupyter notebook) 69
FIGURA 42 – Código para pegar nome da figura (Google Colab) 69
FIGURA 43 – Código para pegar grau de ampliação (Jupyter notebook) 69
FIGURA 44 – Código para selecionar a classe (Google Colab) 70
FIGURA 45 – Código para pegar os tipos (Jupyter notebook) 70

LISTA DE TABELAS

TABELA 1 – "BreakHis"resumo geral 16
TABELA 2 – "BreakHis"por categoria de câncer 16
TABELA 3 – Resumo geral 17
TABELA 4 – Tabela de métricas do gráfico 12 49
TABELA 5 – Tabela de métricas do gráfico 13 50
TABELA 6 – Tabela de métricas do gráfico 14 50
TABELA 7 – Tabela de métricas do gráfico 15 51
TABELA 8 – Tabela de métricas do gráfico 16 52
TABELA 9 – Tabela de métricas do gráfico 17 53
TABELA 10 – Tabela de métricas do gráfico 15 54
TABELA 11 – Tabela de métricas do gráfico 19 55
TABELA 12 – Tabela de métricas do gráfico 20 55
TABELA 13 – Tabela de métricas do gráfico 21 56
TABELA 14 – Tabela de métricas do gráfico 22 57
TABELA 15 – Tabela de métricas do gráfico 23 58
TABELA 16 – Tabela de métricas do gráfico 24 59
TABELA 17 – Tabela de métricas do gráfico 25 60
TABELA 18 – Tabela de métricas do gráfico 26 60
TABELA 19 – Tabela de métricas do gráfico 27 61
TABELA 20 – Tabela de métricas do gráfico 28 62
TABELA 21 – Tabela de métricas do gráfico 29 63
TABELA 22 – Tabela de métricas dos modelos para Categorias de câncer 29 . . 64
TABELA 23 – Tabela de métricas dos modelos para classe do câncer 29 64
TABELA 24 – Tabela de métricas das combinações 65

LISTA DE GRÁFICOS

GRÁFICO 1 – Distribuição de casos e mortes para os 10 principais cânceres mais comuns em 2020 para (A) Ambos os sexos, (B) Homens e (C) Mulheres 12
GRÁFICO 2 – Distribuição de Casos e Mortes por Área Mundial em 2020 para (A)Ambos os Sexos,(B)Homens e (C)Mulheres 13
GRÁFICO 3 – Distribuição de categorias de cancer 22
GRÁFICO 4 – Distribuição de Treino e Teste no *Dataset* 26
GRÁFICO 5 – Quantidade de dados entre Treino e Teste no *Dataset* 26
GRÁFICO 6 – Distribuição de Beigno e Maligno no Dataset 27
GRÁFICO 7 – Quantidade de dados entre Beigno e Maligno no Dataset 27
GRÁFICO 8 – Distribuição em porcentagens das classes 28
GRÁFICO 9 – Distribuição em porcentagens das classes referente aos tipos Malignos 29
GRÁFICO 10 – Distribuição em valores percentuais das classes por *"Magnification"* 30
GRÁFICO 11 – Distribuição em valores percentuais das classes base de treino e teste 31
GRÁFICO 12 – Matriz de Confusão Categoria de Cancer 400X 49
GRÁFICO 13 – Matriz de Confusão Categoria de Cancer 200X 49
GRÁFICO 14 – Matriz de Confusão Categoria de Cancer 50
GRÁFICO 15 – Matriz de Confusão Classe de Cancer 400X 51
GRÁFICO 16 – Matriz de Confusão Classe de Cancer 200X 52
GRÁFICO 17 – Matriz de Confusão Classe de Cancer 53
GRÁFICO 18 – Matriz de Confusão Categoria de Cancer 400X 54
GRÁFICO 19 – Matriz de Confusão Categoria de Cancer 200X 54
GRÁFICO 20 – Matriz de Confusão Tipo de Cancer 55
GRÁFICO 21 – Matriz de Confusão Classe de Cancer 400X 56
GRÁFICO 22 – Matriz de Confusão Classe de Cancer 200X 57
GRÁFICO 23 – Matriz de Confusão Classe de Cancer 58
GRÁFICO 24 – Matriz de Confusão Categoria de Cancer 400X 59
GRÁFICO 25 – Matriz de Confusão Categoria de Cancer 200X 59
GRÁFICO 26 – Matriz de Confusão Categoria de Cancer 60
GRÁFICO 27 – Matriz de Confusão Classe de Cancer 400X 61
GRÁFICO 28 – Matriz de Confusão Classe de Cancer 200X 62

GRÁFICO 29 – Matriz de Confusão Classe de Cancer 63

SUMÁRIO

1 INTRODUÇÃO **10**
1.1 Contextualização 10
1.2 O problema proposto 13

2 Coleta de Dados **15**
2.1 Breast Cancer Histopathological Database (BreakHis) 15
2.2 A histopathological image dataset for grading breast invasive ductal carcinomas 16

3 Processamento/Tratamento de Dados **18**
3.1 Ferramentas utilizadas 18
3.2 Tratamento das bases de dados 18
3.2.1 Tratamento da base de dados BreakHis 18
3.2.2 Tratamento da base de dados Breast invasive ductal car-cinomas . . 21

4 Análise e Exploração dos Dados **24**

5 Criação de Modelos de Machine Learning **32**
5.1 Rede Neural Natural (R.N.A.) 32
5.2 Rede Neural Artificial (R.N.A.) 33
5.2.1 Rede Neural Artificial Multicamadas ou MLPs 36
5.2.2 Funções de Ativação 37
5.2.3 Treinamento de uma Rede Neural 38
5.2.4 Algoritmo Backpropagation 39
5.2.5 Redes neurais convolucionais ou CNN 39
5.2.5.1 Image channels 40
5.2.5.2 Convolution 41
5.2.5.3 Pooling 42
5.2.5.4 Flattening 42
5.2.6 Transferência de aprendizado 43
5.3 Execução dos Modelos 43

6 Apresentação dos Resultados **47**
6.1 Modelo Preditivo - VGG19 48
6.2 Modelo Preditivo - GoogleNet 53
6.3 Modelo Preditivo - ResNet50 58
6.4 Resultado Geral 63
6.5 Combinação de Modelos 64

7 Trabalhos futuros **66**

8 Links **67**

9 Figuras / Graficos **68**

REFERÊNCIAS **71**

1 INTRODUÇÃO

1.1 Contextualização

A medicina é a área da ciência com o maior peso acadêmico. Através dos estudos nessa área a humanidade foi capaz de curar doenças, como, por exemplo, a tuberculose e a varíola. Porém, tão importante quanto curar uma enfermidade é saber como identifica-la, com esse objetivo, foram desenvolvidos vários métodos de detecção de enfermidades ao longo dos anos.

O diagnóstico por imagem é um conjunto de métodos que permite a obtenção de imagens do corpo, seja por inteiro ou de partes específicas. Após a obtenção da imagem, é necessário passar por uma análise clínica, com a finalidade de encontrar anomalias. No entanto, o profissional que exerce essa função se encontra sob grande pressão, dada a responsabilidade inerente de sua tarefa.

A mamografia é um tipo especial de técnica de Raios-x usado para criar imagens detalhadas da mama. É utilizado um aparelho chamado mamógrafo, onde, depois da exposição da mama aos Raios-x, o aparelho transforma a radiação em um sinal elétrico e o envia para um computador. As mamas são comprimidas com espátulas de plástico para otimizar a visualização do tecido da mama e suas anomalias.As imagens geradas ficam armazenadas em formato digital, disponíveis para impressão ou futuras consultas, tanto pela paciente quanto por radiologistas e outros profissionais da saúde .(MORSCH, 2018)

As mamografias de triagem são usadas para rastrear o câncer de mama em mulheres assintomáticas. As mamografias diagnósticas são usadas para diagnosticar doenças mamárias em mulheres com sintomas mamários, nódulos mamários palpáveis, resultados alterados no exame de mamografia que exigem maior investigação. As mamografias diagnósticas podem incluir incidências padrão e especializadas (OLIVEIRA, 2019).

A histologia é uma disciplina que estuda os tecidos dos seres vivos e suas organizações microscópicas, onde é possível entender claramente a maneira pela qual eles se relacionam entre si. Sendo a histologia patológica(ou Histopatologia) uma derivação da histologia.(O. . . , 2019)

> O exame histopatológico é, em linhas gerais, uma análise microscópica de um tecido para a detecção de possíveis alterações ou lesões, com o objetivo básico de informar ao clínico a natureza, a gravidade, a extensão, a evolução e a intensidade de algum problema do doador da amostra, além de sugerir ou confirmar a causa do problema.(O. . . , 2019)

No cenário mundial, o Câncer de Mama é uma das três categorias de maior incidência, com Câncer no Pulmão e o de próstata. Em 2020, foram encontrados 2,2 milhões de novos casos de Câncer de Mama, contribuindo com cerca de 11,7% do total de casos de câncer no mundo.Alcançando em 2020 o posto de câncer com mais incidência mundial.(SUNG et al., 2021, Tradução nossa). Como podemos ver no gráfico 1.

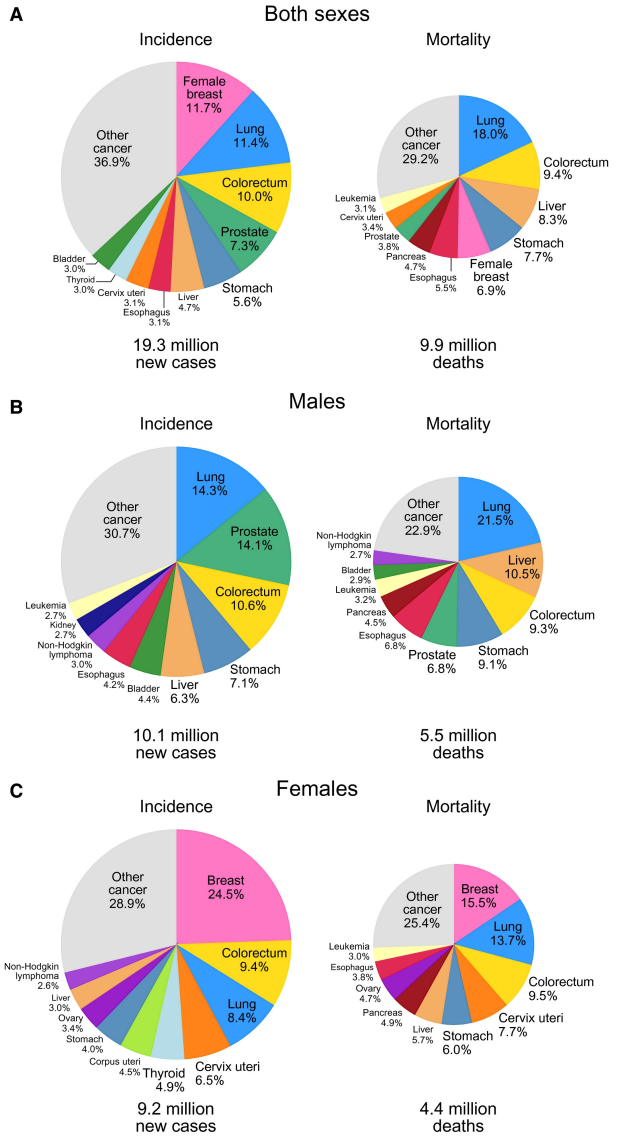
Essa categoria de câncer, é o quarto em questão de mortalidade no mundo, estimadas mais de 684,996 mil mortes em 2020 - o que representa 6,9% do total de mortes por todos os tipos da doença. O artigo também aponta que uma em cada quatro mulheres que têm um caso de câncer diagnosticado têm câncer de mama, representando 24,5% do total.(SUNG et al., 2021, Tradução nossa).Como podemos ver no gráfico 2.

No cenário Nacional, o Câncer de Mama é uma das três categorias de maior incidência INCA (2019), com Câncer de "Cólon e reto"e o de "próstata". Em 2020, foram encontrados 66.280 mil novos casos de Câncer de Mama, contribuindo com cerca de 30% do total de casos de câncer no Brasil.

Essa categoria de câncer, segundo pesquisas do INCA (2019), é o primeiro em questão de mortalidade, estimadas mais de 18 mil mortes em 2019 - o que representa 16,4% do total de mortes por todos os tipos da doença entre mulheres e ocupando o segundo lugar, quando consideramos os 2 sexos.A única região do país em que o câncer de mama não é o mais comum entre as mulheres é a norte, onde o de colo de útero ocupa a primeira posição.

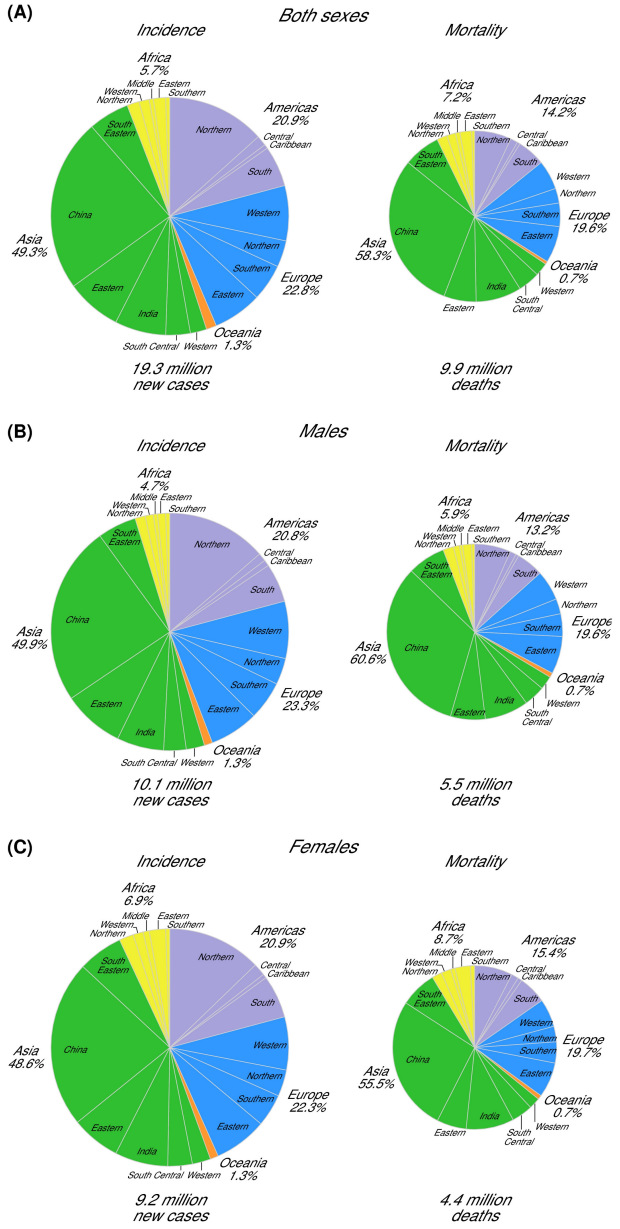
A incidência da doença aumenta em mulheres a partir dos 50 anos, onde poderemos encontrar uma incidência de 4 entre 5 mulheres.

Gráfico 1: Distribuição de casos e mortes para os 10 principais cânceres mais comuns em 2020 para (A) Ambos os sexos, (B) Homens e (C) Mulheres



Fonte: GLOBOCAN 2020

Gráfico 2: Distribuição de Casos e Mortes por Área Mundial em 2020 para (A)Ambos os Sexos,(B)Homens e (C)Mulheres



Fonte: GLOBOCAN 2020

1.2 O problema proposto

O problema proposto consiste na construção de um algoritmo para auxiliar, apenas auxiliar, os profissionais da saúde no diagnóstico de imagens, em particular a análise de imagem histopatológica do câncer de mama. Já que, nenhum algorítimo conseguirá substituir a decisão medica. O trabalho não deve ser visto como um modelo capaz de substituir uma consulta média ou tão menos a análise de um profissional qualificado.

Para facilitar o entendimento do problema e da solução a ser proposta utilizamos

a técnica do "5Ws", que consiste em responder as seguintes perguntas:

Why? Por que auxiliar médicos e profissionais na classificação do câncer a ser tratado e importante?

Segundo dados do Amarante (2021), quando o câncer é descoberto no seu início, a doença tem 95% de hipótese de cura. Segundo o Instituto Nacional de Câncer José Alencar Gomes da Silva (2021), o câncer de mama é responsável por cerca de 30% do total de casos de câncer no Brasil, representando cerca de 16,5% das mortes envolvendo câncer.

Segundo o artigo da Oncológica do Brasil Diagnóstico... (2021), um resultado falso positivo ou negativo interfere diretamente no sucesso do tratamento da doença, podendo gerar um desgaste nem um paciente saudável ou a perca da janela de tratamento em um paciente com câncer.

Who? De quem são os dados analisados?

Os dados utilizados nesse projeto foram gerados e disponibilizados para usos acadêmicos, desde que fossem referenciadas suas origens.

A primeira base de dados, *"Breast Cancer Histopathological Database"* (ou BreakHis), foi gerada para a publicação do artigo *Breast cancer histopathological image classification usingconvolutional neural networks.* (SPANHOL et al., 2016), distribuída pelos autores do artigo.

A segunda base de dados, *"A histopathological image dataset for grading breast invasiveductal carcinomas"*., foi gerada para a publicação do artigo de mesmo nome (BOLHASANI et al., 2020), distribuída pelos autores do artigo.

What? Qual o objetivo do projeto?

Já se sabe que o câncer de mama não é uma doença única, e sim um conjunto de enfermidades que têm em comum a presença da célula maligna originada na mama. Levantando essa ideia, o trabalho realizara duas classificações. Primeiramente ele irá classificar o câncer entre "Maligno"(M) ou "Benigno"(B) e caso ele seja classificado como "Maligno"(M) classificaremos esse câncer com as seguintes categorias: "carcinoma (DC)", "carcinoma lobular(CL)", "Carcinoma mucinoso (MC)"e "Carcinoma papilífero (CP)".

Where? O estudo onde as imagens foram coletadas, foi realizado no Brasil, pelo laboratório "PD Lab".(SPANHOL et al., 2016, tradução nossa)

When? Todas as imagens foram coletadas entre janeiro de 2014 e dezembro de 2014, durante a realização de um estudo clinico.(SPANHOL et al., 2016, tradução nossa)

2 COLETA DE DADOS

Neste trabalho optou-se por utilizar duas fontes de dados, ambas de estudos realizados e publicados em instituição com grande renome.

A primeira fonte de dados, descrita, é a "BreakHis". Um estudo publicado na IEEE (Institute of Electrical and Electronics Engineers) publicado no ano de 2016. Essa fonte de dados é constituída por imagens histopatológicas de cânceres de mama, subdivido entre "adenoses (A)", "Fibroadenoma (F)", "tumor Phyllodes (TP)"e "adenoma tubular (AT)"para câncer benigno e "carcinoma (DC)", "carcinoma lobular(CL)", "Carcinoma mucinoso (MC)"e "Carcinoma papilífero (CP)"para câncer maligno. Poderemos encontrar esse banco de dados em fontes diferentes, no Kaggle (https://www.kaggle.com/kritika397/breast-cancer-dataset-from-breakhis) e na pagina do artigo, no site, da Universidade Federal do Paraná (UFPR) no seguinte link (https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/)

A segunda fonte de dados, descrita, é a "A histopathological image dataset for grading breast invasive ductal carcinomas". Essa fonte contem, enumeras imagens histopatológicas, porem de cânceres benignos. Utilizarei essa fonte de dados, para complementação da "BreakHis", já que a nossa primeira fonte de dados possui um número consideravelmente superior do tipo maligno. E poderia gerar um desbalanço em nosso algorítimo.Essa fonte de dados, como posteriormente explicada, foi disponibilizada para livre uso na "Mendeley Data"

2.1 Breast Cancer Histopathological Database (BreakHis)

The Breast Cancer Histopathological Image Classification (BreakHis) é composto por 7,909 imagens microscópicas de cânceres de mama coletados de 82 pacientes diferentes fatores de ampliação (40X, 100X, 200X, e 400X). O dataset(conjunto de dados) contem 2480 canceres benignos e 5429 canceres malignos. A base de dados, foi construída em parceria entre UFPR (Universidade Federal do Paraná) e o Laboratório de PD - Anatomia Patológica e Citopatologia, Paraná, Brasil.(SPANHOL et al., 2016, tradução nossa)*

Inicialmente as imagens, foram importadas com os dados informativos em seu nome: método de produção da biopsia, classificação do tumor, subtipo do tumor, ampliação utilizada e a identificação do paciente. Como, por exemplo: "SOB-B-TA-14-4659-40-001.png"(SPANHOL et al., 2016, tradução nossa)†

Informação detalhada sobre a quantidade de "espécimes"por ampliação e categorias de cânceres estão presentes nas tabelas 1 e 2. Poderemos ver também uma amostra na figura 1.

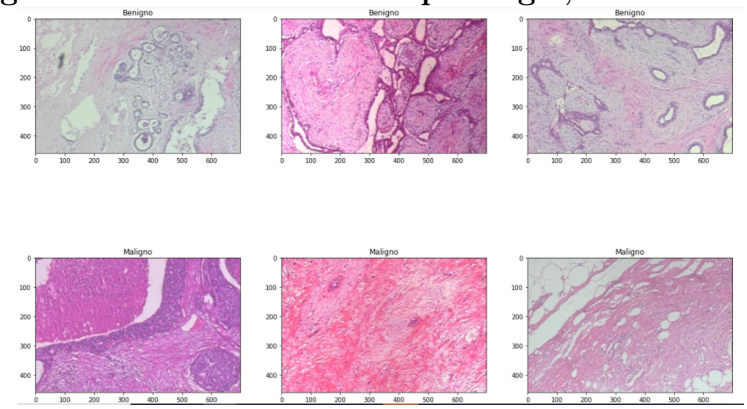
* The Breast Cancer Histopathological Image Classification (BreakHis) is composed of 9,109 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). To date, it contains 2,480 benign and 5,429 malignant samples (700X460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format). This database has been built in collaboration with the PD Laboratory – Pathological Anatomy and Cytopathology, Parana, Brazil.

Tabela 1 – "BreakHis"resumo geral

Ampliação da imagem	Soma
40X	1995
100X	2081
200X	2013
400X	1820
Total	7909

Fonte: Spanhol et al., 2016

Figura 1 – Amostra de Histopatologia, na escala 40x



Fonte: Autor

Tabela 2 – "BreakHis"por categoria de câncer

Ampliação da imagem	Benigno	Maligno
40X	652	1370
100x	644	1437
200x	623	1390
400x	588	1232
Total	2480	5429

Fonte: Spanhol et al., 2016

2.2 A histopathological image dataset for grading breast invasive ductal carcinomas

A histopathological image dataset for grading breast invasive ductal carcinomas, é um "dataset"composto por 922 imagens de 124 pacientes com câncer de mama. Esse "dataset"foi publicado e disponibilizado em um ambiente virtual.diferentes fatores de ampliação (40X, 20X, 10X, e 4X)(BOLHASANI et al., 2020)

A marcação de espécimes neste conjunto de dados foi feita com base em uma classificação histológica de Bloom Richardson modificada. Este método de classificação avalia a quantidade de formação de túbulos, a extensão do pleomorfismo nuclear e a contagem mitótica (ou taxa mitótica).(BOLHASANI et al., 2020)

Cada imagem, ou dado, recebe uma pontuação de 1, 2 ou 3, e as pontuações são somadas para produzir uma nota. A pontuação mitótica é determinada pelo número de figuras mitóticas encontradas em 10 campos consecutivos de alta potência (HPF) na parte

mais mitoticamente ativa do tumor.(BOLHASANI et al., 2020)

Utilizaremos esse "dataset"como um enriquecimento da base de dados original, alcançando assim, uma melhor acurácia do nosso modelo. Obtendo uma menor saturação devido a pouca variedade de ampliações.

> *“ Os erros causados por dados inadequados são muito menores do que aqueles devido à falta total de dados.”* (Babbage Charles, Tradução Nossa)

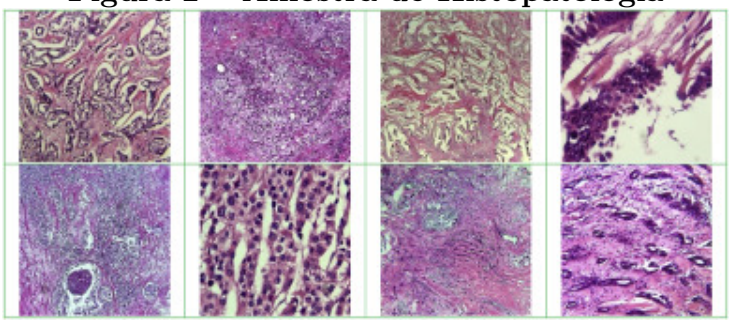
Informação detalhada sobre a quantidade de "espécimes"por ampliação e categorias de cânceres estão presentes nas tabelas 4 . Poderemos ver também duas amostras nas figuras 2 e 3.

Tabela 3 – Resumo geral

Ampliação da imagem	Soma
4X	160
10X	153
20X	155
40X	454
Total	922

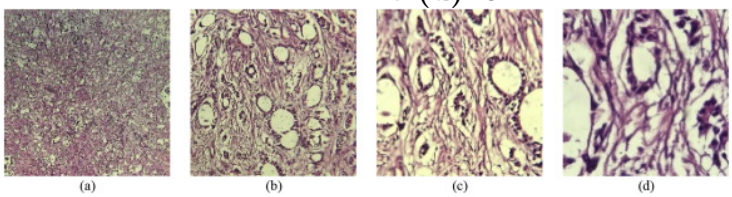
Fonte: Bolhasani et al., 2020

Figura 2 – Amostra de Histopatologia



Fonte: Bolhasani et al., 2020

Figura 3 – Amostra de Histopatologia nas amplificações (a)4x, (b)10x, (c)20x e (d)40x



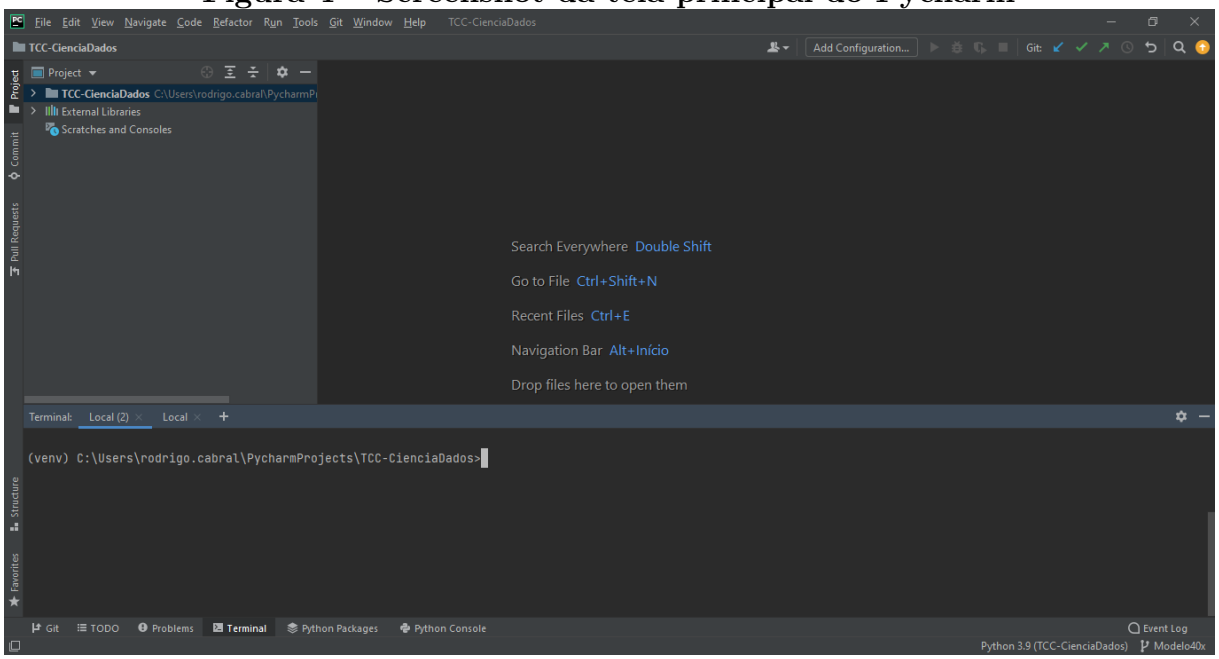
Fonte: Bolhasani et al., 2020

3 PROCESSAMENTO/TRATAMENTO DE DADOS

3.1 Ferramentas utilizadas

Inicialmente utilizamos a para desenvolvimento dos scripts em Python, o PyCharm Community Edition na versão 2021.1.3 (Figura 4) rodando o Python na versão Python 3.9.5. Utilizando também no projeto o Jupyter Notebook na versão 3.0.16. Na linguagem utilizou-se algumas "bibliotecas externas", a mais importante delas: TensorFlow ma versão 2.5.0.

Figura 4 – Screenshot da tela principal do Pycharm



Fonte: Autor

Com o decorrer do projeto, foi estudada a ideia de migrarmos o projeto para o ambiente *"Google Colab"*. O *"Colab"*(ou *"Colaboratory"*) é um sistema oferecido pela Google. No ambiente *"Colab"* são oferecidos, "CPUs","GPUs"e até "TPUs"gratuitos.

Por seu processamento, ser efetuado exclusivamente em ambiente virtual, tornando viável o treinamento paralelamente e também podendo realizar de qualquer máquina com acesso à rede.

3.2 Tratamento das bases de dados

3.2.1 Tratamento da base de dados BreakHis

Após coletados os dados e "descompactados"em máquina própria, a primeira base de dados disponibilizada estava dividida em primeiro nível por grau de amplitude e em um segundo nível, entre benigno e maligno. Primeiramente realizarei uma clusterização das imagens, separando-as por suas respectivas classes. Para isso, precisaremos analisar a composição dos dados, demonstraremos um exemplo aqui, utilizando as imagens com fator de ampliação de 40x.Será mantido um *buckup* da estrutura original, para evitar uma

perda de tempo desnecessária. Separando as imagens dessa forma, poderemos ter uma noção real do nosso *dataset*, sabermos a quantidade real de cada classe e a quantidade em relação aos outros. Para um tratamento mais fácil das imagens, será necessário saber como elas estão dispostas no repositório. Realizaremos isso usando o código da imagem 5.

Figura 5 – Código para mostrar nomes dos arquivos (Jupyter notebook)

```
import os

def mostrar_arquivos(path):
    for p, _, files in os.walk(os.path.abspath(path)):
        for file in files:
            print(os.path.join(p, file))
```

Fonte: Autor

Apos listar os arquivos, foi decido mudar suas nomenclaturas, excluindo valores que possa, se tornar desnecessário e também obsoletos.Podemos ver como é composto os nomes das imagens, como podemos ver na imagem 6. Por exemplo: SOB-B-F-14-14134E-100-004.png, é a quarta imagem, com ampliação de 100X, com a categoria de tumor "benigno", com a classificação de "Fibroadenoma", original do slide 14-14134 sendo coletado utilizando a técnica SOB.

Figura 6 – Amostra com nomes de imagens (Jupyter notebook)

```
SOB_B_A-14-22549AB-40-002.png
SOB_B_A-14-22549AB-40-003.png
SOB_B_A-14-22549AB-40-004.png
SOB_B_A-14-22549AB-40-005.png
SOB_B_A-14-22549AB-40-006.png
SOB_B_A-14-22549AB-40-007.png
SOB_B_A-14-22549AB-40-008.png
SOB_B_A-14-22549AB-40-009.png
```

Fonte: Autor

Como observado na imagem 6, o nome do arquivo leva a seguinte composição:

Procedimento de biopsia : SOB;

Categoria do tumor : M(Maligno) ou B(Benigno);

Classificação do tumor : A|F|TP|AT|DC|CL|MC|CP;*

Ano : Ano do exame;

ID do exame : Composta por números;

Grau de Ampliação : 40x|100x|200x|400x;

SEQ : Identificação do exame;

Em outro momento, foi realizada nova análise das imagens, visando eliminar esses dados, foi utilizado o algorítimo contido na imagem 7 , para a transformação dos nomes das figuras. Como podemos observar os nomes serão compostos por um número e sua categoria. Como demonstrado nas imagens 8,9,10 e 39.†

* "adenoses (A)", "Fibroadenoma (F)", "tumor Phyllodes (TP)"e "adenoma tubu-lar (AT)"para câncer benigno e "carcinoma (DC)", "carcinoma lobular(CL)", "Carcinomamucinoso (MC)"e "Carcinoma papilífero (CP)"para câncer maligno. † Todas as amostras foram tiradas aleatoriamente do *Dataset*

Figura 7 – Função para mudar o nome das imagens (Jupyter notebook)

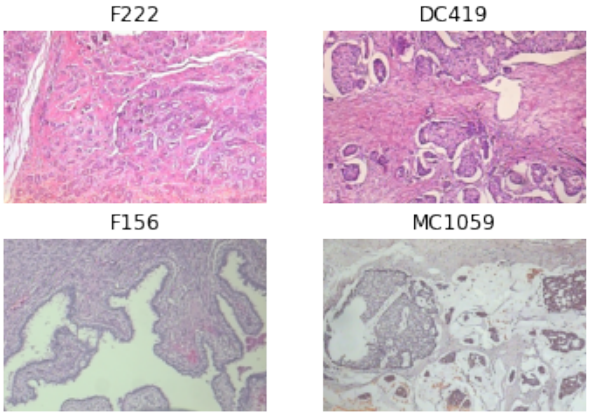
```
import os

dicionario_cancer = {'SOB_B_A': 'A', 'SOB_B_F':'F','SOB_B_PT':'PT','SOB_B_TA':'TA','SOB_M_DC': 'DC',
                     'SOB_M_LC':'LC','SOB_M_MC':'MC','SOB_M_PC':'PC'}

def mudar_label(path):
    x = 0
    for p, _, files in os.walk(os.path.abspath(path)):
        for file in files:
            x += 1
            os.rename(p+"\\"+file,p+"\\"+dicionario_cancer.get(file.split('-')[0], 'Contato não encontrado')+f'{x}.png')
```

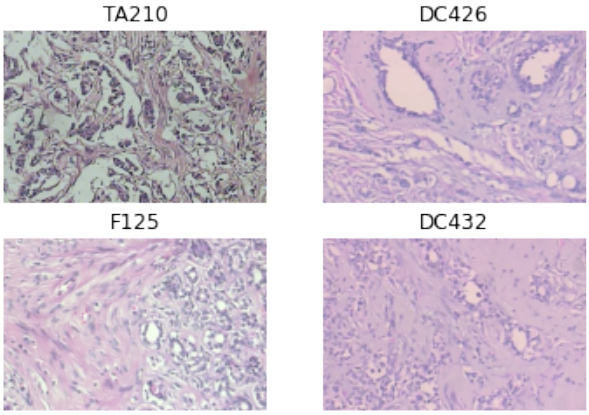
Fonte: Autor

Figura 8 – Amostra com novos nomes, ampliação 40x



Fonte: Autor

Figura 9 – Amostra com novos nomes, ampliação 100x

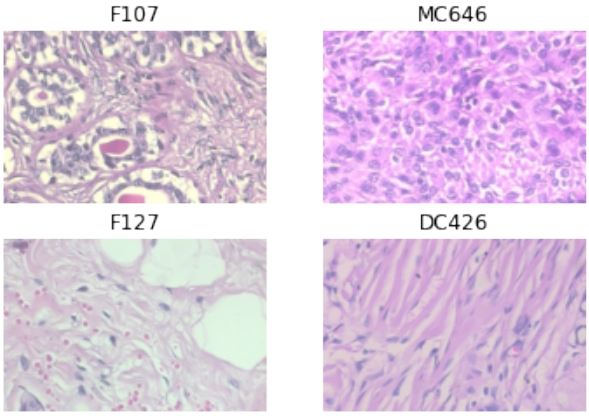


Fonte: Autor

As imagens originalmente, possuem a seguinte resolução 460 x 700 x 3, ou seja, 460 de altura, por 700 de largura e com 3 canais de cores, ou seja, colorida. Após analisar essa informação, foi decidido que seria ideal, a diminuição dessas resoluções. Além de uma melhoria no processamento, se enquadraria nas arquiteturas utilizadas.

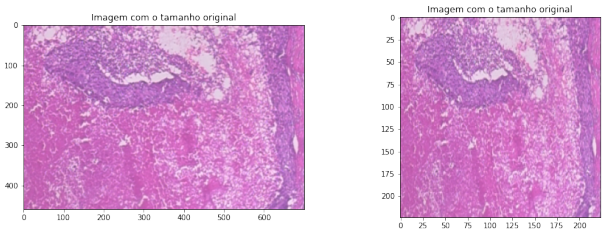
Primeiro passo será remodelar as imagens nas seguintes medidas que avaliarei conforme o andamento do nosso algoritmo, tentando balancear a taxa entre Aprendizagem x Velocidade. A primeira proporção utilizada será 224 x 224 x 3. Agora nos teremos as seguintes proporções: 224 de altura, por 224 de largura e 3 de "channels"(colorida).

Figura 10 – Amostra com novos nomes, ampliação 200x



Fonte: Autor

Figura 11 – Amostra com novos tamanhos das imagens



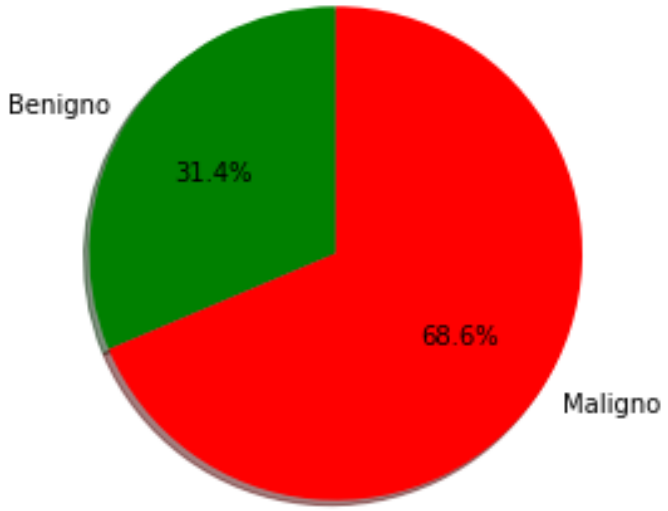
Fonte: Autor

3.2.2 Tratamento da base de dados Breast invasive ductal car-cinomas

Após coletados os dados e "descompactados"em máquina própria, a segunda base de dados disponibilizada também estava dividia, porem, esse nosso "dataset"vem em primeiro grau.dividida em 3 grupos, de acordo com sua "pontuação mitótica". A "pontuação mitótica"não será usada diretamente na em nosso algorítimo. Em um segundo nível, vemos que nossa base de dados é dividida em seus graus de amplitude.

O motivo para o uso da nova fonte de dados, é uma tentativa de realização de um balanceamento natural da base de dados. A Base de dados "BreakHis", é encontrada numa proporção aproximada de 69% para maligno enquanto apenas 31% para benignos. Como podemos ver nos gráficos 3.

Gráfico 3: Distribuição de categorias de cancer



Fonte: Autor

Utilizando o algorítimo da figura 5, conseguiremos ver a nomenclatura aplicada nessas imagens, também seguem um padrão logico:

Numero Simples : número inteiro;

Categoria do tumor : BC(Benigno)-IDC;

Numero do arquivo : número de arquivos de patologia;

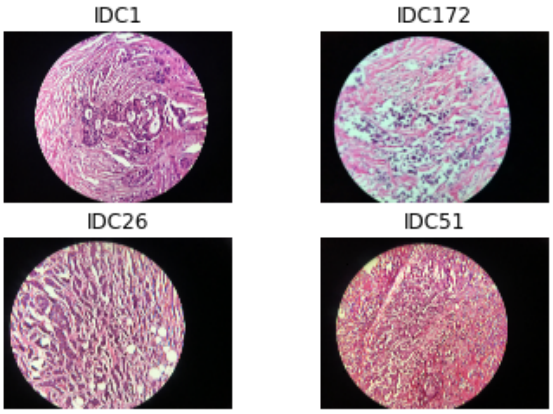
Grau de Ampliação : 40x|100x|200x|400x;

pontuação mitótica : número de figuras mitóticas encontradas em 10 campos consecutivos de alta potência (HPF)*;

Uma das principais dificuldade que encontraremos ao usar essa segunda base é a formatação de cada imagem. Como podemos ver na imagem 41 as imagens são formatadas em forma circular, diferentemente da nossa primeira base que são de formatos quadriculares. Essa diferença pode se tornar uma pode vir a se tornar uma ajuda, pois dados de fontes diferentes, sempre tenderam a vir em formatos diferentes, sendo assim, a BIAS de nossa Rede Neural, conseguirá se sair bem com formatações não padronizadas de imagens.

* A pontuação mitótica é determinada pelo número de figuras mitóticas encontradas em 10 campos consecutivos de alta potência (HPF) na parte mais mitoticamente ativa do tumor. Escores de 3 a 5, bem diferenciados (grau I); escores de 6 a 7, moderadamente diferenciados (grau II); e pontuações de 8 a 9, pouco diferenciadas (grau III)

Figura 12 – Amostra da base dados Breast invasive ductal carcinomas (Jupyter notebook)



Fonte: Autor

As imagens originalmente, possuem a seguinte resolução 3024 x 4032 x 3, ou seja, 3024 de largura, por 4032 de altura e 3 de "channels". Após analisar as imagens, percebi a existência de informações adicionais que não são necessárias para nosso treinamento. Por esse motivo, decidi realizar cortes, centrais nas imagens para melhor aproveitamento. Utilizando a parte do código presente na figura **??**.

Segundo passo será remodelar as imagens em para a proporção utilizada, pelo nosso modelo, que será 224 x 224 x 3, ou seja, 224 de altura, por 224 de largura com 3 canais de cores.

4 ANÁLISE E EXPLORAÇÃO DOS DADOS

> *"A informação é o petróleo do século XXI e a análise de dados é o motor de combustão."** - (Peter Sondergaard, Tradução Nossa)

Para a realização da análise e exploração de dados será utilizado *scripts* em Python e com apresentação no Jupyter Notebook. Os principais módulos utilizados serão: "matplotlib", "seaborn", "cv2", "numpy", "os"e "Pandas".

matplotlib e seaborn : utilizadas para a criação de gráficos;

cv2 : utilizado para leitura e manipulação das imagens;

numpy : utilizado para manipulações dos vetores de imagens;

os : biblioteca natural do python, utilizada para manipular "Paths".

Pandas : uma biblioteca de software criada para a linguagem Python para manipulação e análise de dados e dataframes.

Primeiramente faremos a importação das bibliotecas utilizadas, como podemos ver no código descrito na figura 13.

Figura 13 – Código de importação dos módulos (Google Colab)

```
import seaborn as sns
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Fonte: Autor

Em seguida, realizarei a criação de um dataframe contendo informações primordiais do nosso dataset, informações essas que serão usadas para uma análise exploratória do nosso DataSet. O nosso dataframe terá as seguintes colunas: "FikePath", "Type", "classes","Magnification"e o "Dataset".

FikePath : Caminho onde a imagem foi salva;

Type : Categoria do câncer [B|M];

classes : Classe do tumor [A|F|TP|AT|DC|CL|MC|CP];†

Magnification : grau de ampliação [40|100|200|400];

Dataset : DataSet pertencente "Teste(Te)"ou "Treino(Tr)".

---

* "Information is the oil of the 21st century, and analytics is the combustion engine." † "adenoses (A)", "Fibroadenoma (F)", "tumor Phyllodes (TP)"e "adenoma tubu-lar (AT)"para câncer benigno e "carcinoma (DC)", "carcinoma lobular(CL)", "Carcinomamucinoso (MC)"e "Carcinoma papilífero (CP)"para câncer maligno.

Serão utilizados para criação do dataframe as seguintes funções apresentadas nas figuras 14, 42, 43,44 e 45.

Figura 14 – Código de criação do *"DataFrame"* (Google Colab)

```
def montar_data_frame(path):
    nome_file, file_path = montar_nome(path)
    classes = selecionar_classe(nome_file)
    tipos = pegar_tipos(file_path)
    ampliacaos = pegar_ampliacao(file_path)
    origem = pegar_origem(file_path)
    data = {'FikePath': file_path, 'Type': tipos, 'classes':classes,'Magnification':ampliacaos, 'DataSet':origem}
    df = pd.DataFrame.from_dict(data, orient='index')
    df = df.transpose()
    return df
```

Fonte: Autor

Podemos ver um exemplo do *"DataFrame"* na figura 15.

Figura 15 – Amostra do *"DataFrame"* criado (Google Colab)

	FikePath	Type	classes	Magnification	DataSet
2821	/content/drive/MyDrive/TCC/Imagens/200x/200x_t...	M	DC	200x	Te
1676	/content/drive/MyDrive/TCC/Imagens/400x/400x_t...	B	DCI	400x	Te
825	/content/drive/MyDrive/TCC/Imagens/400x/400x_t...	M	DC	400x	Te
1373	/content/drive/MyDrive/TCC/Imagens/400x/400x_t...	B	PT	400x	Te
375	/content/drive/MyDrive/TCC/Imagens/400x/400x_T...	M	PC	400x	Tr
2663	/content/drive/MyDrive/TCC/Imagens/200x/200x_t...	M	DC	200x	Te
2801	/content/drive/MyDrive/TCC/Imagens/200x/200x_t...	M	DC	200x	Te
2840	/content/drive/MyDrive/TCC/Imagens/200x/200x_t...	M	DC	200x	Te
979	/content/drive/MyDrive/TCC/Imagens/400x/400x_t...	M	DC	400x	Te
3109	/content/drive/MyDrive/TCC/Imagens/200x/200x_t...	B	F	200x	Te

Fonte: Autor

Na figura 16, podemos confirmar que o nosso dataframe segue a composição planejada.

Figura 16 – Valores únicos de cada coluna (Google Colab)

```
[5] df_geral['Type'].unique()
    array(['M', 'B'], dtype=object)

[6] df_geral['classes'].unique()
    array(['DC', 'LC', 'MC', 'PC', 'A', 'F', 'PT', 'TA', 'DCI'], dtype=object)

[7] df_geral['Magnification'].unique()
    array(['400x', '200x'], dtype=object)

[8] df_geral['DataSet'].unique()
    array(['Tr', 'Te'], dtype=object)
```

Fonte: Autor

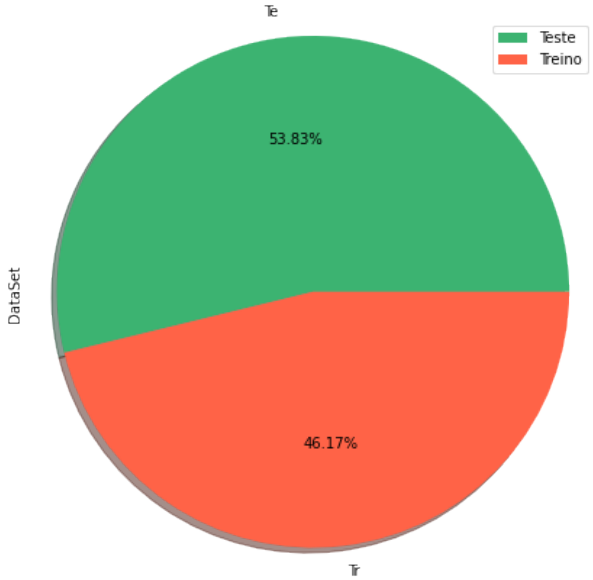
Essa abordagem se mostrou suficiente para garantirmos um estudo mais detalhado do processo de exploração de dados.O primeiro passo na análise exploratória de dados foi entender a composição do nosso *"dataframe"* como mostrado na figura 15 e principalmente na figura 16.

Observamos também, que nosso *"dataframe"* segue o padrão descrito na nossa lista, e com isso poderemos primeiro analisar a divisão do nosso *"dataset"* entre dados

que usaremos para Treinamento e dados que utilizaremos posteriormente para testarmos nosso algorítimo.

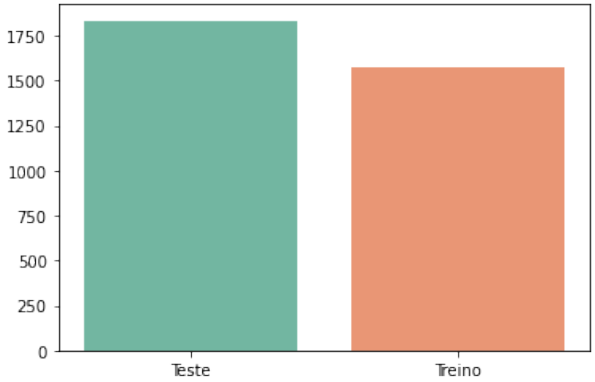
Nosso *"dataset"* segue uma divisão quase que igual entre dados de treino e dados utilizados para teste. Seguindo uma proporção de 47% de para treino e um volume um pouco maior de teste, sendo ele equivalente a 53% do nosso dataset, como demonstrado nos gráficos 4 e 5

Gráfico 4: Distribuição de Treino e Teste no *Dataset*



Fonte: Autor

Gráfico 5: Quantidade de dados entre Treino e Teste no *Dataset*

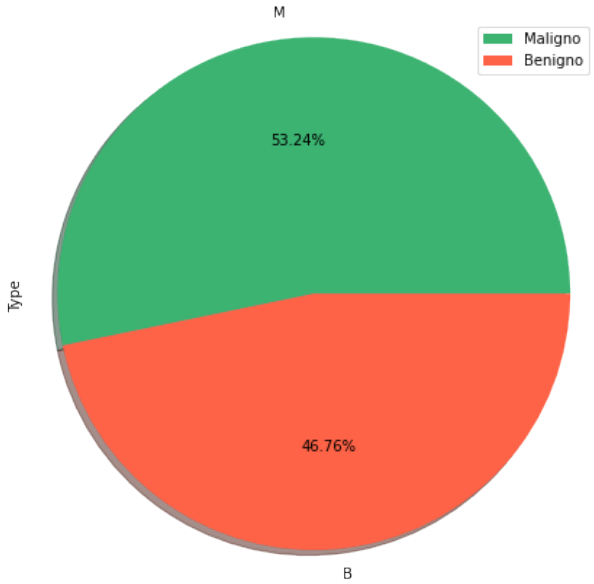


Fonte: Autor

Observamos que antes do *merge* entre as bases de dados *"BreakHis"* com a *"Breast invasive ductal carcinomas"*, era possível ver uma proporção de 31,4% para cânceres Benignos contra 68,9% para cânceres malignos. Com o *merge* das duas bases de dados, conseguimos diminuir um pouco essa proporção, agora encontramos a seguinte proporção

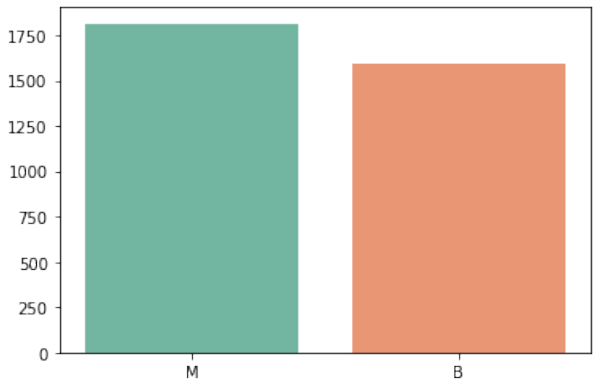
46,17% para cânceres Benignos contra 53,43% para cânceres malignos, como observado nos gráficos 6 e 7.

Gráfico 6: Distribuição de Beigno e Maligno no Dataset



Fonte: Autor

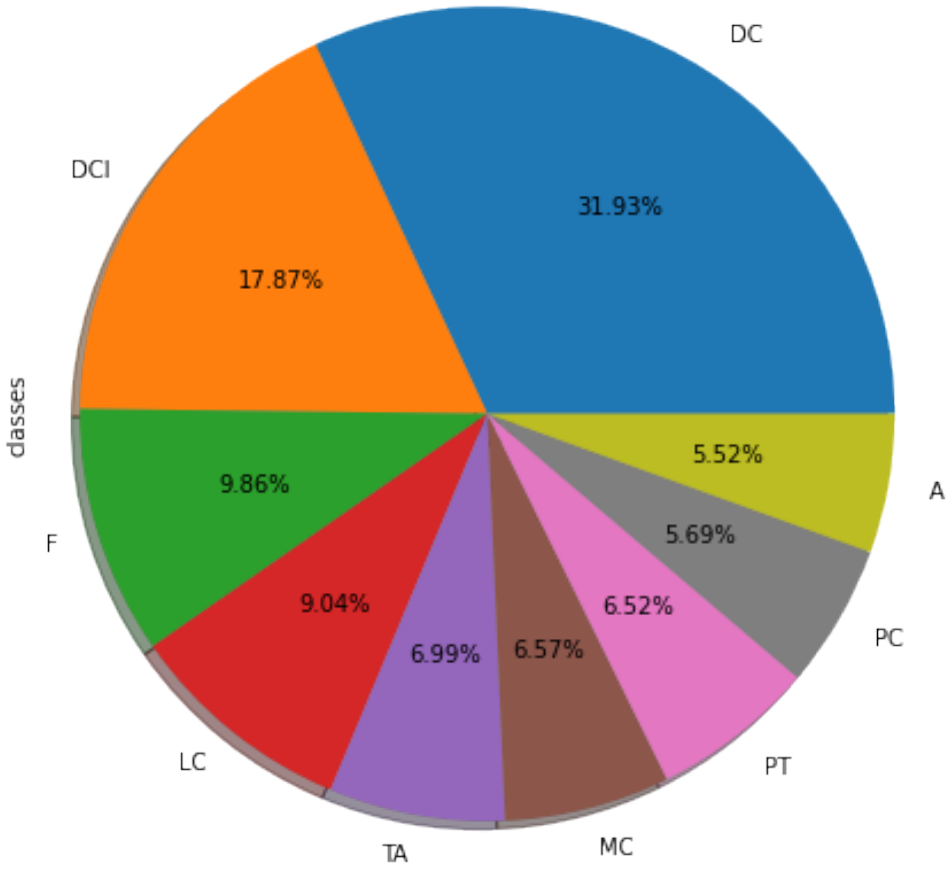
Gráfico 7: Quantidade de dados entre Beigno e Maligno no Dataset



Fonte: Autor

Ao observamos o gráfico 8, podemos perceber a distribuição em porcentagem das classes de cânceres conforme. Podemos ver também, que contemos uma quantidade superior do câncer que contem a classe "carcinoma (DC)". Estando em maior quantidade na base de dados, de ampliação "200x"e "400x".

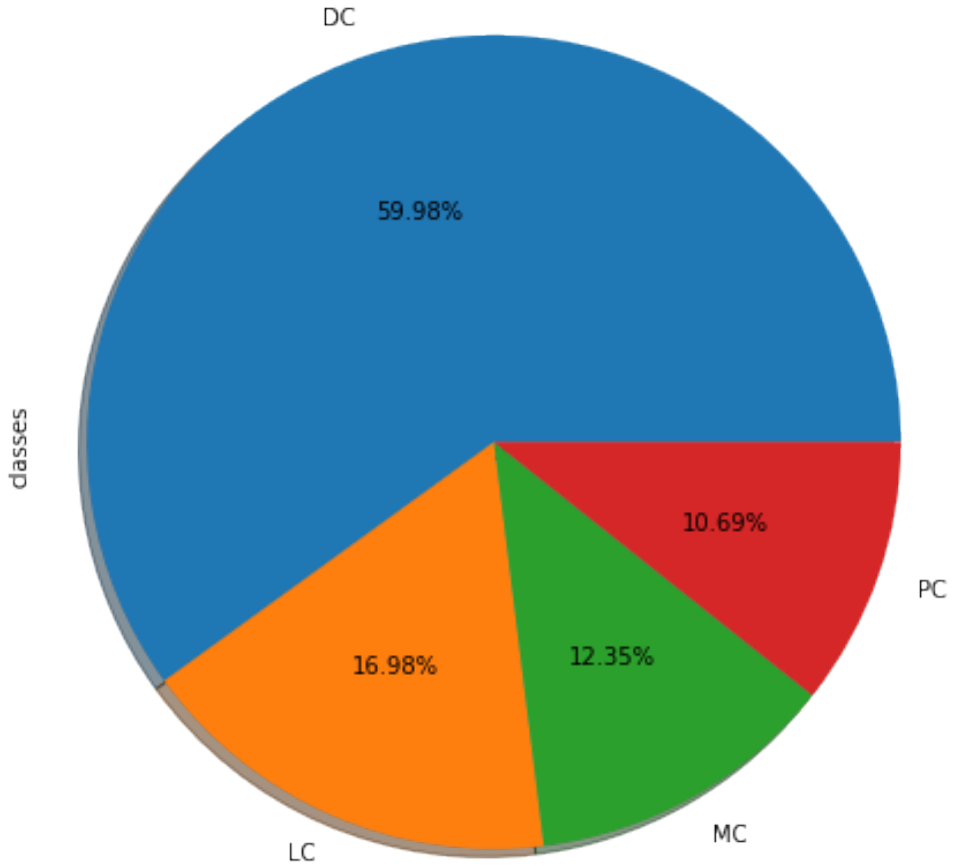
Gráfico 8: Distribuição em porcentagens das classes



Fonte: Autor

Podemos ver também essa superioridade percentual no gráfico 9, quando olhamos apenas as classes que iremos de fato prever.

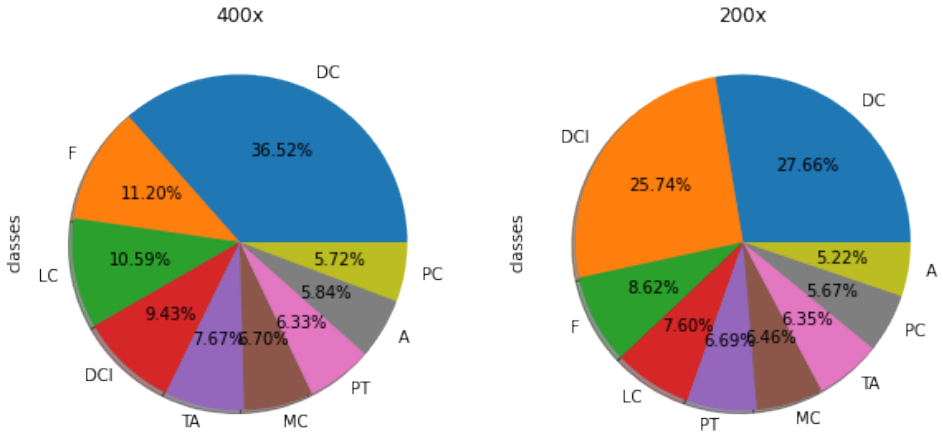
Gráfico 9: Distribuição em porcentagens das classes referente aos tipos Malignos



Fonte: Autor

Quando observamos a nossa massa de dados, percebemos um leve desbalanceamento para a classe "carcinoma (DC)"como visto no gráfico 10. O balanceamento da nossa base de dados será feita com o treinamento da nossa rede neural. O balanceamento será feito pesando no mínimo de perda de dados e o máximo de desempenho da nossa rede.

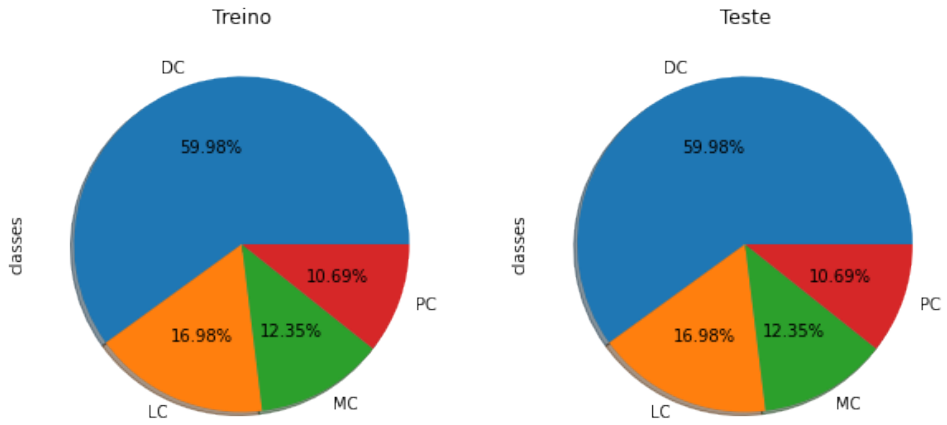
Gráfico 10: Distribuição em valores percentuais das classes por *"Magnification"*



Fonte: Autor

Podemos ver também uma proporção parecida entre os dados de treino e de teste. Como podemos ver no gráfico 11

Gráfico 11: Distribuição em valores percentuais das classes base de treino e teste



Fonte: Autor

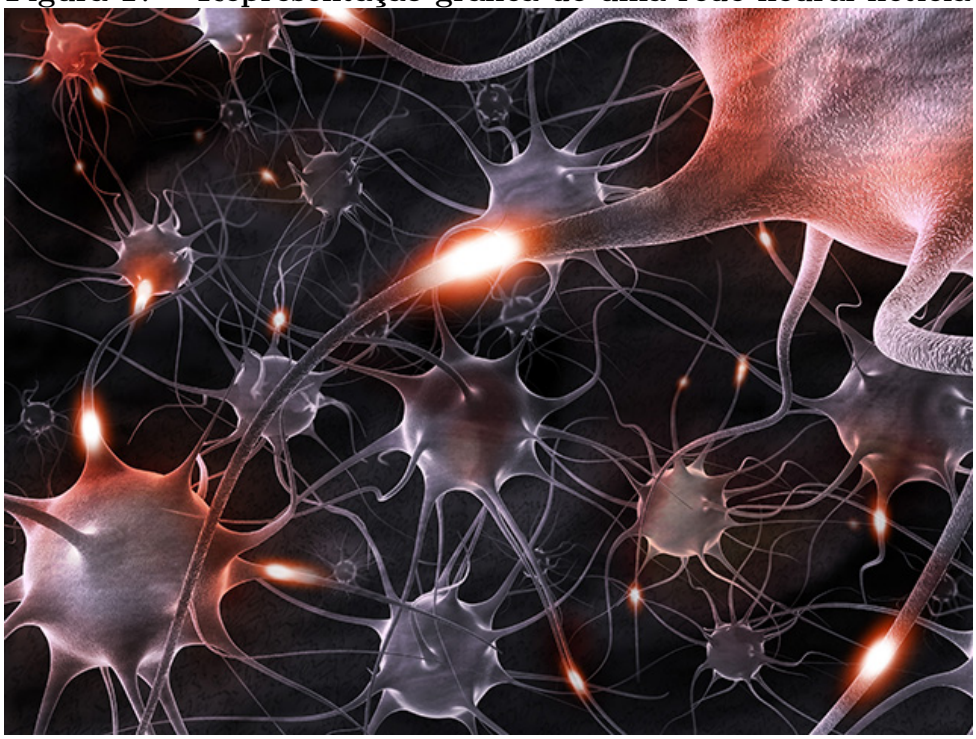
5 CRIAÇÃO DE MODELOS DE MACHINE LEARNING

Neste capítulo será desenvolvido o modelo de "Rede Neural Convolucional"o qual iremos realizar classificação de nossa base de dados. Primeiro teremos alguns conceitos teóricos sobre Aprendizagem profunda(Deep learning) e Rede neural convolucional (Convolutional neural network). As apresentações teóricas serão necessárias, para uma boa compreensão deste trabalho. Primeiro irei explicar o que seria uma rede neural natural que serviu de insparação para a tecnica que utilizaremos.

5.1 Rede Neural Natural (R.N.A.)

No nosso dia a dia, conseguimos realizar ações, como andar, comer, conversar, dirigir e etc.. A capacidade de realizarmos tais ações está ligada a presença de um rede neural. Essa composição, começou a ser estudada em 1902. Podemos ver um exemplo gráfica de uma rede neural na figura 17

Figura 17 – Representação gráfica de uma rede neural fictícia



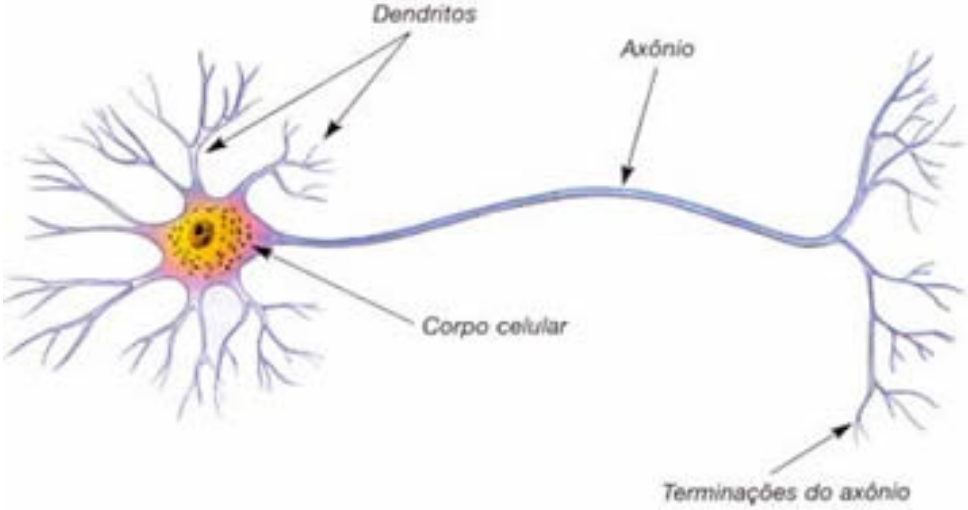
Fonte: Anônimo, 2014

Uma rede neural natural é composta por milhares de células nervosas, denominados neurônios. São células do sistema nervoso que estão relacionadas com a propagação de impulsos elétricos. Um neurônio é considerado a unidade básica desse sistema.(ANôNIMO, 2021).

Os Neurônios são compostos por 3 partes básicas: os dendritos, o axônio e o corpo celular (ANôNIMO, 2021). Como podemos ver representado na figura 18

Dendritos são as terminações dos neurônios, finos e geralmente ramificados, que

Figura 18 – Representação gráfica de um neurônio fictício



Fonte: Anônimo, 2021

conduzem os estímulos captados do ambiente ou de outras células em direção ao corpo celular.(ANôNIMO, 2021)

Axônio é um prolongamento fino, geralmente maior que os dendritos, que possuem a função de transmitir para outras células os impulsos nervosos(impulsos elétricos) provenientes do corpo celular.(ANôNIMO, 2021)

O corpo celular, é a principal parte e mais volumosa do neurônio. Nela é localizada o núcleo e a maioria das estruturas citoplasmáticas.(ANôNIMO, 2021)

No ano de 1949, Donald Hebb escreveu uma obra que apontou que caminhos neurais são fortalecidos cada vez que são usados, um conceito essencial sobre a forma que os seres humanos aprendem. Segundo Donald Hebb, Se dois neurônios disparem em simultâneo, a conexão entre eles é melhorada.(DATA SCIENCE ACADEMY, 2021)

5.2 Rede Neural Artificial (R.N.A.)

Inspirado na arquitetura do cérebro humano, as redes neurais exibem algumas características do cérebro humano, como a aprendizagem de padrões complexos e a generalização da informação aprendida. E assim como as redes neurais naturais as redes neurais artificiais, funcionam de forma parecida, ou seja, há um recebimento de estímulos, o processamento de uma função e o retorno determinado.

Em 1943, o matemático Walter Pitts e neurofisiologista Warren McCulloch (MCCULLOC; PITTS, 1943), tiveram a ideia de criar o artigo *"A Logical Calculus of the Ideas Immament in Nervous Activity"*, onde eles explicariam uma rede neural simples com a ajuda de circuitos elétricos .Essa ideia deu origem a criação do algoritmo de lógica de limiar (threshold logic).(DATA SCIENCE ACADEMY, 2021)

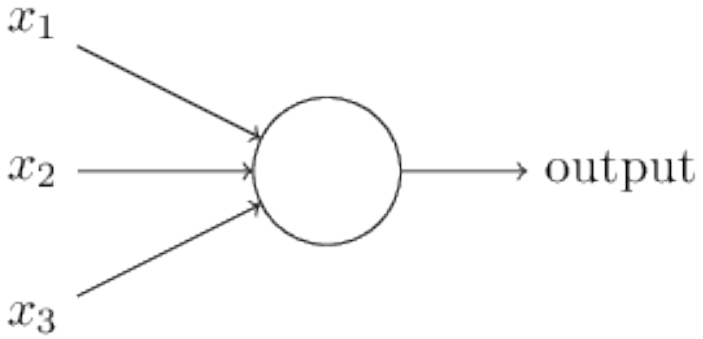
Em 1959 inciou-se os estudos sobre a aprendimento de uma R.N.A. feito por

Donald Hebb (HEBB, 1950), no artigo de nome *"The organization of behavior: A neuropsychological approach"*, nesse artigo, segundo (HEBB, 1950) a aprendizagem de rede é realizada através da inclusão de pesos em cada conexão dos neurônios.

Para Frank Rosenblatt (1958) os modelos de arquiteturas onde pesos sofriam ajuste em seus valores, poderiam ser treinados para classificação de padrões. Esse modelo de arquitetura ficou conhecido como *"perceptron"*.

O perceptron, em um resumo geral é equivalente ao neurônio. Assim como um neurônio, o perceptron possui uma estrutura um pouco mais simples, porem com 3 camadas: entrada equivalente ao Dendrito, oculta(ou intermediário) equivalente ao corpo celular e a saída que seria equivalente ao Axônio. Podemos ver um exemplo gráfico nas figuras 19.

Figura 19 – Representação gráfica de perceptron



Fonte: Data Science Academy, 2021

Um perceptron é um modelo matemático que recebem diversas entras, como no exemplo da figura 19, vemos um perceptron de 3 entradas: x1,x2,x3 que tem como resposta uma única saída binaria.(DATA SCIENCE ACADEMY, 2021)

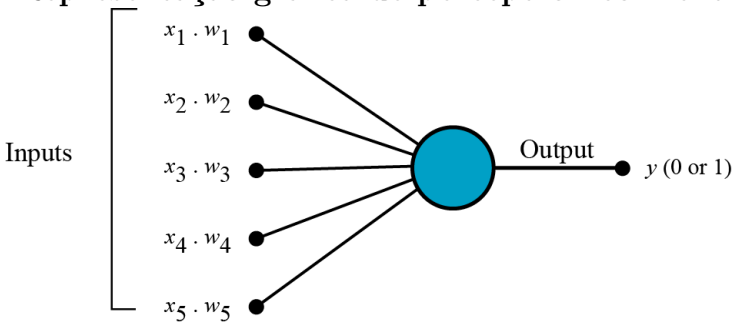
Rosenblatt (1958) propôs uma regra simples para calcular essa saída. Ele introduziu a ideia de pesos, w1,w2, .., wx . Os pesos são compostos de números reais. A saída de um neurônio, 0 ou 1, é determinada pela soma ponderada, $\sum$**jwjxj** , menor ou maior do que algum valor limiar(*threshold logi*).O threshold é um número real que é um parâmetro do neurônio.(DATA SCIENCE ACADEMY, 2021)

Podemos ver essa ideia em forma algébrica mais precisamente na seguinte função :

$$\text{output} = \begin{cases} 0 & \text{se } \sum \text{jwjxj} \leq \text{threshold} \\ 1 & \text{senão} \end{cases}$$

Como vemos, um Perceptron segue o modelo *"feed-forward"*, ou seja, ele é unilateral. As entradas são enviadas para o neuronio , processadas e resultam em uma saída específica, como podemos ver nas figuras 19 e com um grau maior de explicação na figura 20, onde podemos ver as entradas já multiplicadas pelos pesos, pesos esses que podem ser predefinidos ou definidos conforme o treinamento e podemos também ver a saída, que sempre será 0 ou 1.(DATA SCIENCE ACADEMY, 2021)

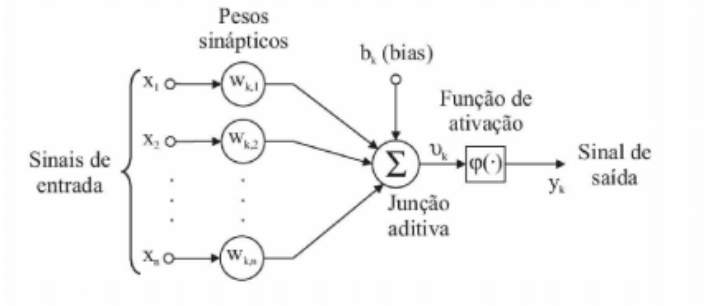
Figura 20 – Representação gráfica de perceptron com entrada e saída



Fonte: Data Science Academy, 2021

Podemos ver na figura 21, o modelo de um neurônio(Perceptron) de McCulloch e Pitts, considerado a arquitetura de uma "RNA" mais simples.

Figura 21 – Modelo de Perceptron não linear



Fonte: HAYKIN, 2007

Como visto na figura 21,um perceptron é composto por alguns elementos, já descritos no trabalho e também : Sinapse ou Peso (w), função de ativação e em alguns casos, é o *bias*.

Os Pesos(ou Sinapses) são usados para guardar o conhecimento. Esses valores são atribuídos e multiplicados pelos valores de entrada, para melhor ajuste ao desejado.

A função de ativação é responsável pela transformação não linear ao longo do sinal de entrada. Esta saída já transformada é então enviada para a camada posterior como entrada. Quando não implementamos uma função e ativação, a rede neural pode apenas obter resultados lineares.

Bias exerce a funcionalidade de um "interceptor"adicionado em uma equação linear. Ele é um parâmetro adicionada na Rede neural cuja função é ajustar a saída junto a fonte ponderada de entradas do neurônio.

A parti da criação do perceptron foram criadas as redes neurais artificiais (RNA).

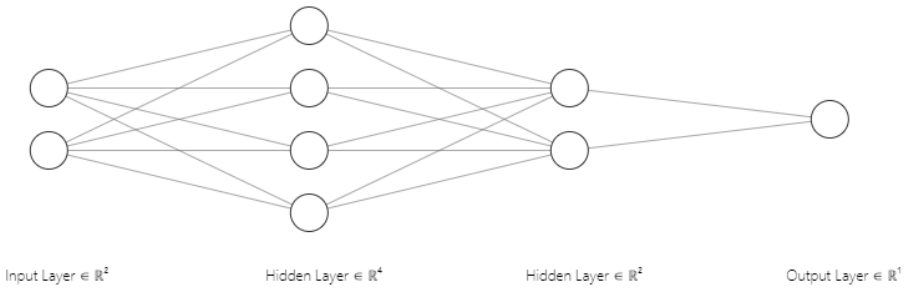
Que são um conjunto de perceptrons que seguindo a ideia das redes neurais naturais, são conectadas entre si.

5.2.1 Rede Neural Artificial Multicamadas ou MLPs

Rede Neural Artificial Multicamadas(ou Perceptron multicamadas) é uma arquitetura onde os neurônios são organizados em duas ou mais camadas.Sendo elas : "Camada de entrada"(*input layer*), onde os seus neurônios recebem o nome de "Neurônios de Entrada". A camada do oculta (hidden layers), cada neurônio dessa camada é conectado com todos os outros da próxima camada. Ela também tem a função de executar cálculos mais complexos e auxiliar a última camada. A terceira e ultima classe é a "Camada de saída"(*output layer*), que possuem os neurônios de saída.

Podemos ver na figura 24 um exemplo de Rede Neural de Múltiplas Camadas.

Figura 22 – Diagrama de uma MLPs



Fonte: Autor usando http://alexlenail.me/

Quanto a forma com que os neurônios artificiais se conectam, redes neurais podem ser classificadas em 3 categorias:

Redes neurais feedforward : Esse é o tipo mais comum de arquitetura em aplicações praticas. Nessa arquitetura a primeira camada é chamada "Entrada"e a última camada é chamada "Saida". Se houver mais de uma camada oculta, ela é chamada redes neurais profundas(ou Deep Learning). As atividades dos "neurônios"são uma função não-linear das atividades da camada anterior;

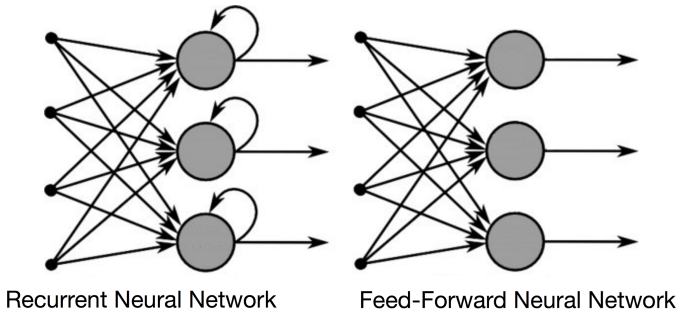
Redes neurais recorrentes : As redes recorrentes tomam como entrada, não somente os valores e os pesos, mas sim o que perceberam anteriormente no tempo.A decisão de uma rede recorrente alcançada na etapa de tempo t-1 afeta a decisão que alcançará um momento mais tarde na etapa de tempo t;

Redes neurais simétricas : Ela é semelhante à Rede Recorrente, porem as conexões entre as unidades são simétricas(Recebem o mesmo peso em ambas as direções). Se tornando uma rede mais fácil de ser analisadas.

(DATA SCIENCE ACADEMY, 2021)

Podemos ver na figura 23 uma representação grafica entre a "Rede neural feed-forward"e a "Rede neural recorrente".

Figura 23 – Diagrama de uma "Rede neural feedforward"e uma "Rede neural recorrente"



Fonte: Data Science Academy, 2021

5.2.2 Funções de Ativação

As "Funções de Ativação"tem um papel importante nas RNAs, já que elas definem saída de cada nó dado o conjunto de entradas.Quando não temos uma função de ativação, os pesos e *bias* transforma a saída em linear.

As principais funções de ativação são: Linear(ou *Linear activation*), Ativação Sigmoide, Softmax , ReLu (ou *Rectified linear unit*) e a Tangente Hiperbólica (ou TanH).

A função Linear, é a função de ativação mais básica porque ela não altera a saída do perceptron. Sendo mais utilizadas em RNAs de regressão. A função algébrica referente a essa função é a:

$$\phi(\mathrm{x}) = \mathrm{x}$$

A função de ativação Sigmoide é utilizada por redes que usem propagações positivas, que precisam de uma saída > 0. Ela é eficiente nas questões de não-linearidade.Essa característica aumenta o custo computacional.A função algébrica referente a essa função é a:

ϕ(x) = $\frac{1}{1+\varrho^{-}x}$

A função de ativação Softmax, é uma função sigmoide. Ela força a saída de uma rede neural a representar a probabilidade dos dados serem de uma das classes definidas. Ela transforma as saídas para cada classe para valores entre 0 e 1 e também divide pela soma das saídas.A função algébrica referente a essa função é a:

σ(x)j = $\frac{\varrho^{-}zj}{\sum k\varrho^{-}zk}$

A função de ativação "ReLU"é a função mais popular e amplamente usada em projetos que tenham Redes Neurais. A função de ativação "ReLU"é uma função não linear, possibilitando ter varias camadas de neurônios ativadas por ela. A função algébrica referente a essa função é a:

f(x) = max (0,x)

A função de ativação "Tangente Hiperbólica"é a função mais próxima da função indenidade, sendo uma alternativa para ativar as camadas ocultas.A função algébrica referente a essa função é a:

f(x) = $2\sigma(2x) - 1$

5.2.3 Treinamento de uma Rede Neural

A faze de aprendizagem de uma rede neural(ou treinamento), consiste em ajustar os pesos, de modo a reduzir o erro observado na saída da rede. Segundo HAYKIN (2007)

as redes neurais são estimuladas pelo ambiente, sofrem modificações nos seus parâmetros livres e responde de uma maneira nova ao ambiente de acordo com suas modificações.

No treinamento é utilizado a técnica de aprendizagem supervisionada, onde é dado um conjunto de dados já rotulados é introduzido na rede neural. É definido um sinal de erro como sendo a diferença entre a resposta verdadeira e a resposta encontrada na saída da rede.

Não existe uma regra exata para definirmos o tempo de treinamento de cada rede, geralmente ela é definida antes do início do treinamento, utilizando uma quantidade máxima de ciclos(ou épocas). Para HAYKIN (2007) a fase de treinamento deve ser finalizada quando encontramos um valor aceitável para a taxa de erro.

5.2.4 Algoritmo Backpropagation

Para corrigir os pesos sinápticos, uma das técnicas populares é o *backpropagation*. Corrigindo os valores dos pesos pela diferença entre o valor encontrado na saída da rede com o valor esperado.

Criado por Rumelhart, Hinton e Williams (1986) o algorítimo *backpropagation*, possibilitava o treinamento de redes neurais com diversas camadas através da Retropropagação do erro. Segundo o Data Science Academy (2021) a técnica do algorítimo *backpropagation* pode deixar o treinamento da rede neural até dez milhões de vezes mais rápido.

"O objetivo do backpropagation é otimizar os pesos para que a rede neural possa aprender a mapear corretamente as entradas para as saídas."(DATA SCIENCE ACADEMY, 2021)

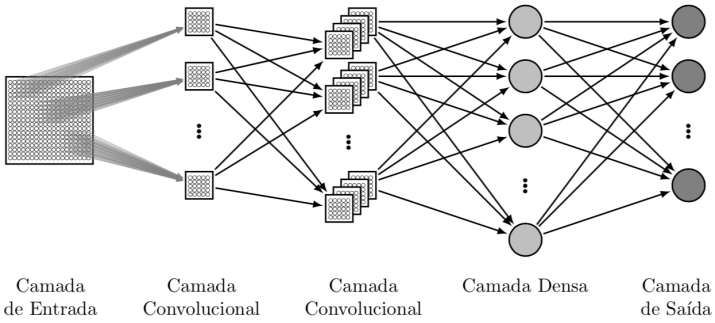
Segundo o Data Science Academy (2021) o algoritmo de backpropagation consiste em duas fases:

- Primeiramente é feito o *forward pass*, onde as entradas fazem o caminho através da rede é retornado uma sai.
- O segundo passo é feito o *backward pass*, onde será calculada a função de perda da última camada. Esse gradiente é usado para aplicar recursivamente a regra da cadeia (ou chain rule), onde iremos atualizar os pesos da Rede.

5.2.5 Redes neurais convolucionais ou CNN

A Rede Neural Convolucional (Convolutional neural network), é uma arquitetura de RNA, sendo uma variante da MLP(Rede Neural Artificial Multicamadas). As CNNs tem um uso mais voltado para fins de classificação, detecção e reconhecimento de imagens e também vídeos. A CNN tem como principal diferencial, a divisão em estágios, como visto na figura 24

Figura 24 – Arquitetura de uma Rede neural convolucional



Fonte: Sakurai, 2017

A CNN é composta de 4 camadas principais, sendo elas: A camada de entrada, camadas convolucionais, camadas densas(ou oculta) e a camada de saída. A camada convolucional, a mais importante dessa arquitetura, é responsável pela extração das principais características da imagem de entrada, essa ação é chamada "mapa de recurso"(ou feature map).

A primeira CNN foi criada por Yann LeCunn, para o reconhecimento de caracteres manuscritos, como, por exemplo, a interpretação de códigos postais. De maneira geral, a imagem é dividia em campos que alimentam a camada convolucional, que extrai então recursos da imagem.(MADHAVAN; JONES, 2017)

Segundo Samaya Madhavan (2017), a CNN é composta por 5 etapas principais, sendo elas:

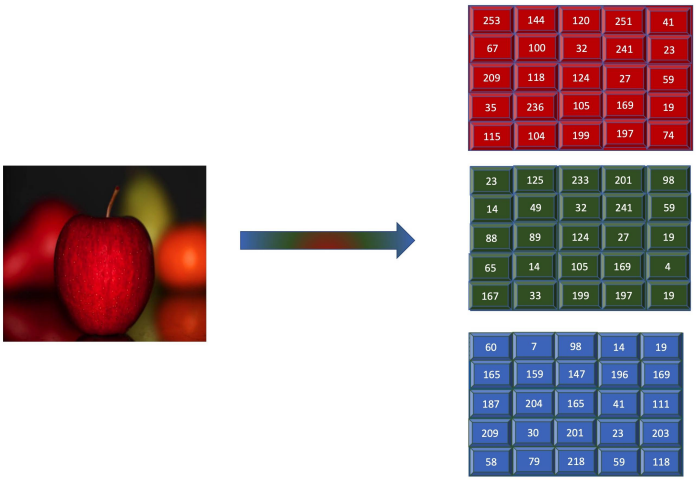
1. *Image channels*;
2. *Convolution*;
3. *Pooling*;
4. *Flattening*;
5. *Full connection.*

5.2.5.1 Image channels

Nessa primeira etapa, encontra-se o processo onde tornamos a imagem compatível com a com o algorítimo CNN. A imagem precisara ser transformada em píxeis. Cada píxel é mapeado usando números de 0 até 255, onde 0 significa branco e 255 preto.

Para imagens em preto e branco, será representada em uma matriz bidimensional. Cada célula da matriz irá conter seu valor correspondente. Para imagens coloridas, será representada em uma matriz tridimensional. Sendo cada *"channels"* da matriz uma cor, sendo elas: *red, green* e *blue* respectivamente. Poderemos ver um exemplo, na figura 26.

Figura 25 – Imagem representada em forma de matriz.



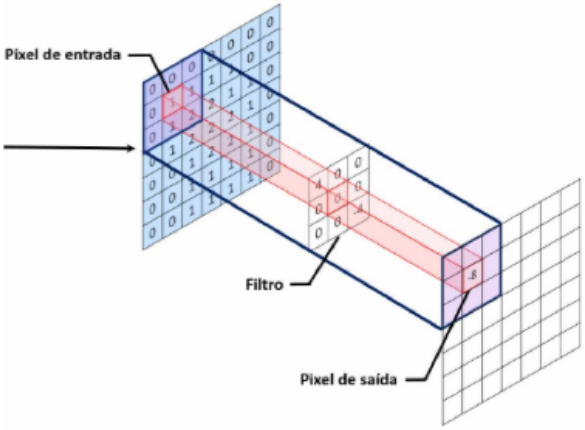
Fonte: Madhavan, 2017

5.2.5.2 Convolution

Segundo Samaya Madhavan (2017), a convolução é uma operação onde uma função modifica a forma de outra. Ainda segundo Samaya Madhavan (2017), as convoluções nas imagens são aplicadas por vários motivos, sendo eles: aumentar a nitidez, suavizar e intensificar.Na CNN, as convoluções são aplicadas para extrair os recursos proeminentes nas imagens.

A extração dos recursos principais da imagem é realizada através do uso de um filtro ou *Kernel*.Esse filtro consiste em uma matriz que será o conteúdo a ser extraído. O filtro(ou Kenel) é aplicado sobre a matriz de entrada, e a convolução resultante é uma matriz bidimensional que contém a correlação da imagem em relação ao filtro(ou Kenel). Segundo Samaya Madhavan (2017) A matriz de saída é conhecida como *"feature map"*.

Figura 26 – Representação da camada de convolução



Fonte: USHIZIMA, 2017

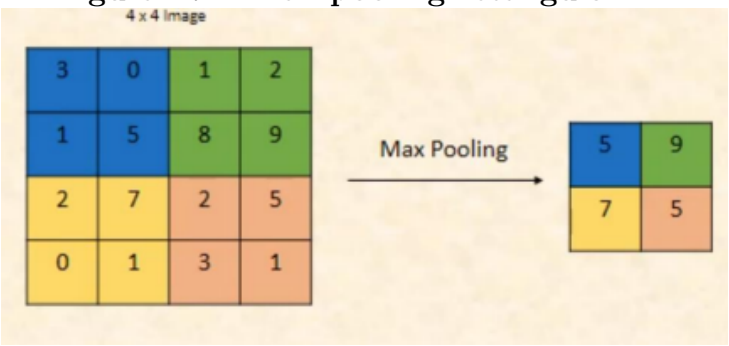
5.2.5.3 Pooling

Após as camadas convolucionais, com o objetivo de reduzir a dimensão espacial do mapa gerado, é aplicada a camada de *Pooling*. A camada possibilita que as camadas mais profundas possam ter uma quantidade maior de filtros sem existir um custo computacional muito elevado. Segundo Madhavan (2017) esse procedimento ajuda a compactar ainda mais a dimensão do *feature map*.

O *"Pooling"* pode ser obtido pegando o máximo, o mínimo ou a média dentro de um grupo de células. Sendo Conhecidos como: *Max Pooling, Min Pooling* e *Average Pooling*. Sendo os *Max Pooling* e *Average Pooling* mais utilizados.

No *Max Pooling*, o valor maximo da matriz é retornado para cada retangulo percorrido. Segundo essa técnica é a mais utilizada.Podemos ver um exemplo na figura 27.

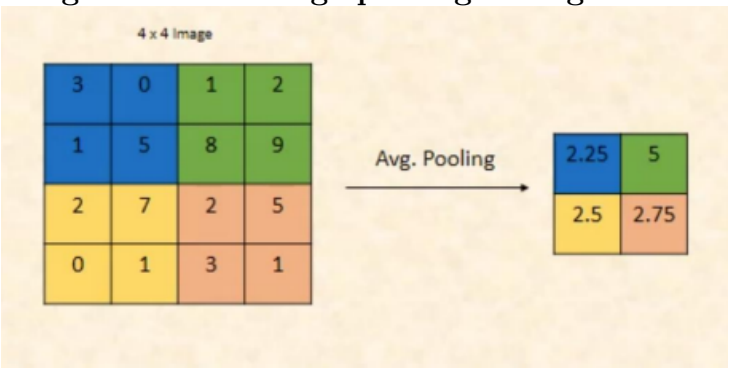
Figura 27 – Max pooling retângulo 2x2



Fonte: RIZWAN, 2018

No *Average Pooling*, a média dos valores no interior do retângulo é retornado para cada retângulo percorrido. Segundo GOODFELLOW Yoshua BENGIO (2016) essa técnica é a mais utilizada.Podemos ver um exemplo na figura 28.

Figura 28 – Average pooling retângulo 2x2



Fonte: RIZWAN, 2018

5.2.5.4 Flattening

A camada *Flattening* tem como objetivo transformar os resultados da CNN em dados compatíveis a um RNA(Rede Neural Artificial). Um RNA tem como entrada valores

em forma de vetor.

5.2.6 Transferência de aprendizado

A transferência de aprendizado consiste em reaproveitar apenas as camadas que deseja ("parte de baixo" ou *"floor"*) de uma rede já treinada, podendo aproveitar os pesos alinhados e com isso treinar apenas as camadas acrescentadas("parte de cima"ou *"top"*).

O processo de "transferência de aprendizado" é útil, pois economiza uma quantidade enorme de processamento, já que, o *"floor"* só processará cada padrão uma única vez. No nosso projeto utilizaremos esse processo em 3 redes diferentes, sendo elas ,"ResNet50", "VGG19"e "InceptionV3". Elas serão treinadas utilizando os pesos do treinamento "ImageNet".

5.3 Execução dos Modelos

Nessa seção apresentarei o "passo a passo"da execução dos modelos. Começaremos apresentando como foi realizado a criação do modelo *"ResNet50"*, logo em seguida será apresentado a criação do modelo *"VGG19"* e por fim a utilização da arquitetura *"InceptionV3"*.

Hoje sabemos que o quanto mais profundo uma modelo pode ir, melhor seu desempenho. A arquitetura *Resnet* foi avaliada como um dos trabalhos mais inovadores em visão computacional, conseguindo ganhar a competição na tarefa de classificação *"ILSVRC-2015"*.

Criado pelo Departamento de Engenharia, Universidade de Oxford(Department of Engineering Science, University of Oxford.) a "VGG19"é uma arquitetura de rede neural convolucional com 19 camadas de profundidade. 16 camadas convolucionais e 3 camadas totalmente conectadas.

A proposta desse modelo consiste em identificar e classificar as categorias de câncer e após sua classificação como "Maligno", utilizaremos a técnica de transferia de aprendizagem para a clusterização por classes.

Inicialmente, importaremos as Bibliotecas necessárias para a criação do modelo. A principais bibliotecas serão *"TensorFlow"* para criação das estruturas da "RedeNeural"e *"Pandas"* para criação e manipulação dos *"DataFremes"*. Poderemos ver o código referente a essas importações na figura 29.

Logo apos as importações das Bibliotecas será feito também a declaração de algumas constantes, como: caminho principal dos dados, tamanho das imagens, o "Batch Size"e o número de "Epocas"padrão no treinamento. Usamos por padrão "25 épocas"e formos aumentando conforme o necessário. Podemos ver esse trecho de código na figura 30.

Figura 29 – Importações Biblioteca (Google Colab)

```
[1] import re
    from tensorflow.keras.layers import *
    from tensorflow.keras.optimizers import *
    import seaborn as sns
    from tensorflow.keras.models import *
    from tensorflow.keras.preprocessing.image import *
    from tensorflow.keras.callbacks import *
    from tensorflow.keras import layers, models, metrics
    from tensorflow.keras.applications.efficientnet import *
    from sklearn.metrics import confusion_matrix, classification_report, accuracy_score
    from sklearn.model_selection import train_test_split
    from tensorflow.keras.applications.inception_v3 import InceptionV3
    from tensorflow.keras.applications.vgg19 import VGG19
    from tensorflow.keras.applications.resnet50 import ResNet50
    import matplotlib.pyplot as plt
    from sklearn.model_selection import train_test_split, StratifiedKFold
    import cv2
    from glob import glob
    from random import shuffle
    import tensorflow as tf
    import numpy as np
    import pandas as pd
    from pathlib import Path
    import os.path
```

Fonte: Autor

Figura 30 – Declaração das constantes (Google Colab)

```
[37]
    AUTOTUNE = tf.data.experimental.AUTOTUNE
    GCS_PATH = r"/content/drive/MyDrive/TCC"
    BATCH_SIZE = 16
    IMAGE_SIZE = [224, 224]
    EPOCHS = 25
```

Fonte: Autor

Apos importarmos as bibliotecas necessárias, e declararmos as constantes, será iniciado a fase onde será preparado os dados para o treinamento, validação e teste. Como mostrado no capítulo 4 na figura 15 será criado um "Dataframe"com as principais informações do nosso *"DataSet"*, sendo elas *"FikePath","Type","class","Magnification"* e *"Dataset"* como descritos na lista 4. Serão utilizados para criação do dataframe as seguintes funções apresentadas nas figuras 14, 42, 43,44 e 45.

Com o "dataframe"conseguiremos segmentar os dados de acordo com sua *"Magnification"*. Como já explicado começarei com o modelo utilizando apenas *"Magnification"= 400x* e depois será feita a separação entre *"Dataset de treino"* e *"Dataset de teste"*. Poderemos observar o código referente a essas segmentações na figura 31.

Figura 31 – Segmentação Dataframe 400x (Google Colab)

```
[ ] df_treino_400x = df_geral_400x[df_geral_400x['DataSet'] == 'Tr'][df_geral_400x['Magnification'] == '400x']
    df_teste_400x = df_geral_400x[df_geral_400x['DataSet'] == 'Te'][df_geral_400x['Magnification'] == '400x']
    df_treino_400x = df_treino_400x.drop(columns=['Magnification','DataSet'])
    df_teste_400x = df_teste_400x.drop(columns=['Magnification','DataSet'])
```

Fonte: Autor

Foi realizado o mesmo procedimento para buscarmos as imagens com *"Magnification"= 200x* e também o separaremos pelo tipo de dataset, sendo eles: *"Dataset de treino"* e *"Dataset de teste"*.Poderemos observar o código referente a essas segmentações na figura 32.

Para realizarmos o treinamento, precisaremos definir a taxa de aprendizagem, que no nosso caso será 10e-4 e a separação das imagens em *batchs*, que nesse caso terão o tamanho de 16.

Todos os nossos modelos de Rede serão uma junção da *"parte de baixo"("floor")* das arquiteturas. Pegaremos todas as camadas convolucionais e também o resultado dos pesos adquiridos no treinamento com a base de dados "ImageNet". Integraremos com a *"parte de cima"("top")* criada para criar um pós-processamentos vindo das camadas convolucionais.

O método de optimização utilizado no projeto será o *"ADAM"*.Segundo Ribeiro e Junior (2020) *"ADAM"* é um método de otimização que calcula taxas de aprendizado

Figura 32 – Segmentação Dataframe 200x (Google Colab)

```
[17] df_geral =  montar_data_frame(r'/content/drive/MyDrive/TCC/200x')
     df_treino = df_geral[df_geral['DataSet'] == 'Tr'][df_geral['Magnification'] == '200x']
     df_teste = df_geral[df_geral['DataSet'] == 'Te'][df_geral['Magnification'] == '200x']
```

Fonte: Autor

adaptativas para cada parâmetro. *"ADAM"* mantêm uma média exponencial decrescente dos gradientes anteriores.

Como demonstrado no capítulo 3 na subseção 3.2 a imagem será redimensionada para "224 x 224 x 3". Sendo esse o padrão da camada de entrada(*input layer*) de todos os nossos modelos.

A sessão de treinamento é divida em 2 processos, sendo o primeiro a separação entre cânceres malignos e benignos a segunda a identificação de suas classes.

Na primeira sessão de treinamento como falado antes, prepararemos a nossa rede neural para identificar um câncer maligno. Seguiremos o seguinte passo-a-passo para esse treinamento:

1. Realização do treinamento para imagens com amplitude de "400x"por 25 épocas;
2. Realização de uma nova rodada de treinamento agora por 15 épocas caso a anterior não alcance uma taxa minima de 93% de acurácia;;
3. Uma última rodada de treinamento com apenas 10 épocas caso a anterior não alcance uma taxa minima de 93% de acurácia;
4. validação do treinamento com o *dataset* de teste.
5. Criação de um *dataset* contendo apenas canceres malignos.
6. Importação dos pesos adquiridos nos processos anteriores;
7. Realização uma rodada de treinamento para imagens com amplitude de "200x"por 10 épocas;
8. validação do treinamento com o *dataset* de teste.
9. Criação de um *dataset* contendo apenas canceres malignos.

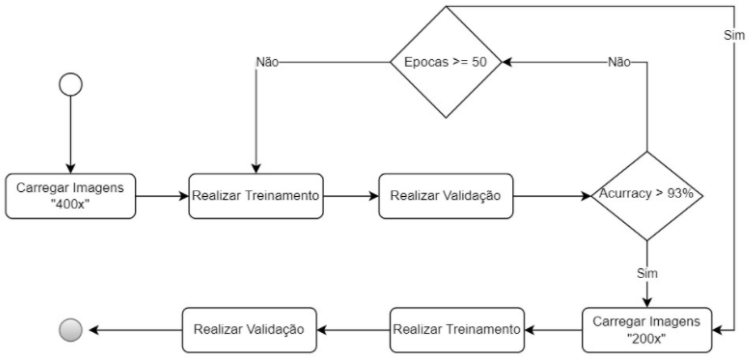
Apos o primeiro processo de treinamento, visto no conjunto de passos 5.3, será feita a segunda etapa onde classificaremos os canceres.Já tendo realizado a primeira etapa de treinamento e validação realizaremos as novas etapas de treino:

1. Realização do treinamento para imagens com amplitude de "400x"por 50 épocas;
2. Realização de uma nova rodada de treinamento agora por 25 épocas, aso a anterior não alcance uma taxa minima de 93% de acurácia;

3. Uma última rodada de treinamento com apenas 25 épocas caso a anterior não alcance uma taxa minima de 93% de acurácia;
4. validação do treinamento com o *dataset* de teste.
5. Importação dos pesos adquiridos nos processos anteriores;
6. Realização uma rodada de treinamento para imagens com amplitude de "200x"por 25 épocas;
7. validação do treinamento com o *dataset* de teste.
8. Criação de um *dataset* contendo os canceres e suas classes.

Podemos ver ilustrativamente, na figura 33, o processo descrito no passo a passo da criação do modelo de predição para categorias.

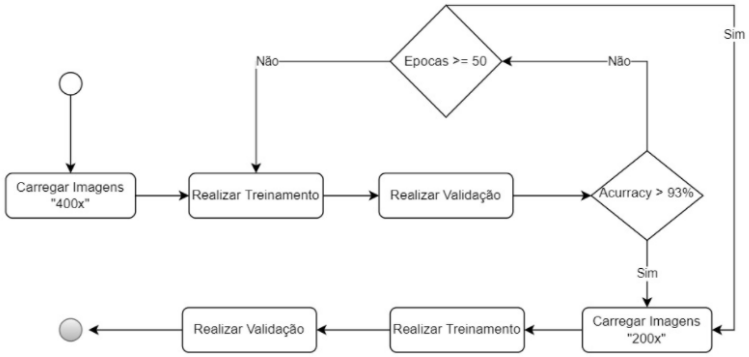
Figura 33 – Diagrama de fluxo criação do modelo categoria



Fonte: Autor

Podemos ver ilustrativamente, na figura 35, o processo descrito no passo a passo da criação do modelo de predição para classes.

Figura 34 – Diagrama de fluxo criação do modelo classe

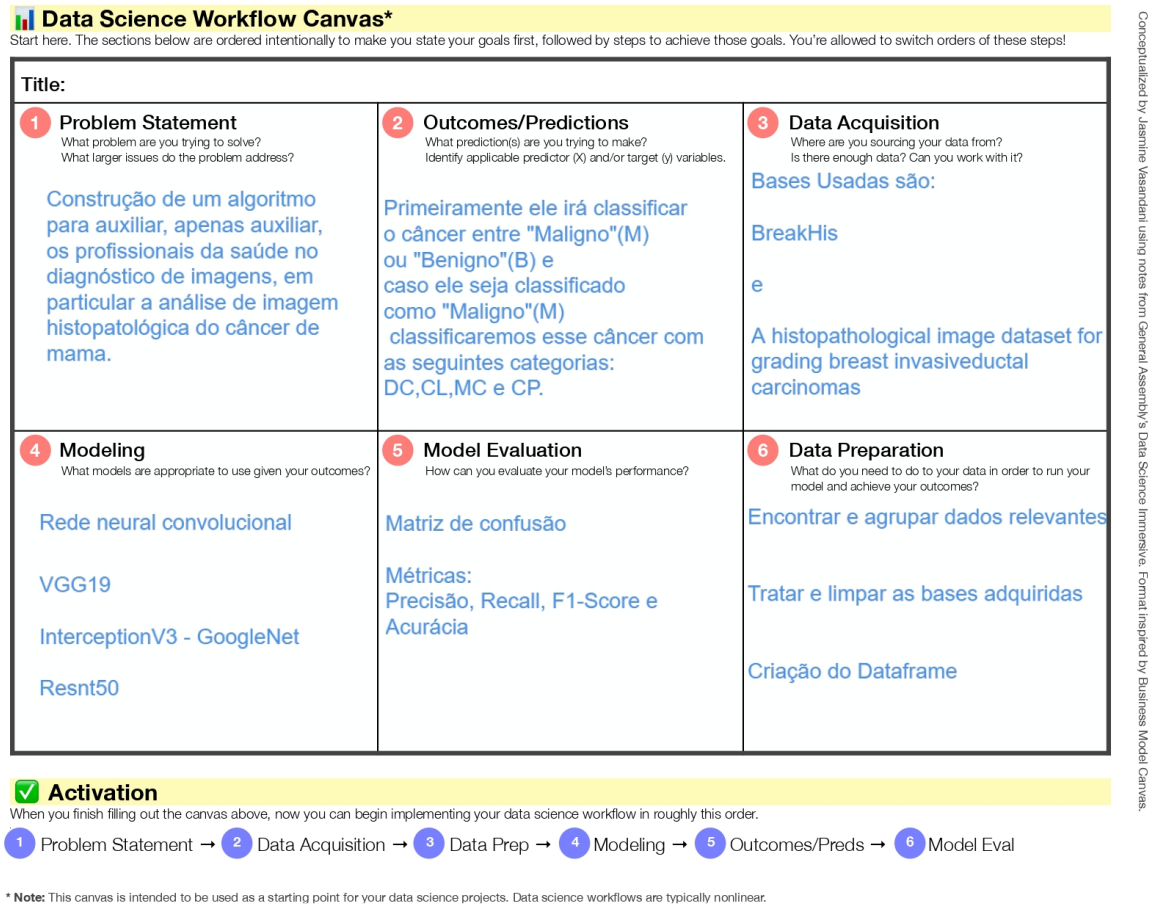


Fonte: Autor

6 APRESENTAÇÃO DOS RESULTADOS

Os resultados do projeto foram obtidos a partir do estabelecimento das metas mapeadas no modelo Canvas de Vasandani conforme apresentado na figura 35

Figura 35 – Modelo Canvas



📊 **Data Science Workflow Canvas***

Start here. The sections below are ordered intentionally to make you state your goals first, followed by steps to achieve those goals. You're allowed to switch orders of these steps!

Title:

1 Problem Statement What problem are you trying to solve? What larger issues do the problem address?	2 Outcomes/Predictions What prediction(s) are you trying to make? Identify applicable predictor (X) and/or target (y) variables.	3 Data Acquisition Where are you sourcing your data from? Is there enough data? Can you work with it?
Construção de um algoritmo para auxiliar, apenas auxiliar, os profissionais da saúde no diagnóstico de imagens, em particular a análise de imagem histopatológica do câncer de mama.	Primeiramente ele irá classificar o câncer entre "Maligno"(M) ou "Benigno"(B) e caso ele seja classificado como "Maligno"(M) classificaremos esse câncer com as seguintes categorias: DC,CL,MC e CP.	Bases Usadas são: BreakHis e A histopathological image dataset for grading breast invasiveductal carcinomas
4 Modeling What models are appropriate to use given your outcomes?	**5 Model Evaluation** How can you evaluate your model's performance?	**6 Data Preparation** What do you need to do to your data in order to run your model and achieve your outcomes?
Rede neural convolucional VGG19 InterceptionV3 - GoogleNet Resnt50	Matriz de confusão Métricas: Precisão, Recall, F1-Score e Acurácia	Encontrar e agrupar dados relevantes Tratar e limpar as bases adquiridas Criação do Dataframe

✅ **Activation**

When you finish filling out the canvas above, now you can begin implementing your data science workflow in roughly this order.

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 Modeling → 5 Outcomes/Preds → 6 Model Eval

* **Note:** This canvas is intended to be used as a starting point for your data science projects. Data science workflows are typically nonlinear.

Conceptualized by Jasmine Vasandani using notes from General Assembly's Data Science Immersive. Format inspired by Business Model Canvas.

Fonte: Autor

A seguir apresentaremos os resultados adquiridos pelos modelos das nossas *"CNN"*. Apresentaremos os resultados em uma "Matriz de confusão"e uma tabela mostrando sua eficiência.

Antes de seguirmos com a apresentação dos resultados, precisarmos realizar as importações das bibliotecas necessárias, como visto na figura 29

Os resultados serão apresentados seguindo a seguinte ordem:

1. Resultados para "Tipo de Câncer"com o modelo VGG19
2. Resultados para "Classe de Câncer"com o modelo VGG19
3. Resultados para "Tipo de Câncer"com o modelo GoogleNet
4. Resultados para "Classe de Câncer"com o modelo GoogleNet

5. Resultados para "Tipo de Câncer"com o modelo ResNet

6. Resultados para "Classe de Câncer"com o modelo ResNet

Primeiro passo para apresentação do resultado será a criação da rede e o carregamento dos pesos adquiridos nos treinamentos. Na imagem a seguir, mostraremos um exemplo desse código.

Figura 36 – Criação e carregamento de pesos (Google Colab)

```
[14] VGG19 = build_model("VGG19")
     VGG19.load_weights('/content/drive/MyDrive/TCC/VGG19_MB_TESTE.h5')
```

Fonte: Autor

Realizaremos os testes, para as amplitudes 400x, 200x e para o *dataset* completo.

Figura 37 – Código para teste do dataset (Google Colab)

```
[16] pred = VGG19.predict(test_images_VGG19_MB)
     pred=np.argmax(pred,axis=1)
```

Fonte: Autor

Apos testados, criaremos o *dataset* com o resultado.

Figura 38 – Criação e carregamento de pesos (Google Colab)

```
[22] pred_df_400=df_teste_400.copy()
     labels={}
     for l,v in test_images_VGG19_MB.class_indices.items():
         labels.update({v:l})
     pred_df_400['pred']=pred
     pred_df_400['pred']=pred_df_400['pred'].apply(lambda x: labels[x])
```

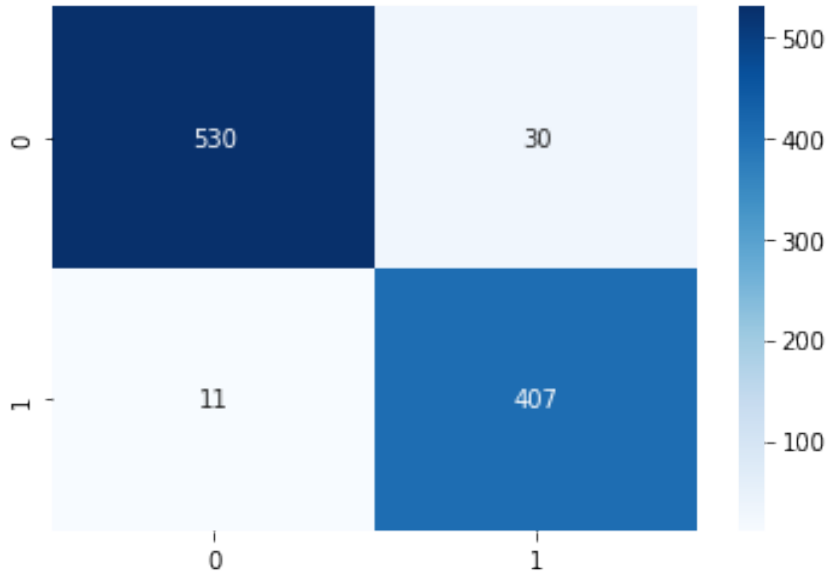
Fonte: Autor

Os exemplos presentes nas figuras 36,37 e 38 são referentes ao modelo *"VGG19"*, porem o padrão é replicado para os modelos *"GoogleNet"* e *"ResNet"*.

6.1 Modelo Preditivo - VGG19

A matriz de confusão abaixo (gráfico 12) mostra o resultado das predições efetuadas nos testes, na amostra onde temos apenas imagens com 400x.

Gráfico 12: Matriz de Confusão Categoria de Cancer 400X



Fonte: Autor

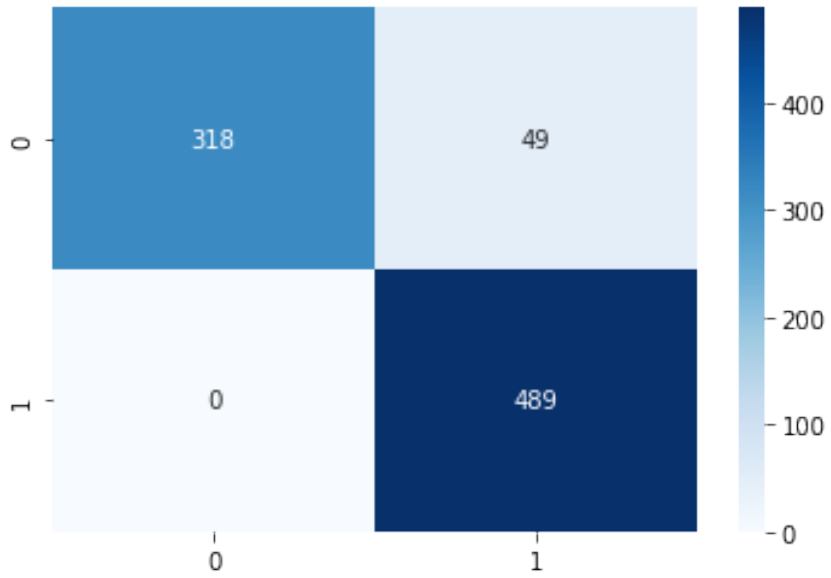
Tabela 4 – Tabela de métricas do gráfico 12

Tipo de Câncer	Precisão	Recall	F1-Score	Support
Benigno	98%	95%	96%	560
Maligno	93%	97%	95%	418

Fonte: Autor

A matriz de confusão abaixo (gráfico 13) mostra o resultado das predições efetuadas nos testes, na amostra onde temos apenas imagens com 200x.

Gráfico 13: Matriz de Confusão Categoria de Cancer 200X



Fonte: Autor

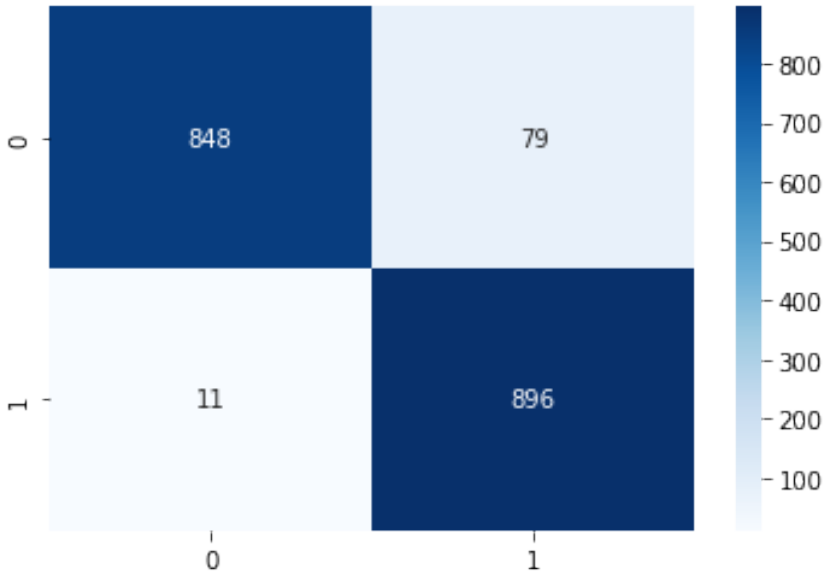
Tabela 5 – Tabela de métricas do gráfico 13

Tipo de Câncer	Precisão	Recall	F1-Score	Support
Benigno	100%	87%	93%	367
Maligno	91%	100%	95%	489

Fonte: Autor

A matriz de confusão abaixo (gráfico 14) mostra o resultado das predições efetuadas nos testes com todas as imagens.

Gráfico 14: Matriz de Confusão Categoria de Cancer



Fonte: Autor

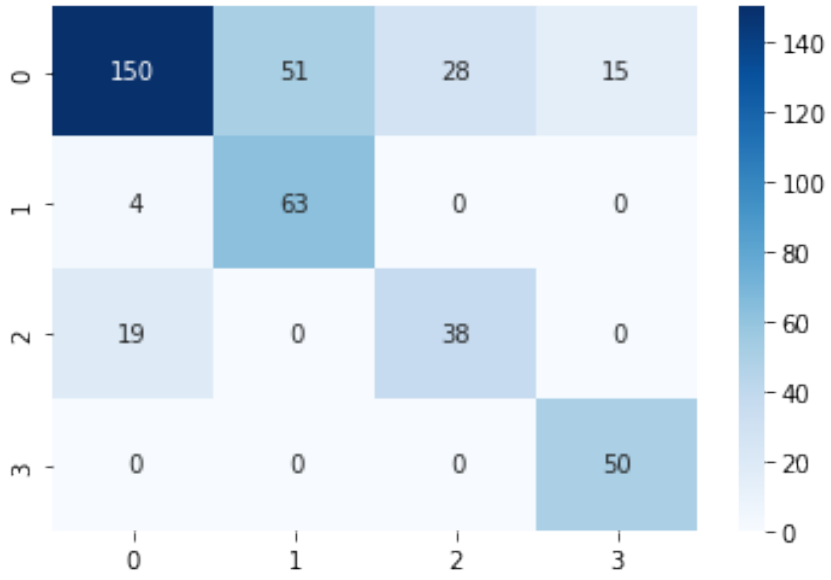
Tabela 6 – Tabela de métricas do gráfico 14

Tipo de Câncer	Precisão	Recall	F1-Score	Support
Benigno	99%	91%	95%	1301
Maligno	92%	99%	95%	1406

Fonte: Autor

A matriz de confusão abaixo (gráfico 15) mostra o resultado das predições efetuadas nos testes, para classes de câncer na amostra onde temos apenas imagens com 400x.

Gráfico 15: Matriz de Confusão Classe de Cancer 400X



Fonte: Autor

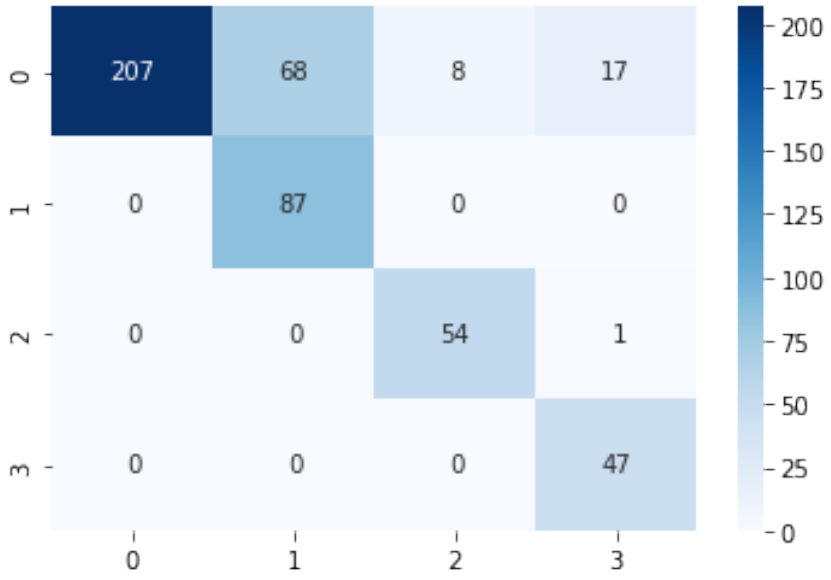
Tabela 7 – Tabela de métricas do gráfico 15

Classe do Câncer	Precisão	Recall	F1-Score	Support
DC	87%	61%	72%	244
LC	55%	94%	70%	67
MC	58%	67%	62%	57
PC	77%	100%	87%	50

Fonte: Autor

A matriz de confusão abaixo (gráfico 16) mostra o resultado das predições efetuadas nos testes, para classes de câncer na amostra onde temos apenas imagens com 200x.

Gráfico 16: Matriz de Confusão Classe de Cancer 200X



Fonte: Autor

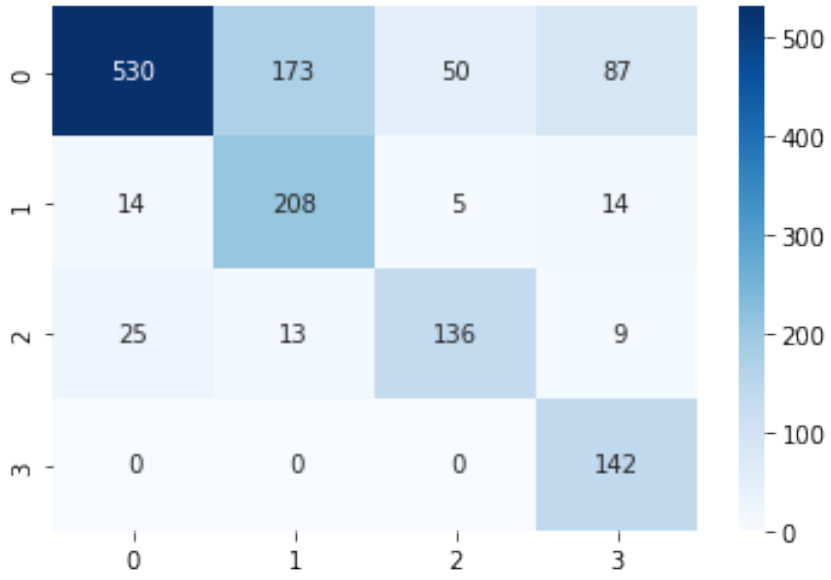
Tabela 8 – Tabela de métricas do gráfico 16

Classe do Câncer	Precisão	Recall	F1-Score	Support
DC	100%	69%	82%	300
LC	56%	100%	72%	87
MC	87%	98%	92%	55
PC	72%	100%	84%	47

Fonte: Autor

A matriz de confusão abaixo (gráfico 17) mostra o resultado das predições efetuadas nos testes com todas as imagens.

Gráfico 17: Matriz de Confusão Classe de Cancer



Fonte: Autor

Tabela 9 – Tabela de métricas do gráfico 17

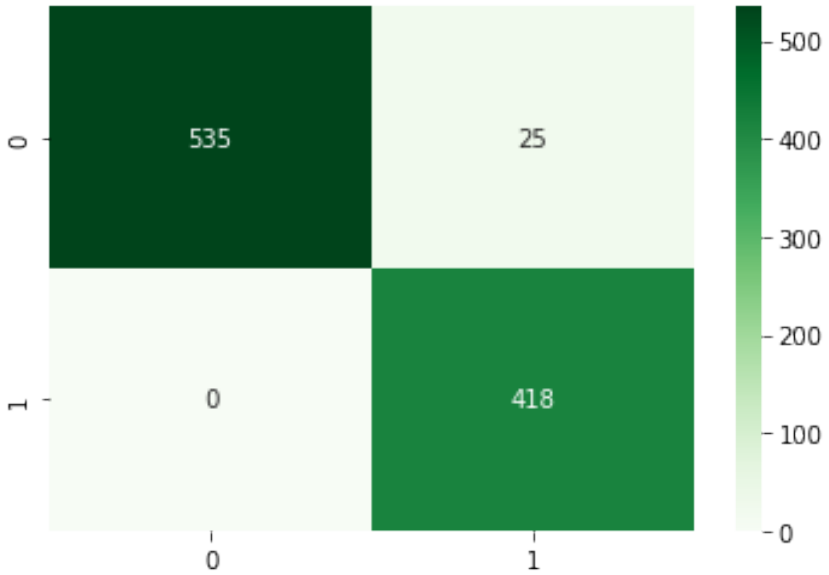
Classe do Câncer	Precisão	Recall	F1-Score	Support
DC	93%	63%	75%	840
LC	53%	86%	66%	241
MC	71%	74%	73%	183
PC	56%	100%	72%	142

Fonte: Autor

6.2 Modelo Preditivo - GoogleNet

A matriz de confusão abaixo (gráfico 18) mostra o resultado das predições efetuadas nos testes, na amostra onde temos apenas imagens com 400x.

Gráfico 18: Matriz de Confusão Categoria de Cancer 400X



Fonte: Autor

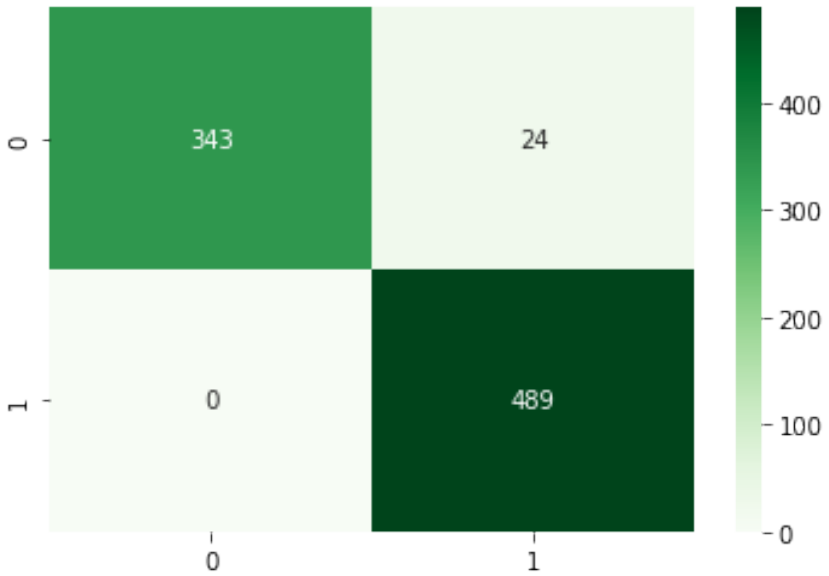
Tabela 10 – Tabela de métricas do gráfico 15

Tipo de Câncer	Precisão	Recall	F1-Score	Support
Benigno	100%	96%	98%	560
Maligno	94%	100%	97%	418

Fonte: Autor

A matriz de confusão abaixo (gráfico 19) mostra o resultado das predições efetuadas nos testes, na amostra onde temos apenas imagens com 200x.

Gráfico 19: Matriz de Confusão Categoria de Cancer 200X



Fonte: Autor

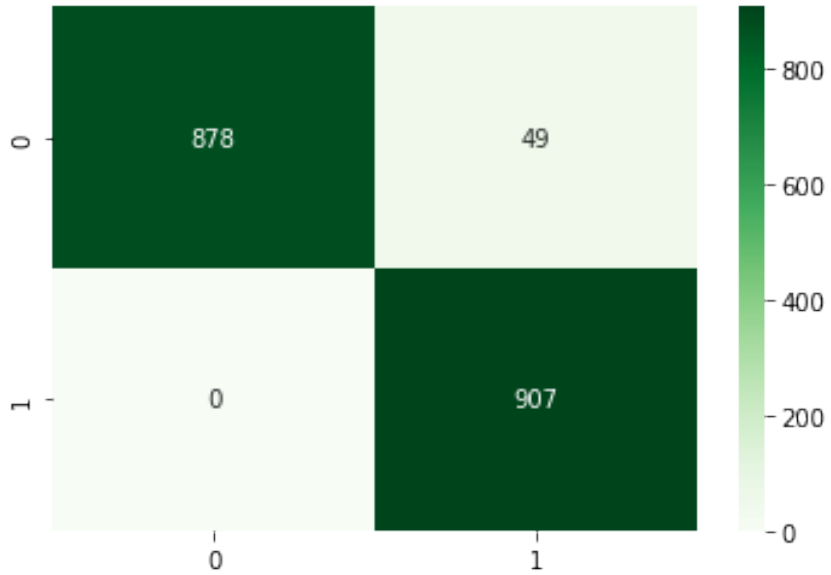
Tabela 11 – Tabela de métricas do gráfico 19

Tipo de Câncer	Precisão	Recall	F1-Score	Support
Benigno	100%	93%	97%	367
Maligno	95%	100%	98%	489

Fonte: Autor

A matriz de confusão abaixo (gráfico 20) mostra o resultado das predições efetuadas nos testes com todas as imagens.

Gráfico 20: Matriz de Confusão Tipo de Cancer



Fonte: Autor

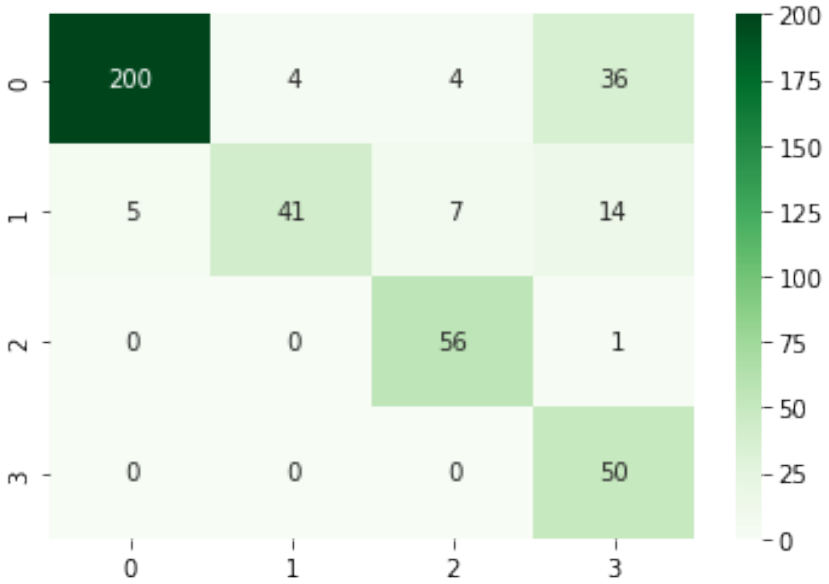
Tabela 12 – Tabela de métricas do gráfico 20

Tipo de Câncer	Precisão	Recall	F1-Score	Support
Benigno	100%	95%	97%	927
Maligno	96%	100%	98%	907

Fonte: Autor

A matriz de confusão abaixo (gráfico 21) mostra o resultado das predições efetuadas nos testes, para classes de câncer na amostra onde temos apenas imagens com 400x.

Gráfico 21: Matriz de Confusão Classe de Cancer 400X



Fonte: Autor

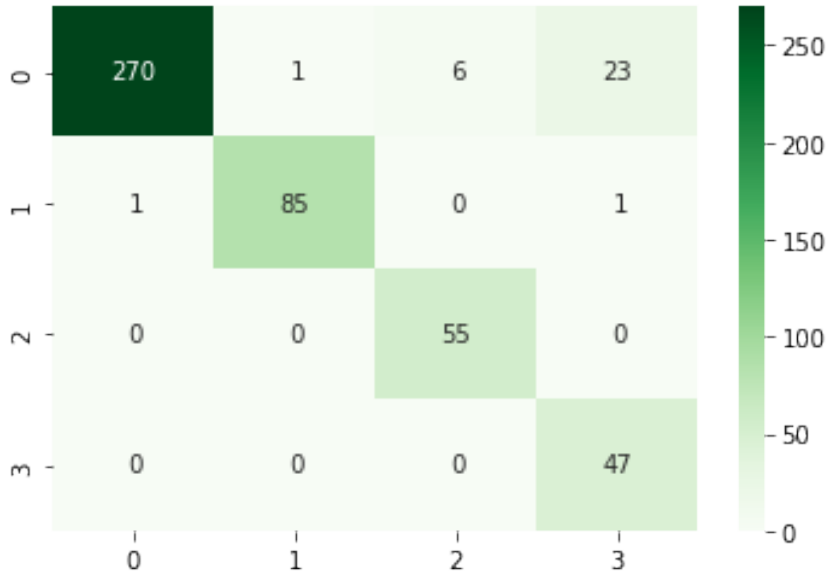
Tabela 13 – Tabela de métricas do gráfico 21

Classe do Câncer	Precisão	Recall	F1-Score	Support
DC	98%	82%	89%	244
LC	91%	61%	73%	67
MC	84%	98%	90%	57
PC	50%	100%	66%	50

Fonte: Autor

A matriz de confusão abaixo (gráfico 22) mostra o resultado das predições efetuadas nos testes, para classes de câncer na amostra onde temos apenas imagens com 200x.

Gráfico 22: Matriz de Confusão Classe de Cancer 200X



Fonte: Autor

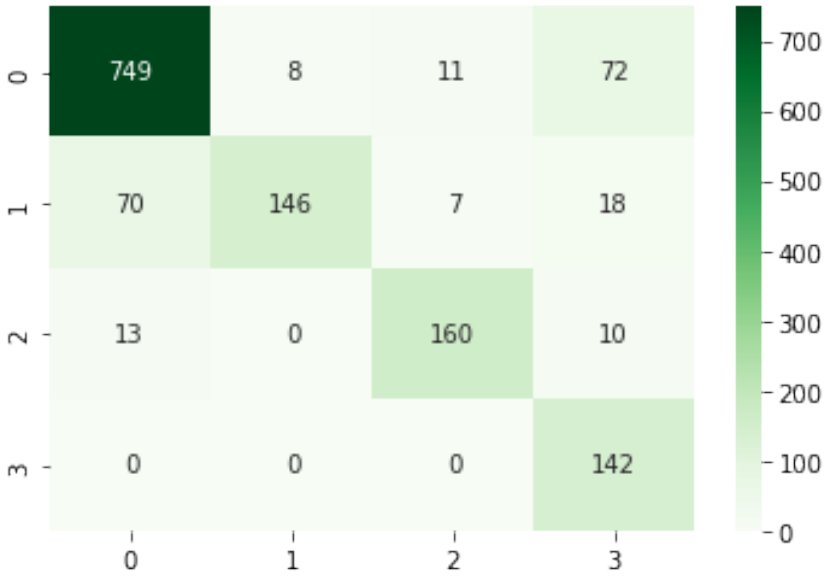
Tabela 14 – Tabela de métricas do gráfico 22

Classe do Câncer	Precisão	Recall	F1-Score	Support
DC	100%	90%	95%	300
LC	99%	98%	98%	87
MC	90%	100%	95%	55
PC	66%	100%	80%	47

Fonte: Autor

A matriz de confusão abaixo (gráfico 23) mostra o resultado das predições efetuadas nos testes com todas as imagens.

Gráfico 23: Matriz de Confusão Classe de Cancer



Fonte: Autor

Tabela 15 – Tabela de métricas do gráfico 23

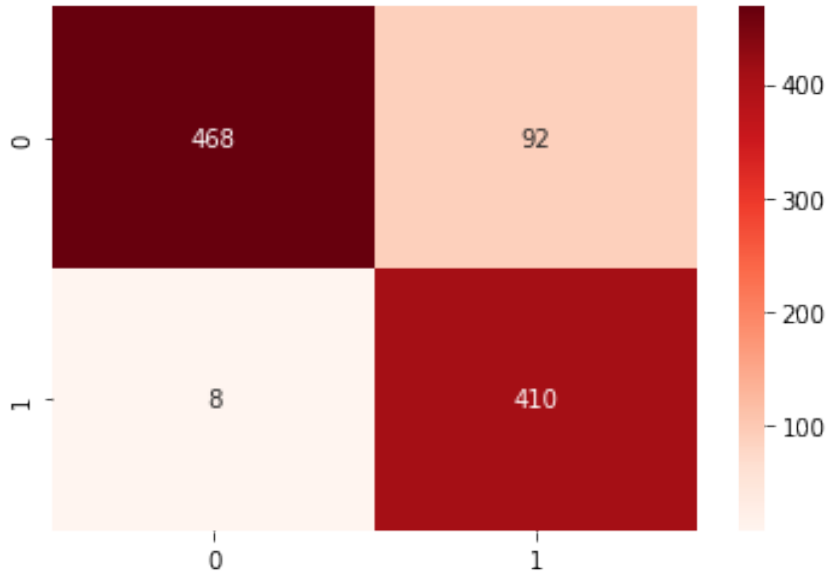
Classe do Câncer	Precisão	Recall	f1-score	Support
DC	90%	89%	90%	840
LC	95%	61%	74%	241
MC	90%	87%	89%	183
PC	59%	100%	74%	142

Fonte: Autor

6.3 Modelo Preditivo - ResNet50

A matriz de confusão abaixo (gráfico 24) mostra o resultado das predições efetuadas nos testes, na amostra onde temos apenas imagens com 400x.

Gráfico 24: Matriz de Confusão Categoria de Cancer 400X



Fonte: Autor

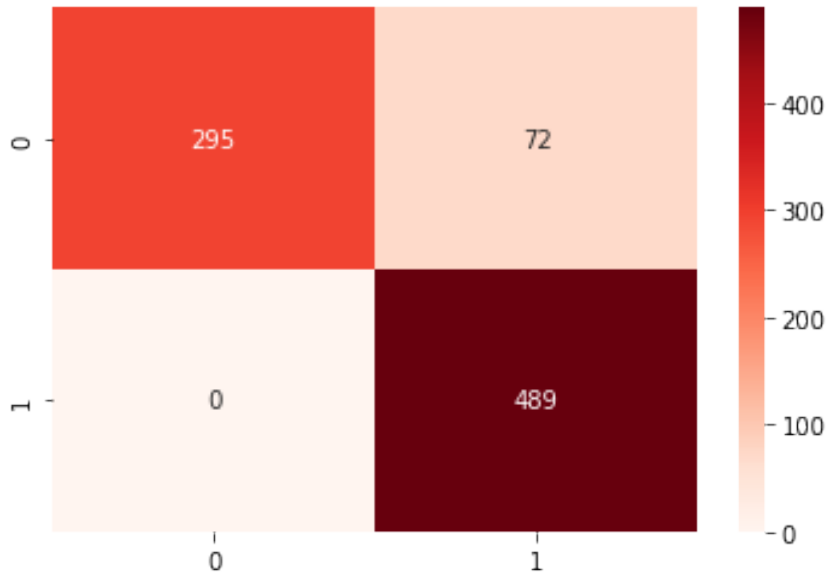
Tabela 16 – Tabela de métricas do gráfico 24

Tipo de Câncer	Precisão	Recall	F1-Score	Support
Benigno	98%	84%	90%	560
Maligno	82%	98%	89%	418

Fonte: Autor

A matriz de confusão abaixo (gráfico 25) mostra o resultado das predições efetuadas nos testes, na amostra onde temos apenas imagens com 200x.

Gráfico 25: Matriz de Confusão Categoria de Cancer 200X



Fonte: Autor

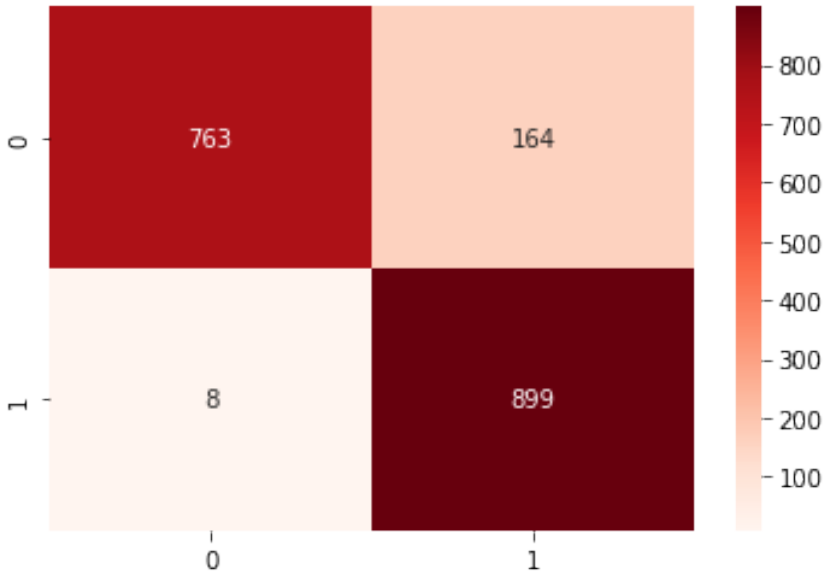
Tabela 17 – Tabela de métricas do gráfico 25

Tipo de Câncer	Precisão	Recall	F1-Score	Support
Benigno	100%	80%	89%	367
Maligno	87%	100%	93%	489

Fonte: Autor

A matriz de confusão abaixo (gráfico 26) mostra o resultado das predições efetuadas nos testes com todas as imagens.

Gráfico 26: Matriz de Confusão Categoria de Cancer



Fonte: Autor

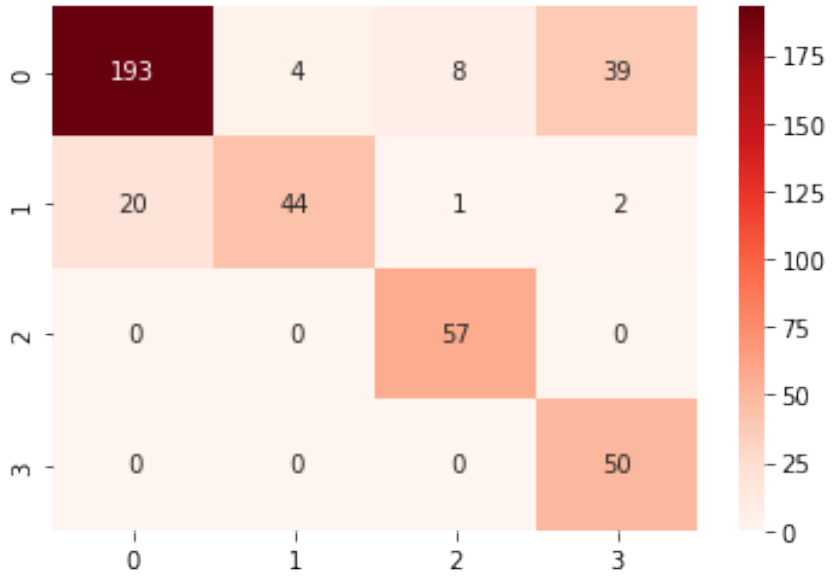
Tabela 18 – Tabela de métricas do gráfico 26

Tipo de Câncer	Precisão	Recall	F1-Score	Support
Benigno	99%	82%	90%	927
Maligno	85%	99%	91%	907

Fonte: Autor

A matriz de confusão abaixo (gráfico 27) mostra o resultado das predições efetuadas nos testes, para classes de câncer na amostra onde temos apenas imagens com 400x.

Gráfico 27: Matriz de Confusão Classe de Cancer 400X



Fonte: Autor

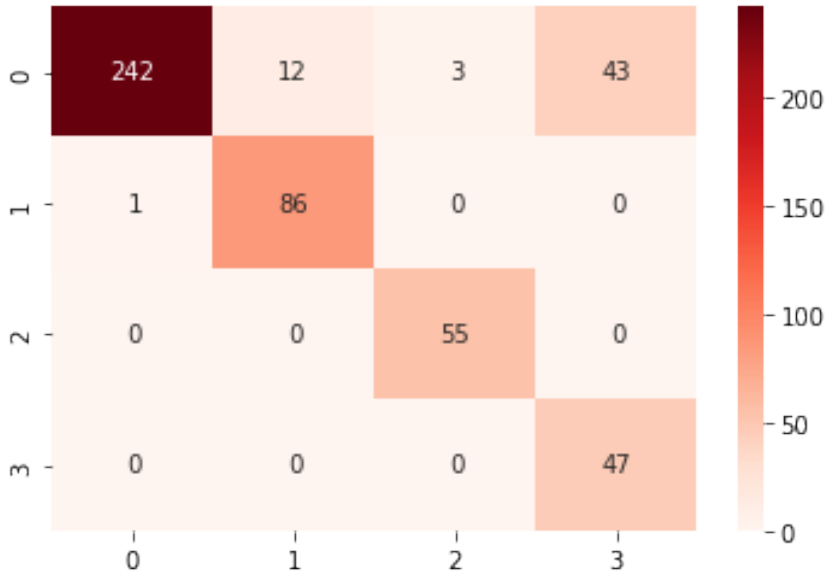
Tabela 19 – Tabela de métricas do gráfico 27

Classe do Câncer	Precisão	Recall	F1-Score	Support
DC	91%	79%	84%	224
LC	92%	66%	77%	67
MC	86%	100%	93%	57
PC	55%	100%	71%	50

Fonte: Autor

A matriz de confusão abaixo (gráfico 28) mostra o resultado das predições efetuadas nos testes, para classes de câncer na amostra onde temos apenas imagens com 200x.

Gráfico 28: Matriz de Confusão Classe de Cancer 200X



Fonte: Autor

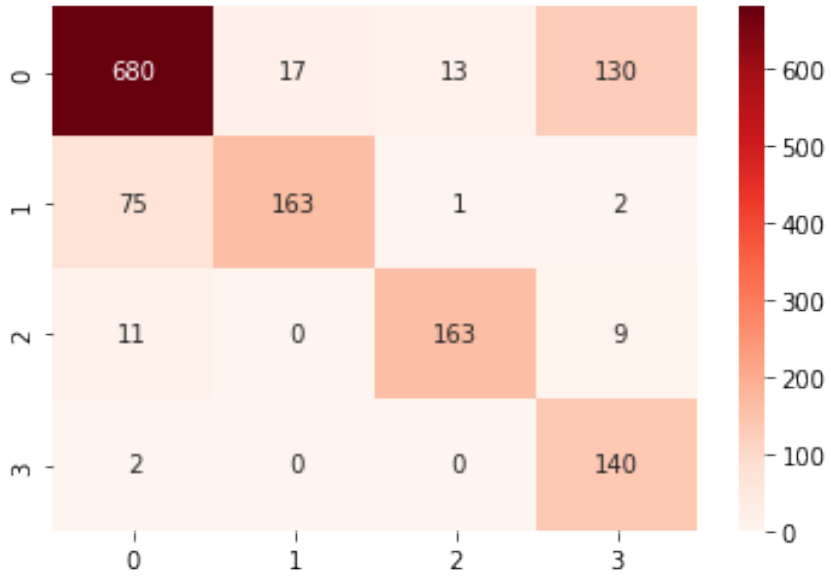
Tabela 20 – Tabela de métricas do gráfico 28

Classe do Câncer	Precisão	Recall	F1-Score	Support
DC	100%	81%	89%	300
LC	88%	99%	93%	87
MC	95%	100%	97%	55
PC	52%	100%	69%	47

Fonte: Autor

A matriz de confusão abaixo (gráfico 29) mostra o resultado das predições efetuadas nos testes com todas as imagens.

Gráfico 29: Matriz de Confusão Classe de Cancer



Fonte: Autor

Tabela 21 – Tabela de métricas do gráfico 29

Classe do Câncer	Precisão	Recall	F1-Score	Support
DC	89%	81%	85%	840
LC	91%	68%	77%	241
MC	92%	89%	91%	183
PC	50%	99%	66%	142

Fonte: Autor

6.4 Resultado Geral

Nessa seção apresentaremos os resultados médios de cada previsão apresentadas nas seções anteriores. Serão presentados em duas tabelas. A primeira Representado a classificação por categoria de câncer a segunda referente a classificação por classes de câncer.

Os valores apresentados nas tabelas são médias ponderadas (ou *"weighted avg"*) das métricas obtidas em cada modelo ao prever sua saída (ou *"target"*). As métricas apresentadas serão: "Precisão", "Recall", "F1-Score"e "Acurracy".

Na tabela abaixo(tabela 22) é apresentado os resultados referente a predição da categoria do câncer.Nesta tabela, teremos uma visão de cada modelo prevendo as seguintes saídas: "M"e "B".*

* "Maligno (M)", "Benigno(B)".

Tabela 22 – Tabela de métricas dos modelos para Categorias de câncer 29

Modelo de CNN	Precisão	Recall	F1-Score	Acurácia
VGG19	95%	95%	95%	95%
GoogleNet	97%	97%	97%	97%
ResNet50	92%	91%	91%	91%

Fonte: Autor

Na tabela abaixo(tabela 23) é apresentado os resultados referente a predição da classe do câncer. Nesta tabela, teremos uma visão de cada modelo prevendo as seguintes saídas: "DC","LC","MC"e "PC"*

Tabela 23 – Tabela de métricas dos modelos para classe do câncer 29

Modelo de CNN	Precisão	Recall	F1-Score	Acurácia
VGG19	68%	81%	71%	72%
GoogleNet	92%	89%	89%	89%
ResNet50	89%	85%	86%	85%

Fonte: Autor

6.5 Combinação de Modelos

Após as execuções dos modelos, e tendo visto separadamente cada avaliação percebemos que os melhores modelos até então são, o GoogleNet(InterceptionV3) para categoria e GoogleNet(InterceptionV3) para classe.

Agora realizaremos a combinação dos modelos, para definirmos a conjunção final do Projeto ou confirmamos o que já tínhamos encontrado. Em um primeiro momento descartaremos os Modelos "ResNet Categoria"e "VGG19 Classes", por não atingirem resultados mínimos esperados.

Primeira conjunto de combinações a serem realizadas, serão as combinações entre "VGG19 Categoria"com o modelo "GoogleNet Classe". Essa combinação será chamada: "Combinação 1". A segunda combinação será entre "VGG19 Categoria"e o modelo "ResNet Classe". Essa combinação receberá o nome de "Combinação 2".

O segundo conjunto de combinações, terá como base o modelo "GoogleNet Categoria". A "Combinação 3"será a junção dos modelos "GoogleNet Categoria"com o modelo "ResNet Classe"e por último a "Combinação 4"composta por "GoogleNet Categoria"e "GoogleNet Classe".

Podemos observar na tabela 24 as métricas obtidas de "cada combinação".

A parti dos resultados apresentados no decorrer do capítulo, é possível perceber que a melhor combinação de modelos é a "Combinação 4", como já tínhamos analisado anteriormente. Mesmo com o desbalanceamento de classes, foi encontrado Resultados suficientemente bons para aplicarmos em nosso modelo final.

Conseguirmos concluir o modelo com uma métrica média de F1-Score em 90%,

* "carcinoma (DC)", "carcinoma lobular(CL)", "Carcinomamucinoso (MC)"e "Carcinoma papilífero (CP)"

Tabela 24 – Tabela de métricas das combinações

Modelo de CNN	Precisão	Recall	F1-Score	Acurácia
Combinação 1	88%	90%	88%	90%
Combinação 2	86%	88%	87%	92%
Combinação 3	89%	90%	89%	90%
Combinação 4	91%	92%	90%	92%

Fonte: Autor

métrica obtida através de uma média simples entre o modelo de categoria e de classe, um resultado bom considerando nossa base de dados e modelos, e uma acurácia média de 92%, considerando também uma média simples entre o modelo de categoria e de classe.

Conforme os dados anteriores, conseguimos encontrar uma combinação, considerada até onde podemos ver, "ótima"para nosso modelo. Uma combinação onde poderemos encontrar uma taxa de sucesso de 90%.

Com o modelo descrito nesse trabalho, poderemos ajudar a diminuir a taxa de falsos negativos, o que trazem o risco de um diagnóstico tardio da doença o que aumenta o risco de óbito, devido ao câncer, sendo esse o quarto em questão de mortalidade.

7 TRABALHOS FUTUROS

8 LINKS

A seguir, estão os links do vídeo de apresentação e do repositório contendo os dados utilizados no projeto.

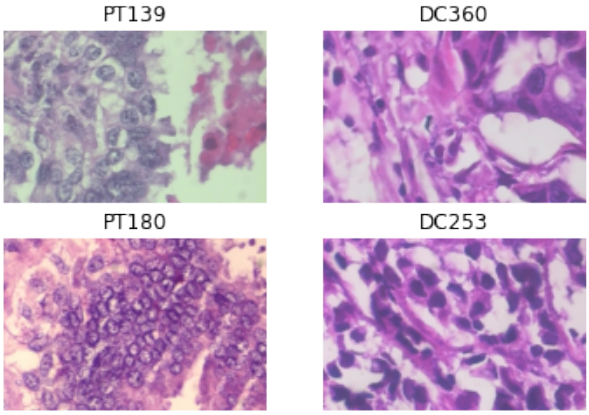
Link para base de dados(imagens de teste): <https://drive.google.com/drive/folders/1VQpWhlzK__2EUfd4CMeUWQ8-wAHaWGeT?usp=sharing>

Link para repositório: <https://github.com/RodrigoOBC/TCC-CienciaDados>

Apresentação:

9 FIGURAS / GRAFICOS

Figura 39 – Amostra com novos nomes, ampliação 400x



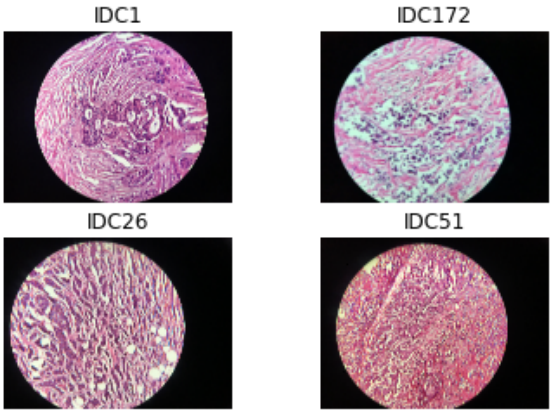
Fonte: Autor

Figura 40 – Amostra com nomes de imagens (Jupyter notebook)

```
01_BC_G1_9057_10x_1.JPG
02_BC_G1_9103_10x_1.JPG
03_BC_G1_10959_10x_1.JPG
04_BC_G1_10783_10x_1.JPG
05_BC_G1_9610_10x_1.JPG
06_BC_G1_7967_10x_1.JPG
01_BC_G1_9057_20x_1.JPG
01_BC_G1_9057_40x_1.JPG
01_BC_G1_9057_4x_1.jpg
```

Fonte: Autor

Figura 41 – Amostra da base dados Breast invasive ductal carcinomas (Jupyter notebook)



Fonte: Autor

Figura 42 – Código para pegar nome da figura (Google Colab)

```
def montar_nome(path):
    nome_file = []
    file_path = []
    x = 0
    for p, _, files in os.walk(os.path.abspath(path)):
        for file_name in files:
            if 'ResNet'  in file_name or "VGG19" in file_name or "Interception" in file_n
              continue
            else:
              nome_file.append(file_name)
              file_path.append(p+'//'+file_name)
    return nome_file, file_path
```

Fonte: Autor

Figura 43 – Código para pegar grau de ampliação (Jupyter notebook)

```
def pegar_ampliacao(file_path):
    ampliacao = []
    for path in file_path:
        if '40x' in path:
            ampliacao.append('40x')
        elif '100x' in path:
            ampliacao.append('100x')
        elif '200x' in path:
            ampliacao.append('200x')
        elif '400x' in path:
            ampliacao.append('400x')
    return ampliacao
```

Fonte: Autor

Figura 44 – Código para selecionar a classe (Google Colab)

```
def selecionar_classe(nomes):
    classe = []
    for nome in nomes:
        nome = ''.join([i for i in nome if not i.isdigit()])
        nome = nome.split('.')[0]
        if "TA" == nome:
            classe.append('TA')
        elif "DCI" == nome:
            classe.append('DCI')
        elif "F" == nome:
            classe.append('F')
        elif "PT" == nome:
            classe.append('PT')
        elif "DC" == nome:
            classe.append('DC')
        elif "LC" == nome:
            classe.append('LC')
        elif "MC" == nome:
            classe.append('MC')
        elif "PC" == nome:
            classe.append('PC')
        elif 'A' == nome:
            classe.append('A')
    return classe
```

Fonte: Autor

Figura 45 – Código para pegar os tipos (Jupyter notebook)

```
def pegar_tipos(file_path):
    tipo = []
    for path in file_path:
        path = path.split('/')[7]
        if 'B' in path:
            tipo.append('B')
        if 'M' in path:
            tipo.append('M')
    return tipo
```

Fonte: Autor

REFERÊNCIAS

AMARANTE, S. Câncer de mama: a importância do diagnóstico precoce. In: IFF/FIOCRUZ. Rio de Janeiro, 2021. Disponível em: <http://www.iff.fiocruz.br/index.php/8-noticias/274-cancerdemama>. Acesso em: 12 set. 2021. 14

ANôNIMO. Como funcionam os neurônios. p. 1, 2014. Disponível em: <https://metodosupera.com.br/como-funcionam-os-neuronios/>. Publicado em: 17/09/2014. 32

ANôNIMO. "células nervosas"em só biologia. p. 1, 2021. Consultado em 19/08/2021 às 19:30. Disponível na Internet em <https://www.sobiologia.com.br/conteudos/FisiologiaAnimal/nervoso2.php>. 32, 33

BOLHASANI, H. et al. “a histopathological image dataset for grading breast invasive ductal carcinomas”. v. 19, n. 100341, p. 1, 2020. 14, 16, 17

DATA SCIENCE ACADEMY. *Deep Learning Book*. 1. ed. [S.l.]: Data Science Academy, 2021. Disponível em: <https://www.deeplearningbook.com.br/>.Acesso em: 19/08/2021 às 16:10. 33, 34, 35, 36, 37, 39

DIAGNóSTICO: o que interfere no falso positivo ou negativo. 2021. Disponível em: <https://www.oncologicadobrasil.com.br/blog/diagnostico-falso-positivo-negativo/>. acessado em: 11 de set.de 2021. 14

GOODFELLOW YOSHUA BENGIO, A. C. I. Deep learning. In: . [S.l.]: MIT., 2016. 42

HAYKIN, S. *Redes neurais: princípios e prática.* 1. ed. [S.l.]: Bookman Editora, 2007. 35, 38, 39

HEBB, D. *The organization of behavior: A neuropsychological approach.* 1. ed. [S.l.]: John Wiley and Sons, Inc., 1950. 34

INCA. *Atlas on-line de mortalidade.* 1. ed. [S.l.]: INSTITUTO NACIONAL DE CÂNCER JOSÉ ALENCAR GOMES DA SILVA, 2019. Disponível em: <https://www.inca.gov.br/aplicativos/atlas-de-mortalidade-por-cancer/>.Acesso em: 19/08/2021 às 16:10. 11

MADHAVAN, S. Introduction to convolutional neural networks. p. 1, 2017. Disponível em: <https://developer.ibm.com/articles/introduction-to-convolutional-neural-networks/>. Publicado em: 13/07/2021. 40, 41, 42

MADHAVAN, S.; JONES, M. T. Deep learning architectures. p. 1, 2017. Disponível em: <https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/>. Publicado em: 08/09/2017 e atualizado em: 25/01/2021. 40

MCCULLOC, W.; PITTS, W. The bulletin of mathematical biophysics. v. 5, n. 4, p. 115–133, 1943. 33

MORSCH, J. Como é feita a mamografia digital: para que serve, preparo e vantagens. In: . Rio Grande do Sul: [s.n.], 2018. Disponível em: <https://telemedicinamorsch.com.br/blog/como-e-feita-a-mamografia-digital>. Acesso em: 12 set. 2021. 10

O que é o diagnóstico histopatológico e qual sua importância? Anticorpos Laboratório, 2019. Disponível em: <https://www.anticorpos.com.br/artigos/o-que-e-o-diagnostico-histopatologico-e-qual-sua-importancia/>. acessado em: 13 de dez.de 2021. 10

OLIVEIRA, A. VisÃo computacional aplicada À anÁlise de mamografias. v. 1, n. 1, p. 31 – 31, 2019. TCC (Bacharelado em Ciência da Computação) - Centro Universitário Carioca, Rio de Janeiro - Rio de Janeiro. 10

RIBEIRO, A.; JUNIOR, F. Um estudo comparativo entre cinco métodos de otimização aplicados em uma rnc voltada ao diagnóstico do glaucoma. *Revista de Sistemas e Computação*, CS, v. 10, n. 1, p. 122–130, 2020. Disponível em: <http://www.revistas.unifacs.br/index.php/rsc>. 44

RIZWAN, M. Convolutional neural networks – in a nut shell. In: ENGMRK. [S.l.], 2018. Disponível em: <https://engmrk.com/convolutional-neural-network-3/> Acesso em: 20/9/2021. 42

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, v. 6, n. 65, p. 386–408, 1958. 34

RUMELHART, D.; HINTON, G.; WILLIAMS, R. J. Learning internal representations by error propagation. In: . [S.l.: s.n.], 1986. 39

SAKURAI, R. Implementando a estrutura de uma rede neural convolucional utilizando o mapreduce do spark. In: . [S.l.: s.n.], 2017. Disponível em: <http://www.sakurai.dev.br/cnn-mapreduce/>. Publicado em: 20/12/2017. 40

SILVA, I. N. de Câncer José Alencar Gomes da. Detecção precoce. In: INCA. Rio de Janeiro, 2021. Disponível em: <https://www.inca.gov.br/controle-do-cancer-de-mama/acoes-de-controle/deteccao-precoce>. Acesso em: 12 set. 2021. 14

SPANHOL, F. A. et al. Breast cancer histopathological image classification using convolutional neural networks. p. 1, 2016. 14, 15, 16

SUNG, H. et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, ACS JOURNALS, v. 71, n. 3, p. 209–249, 2021. 10, 11

USHIZIMA. Redes neurais convolucionais com tensorflow: Teoria e prática. p. 382–406, 2017. 41