

UNIVERSIDADE PAULISTA
CIÊNCIA DA COMPUTAÇÃO

MELQUISEDEC FELIPE COUTINHO

**UTILIZANDO REDES NEURAIS CONVOLUCIONAIS PARA
RECONHECIMENTO DE CÂNCER**

BAURU
2020

MELQUISEDEC FELIPE COUTINHO

**UTILIZANDO REDES NEURAIS CONVOLUCIONAIS PARA
RECONHECIMENTO DE CÂNCER**

Trabalho de conclusão de curso apresentado
como requisito parcial de obtenção do título
de Bacharel em Ciência da Computação, do
Instituto de Ciência Exatas e Tecnologia, da
Universidade Paulista, campus de Bauru interior
do Estado de São Paulo.

Orientador: Prof. Me. Robson Fernandes da
Silva

Melquisedec Felipe Coutinho Utilizando Redes Neurais Convolucionais para
reconhecimento de Câncer/ Melquisedec Felipe Coutinho. – Bauru, 2020- 45
p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Me. Robson Fernandes da Silva

Trabalho de Conclusão de Curso – Universidade Paulista

Bacharelado em Ciência da Computação, 2020.

1. Ciência de Dados 2. Aprendizado de Máquina 3. Redes Neurais Convolucionais
4. Câncer de Mama 5. Saúde

Melquisedec Felipe Coutinho

Utilizando Redes Neurais Convolucionais para reconhecimento de Câncer

Trabalho de conclusão de curso apresentado
como requisito parcial de obtenção do título
de Bacharel em Ciência da Computação, do
Instituto de Ciência Exatas e Tecnologia, da
Universidade Paulista, campus de Bauru interior
do Estado de São Paulo.

Banca Examinadora

Prof. Me. Angela Teresa Rochetti
Universidade Paulista – UNIP

Prof. Me. Robson Fernandes da Silva
Orientador
Universidade Paulista - UNIP

Prof. Me. Victor de Assis Rodrigues
Universidade Paulista – UNIP

Bauru, _____ de _____ de _____.

Agradecimentos

Agradeço a minha família por sempre incentivar meus estudos, mesmo nos dias mais difíceis, agradeço ao meu orientador e amigo Robson Fernandes, por ter feito parte desse caminho desde a nossa contribuição na mesma empresa, agradeço minha namorada por ajudar na revisão do trabalho como todo e também por me apoiar sempre, agradeço aos amigos que fiz durante esses anos de estudos, aos professores que compartilharam seu conhecimento conosco e a cada um que de alguma forma fez e está fazendo parte dessa história.

"What I cannot create, I do not understand."

Richard Feynman

Resumo

Desde muitos anos o câncer é um grande causador de mortes no mundo, atingindo desde o país mais desenvolvido ao menos desenvolvido, visto que a doença nem sempre é diagnosticada precocemente. Um dos principais fatores que podem levar o paciente a óbito é o tempo, pois este, é crucial para o tratamento, uma vez que quanto antes se descobre o grau em que a doença se encontra, mais rápido pode-se iniciar o tratamento. Portanto, este trabalho busca amenizar o tempo de descoberta, com ênfase no câncer de mama, utilizando de aprendizado de máquina, possibilitando caracterizar facilmente o estágio que se encontra o câncer no paciente e, consequentemente, facilitar o tratamento, podendo acarretar na cura do mesmo. Serão realizados três processos para obter este resultado: um pré processamento das imagens do *dataset*; análise exploratória e criação de modelo para caracterização do estágio da doença, respectivamente.

Palavras-chave: Ciência de Dados, Aprendizado de Máquina, Redes Neurais Convolucionais, Câncer de Mama, Saúde.

Abstract

For many years cancer has been a major cause of death in the world, reaching from the most developed to the least developed country, since the disease is not always diagnosed early. One of the main factors that can lead the patient to death is time, as this is crucial for treatment, since the sooner the degree to which the disease is found, the faster treatment can be started. Therefore, this work seeks to shorten the time of discovery, with an emphasis on breast cancer, using machine learning, making it possible to easily characterize the stage that the cancer is found in the patient and, consequently, facilitate the treatment, which may result in the cure of the same . Three processes will be carried out to obtain this result: a pre-processing of the dataset images; exploratory analysis and creation of a model to characterize the disease stage, respectively.

Keywords: Data Science, Machine Learning, Convolutional Neural Networks, Breast Cancer, Health.

Listas de figuras

Figura 1 – Mortalidade Entre 2000 e 2009.	16
Figura 2 – Mortalidade Entre 2010 e 2018.	17
Figura 3 – La Notte.	18
Figura 4 – Mecanismo de Anticítera.	19
Figura 5 – Ada Lovelace.	20
Figura 6 – Tabulador e classificador de Herman Hollerith.	20
Figura 7 – "Mother of All Demos".	21
Figura 8 – Tecnologias Mais Populares.	22
Figura 9 – Neurônio.	24
Figura 10 – Perceptron.	25
Figura 11 – Perceptron Multicamadas.	26
Figura 12 – Arquitetura de uma CNN Tradicional.	27
Figura 13 – Cronograma.	32
Figura 14 – Estrutura do <i>dataset</i> .	33
Figura 15 – Função Pré Processamento	34
Figura 16 – Dataframe.	34
Figura 17 – Células <i>IDC</i> Negativo.	35
Figura 18 – Células <i>IDC</i> Positivo.	35
Figura 19 – Célula <i>IDC</i> Negativo.	36
Figura 20 – Célula <i>IDC</i> Positivo.	36
Figura 21 – Balanceamento do <i>dataset</i> .	37
Figura 22 – Separação do <i>dataset</i> .	37
Figura 23 – Normalização do <i>dataset</i> .	38
Figura 24 – Topologia CNN.	38
Figura 25 – Compilação do Modelo.	39
Figura 26 – Treinamento do Modelo.	39
Figura 27 – Acurácia x Loss.	40
Figura 28 – Matriz de Confusão.	40
Figura 29 – Predizendo Amostra.	41

Lista de abreviaturas e siglas

API	<i>Application Programming Interface</i>
CNN	<i>Convolutional Neural Networks</i>
IA	Inteligência Artificial
IDC	<i>Invasive Ductal Carcinoma</i>
ILC	<i>Invasive Lobular Cancer</i>
INCA	Instituto Nacional de Câncer
RNA	<i>Rede Neural Artificial</i>
RNN	<i>Recurrent Neural Network</i>

Sumário

1	INTRODUÇÃO	12
1.1	Objetivo	13
1.2	Problema	13
1.3	Hipótese	13
1.4	Justificativa	13
2	FUNDAMENTAÇÃO TEÓRICA	15
2.1	Câncer	15
2.1.1	Tipos Mais Comuns	15
2.1.2	No Brasil	16
2.1.3	Câncer de Mama	17
2.2	Tecnologia	19
2.2.1	História	19
2.2.2	Na Saúde	22
2.3	Inteligência Artificial	23
2.4	Aprendizado de Máquina	23
2.5	Redes Neurais	24
2.5.1	<i>Perceptron</i>	25
2.5.2	<i>Perceptron Multicamadas</i>	25
2.5.3	Redes Neurais Recorrentes	26
2.5.4	Redes Neurais Convolucionais	26
2.5.4.1	Métricas	28
3	METODOLOGIA	30
3.1	Etapas	30
3.2	Ferramentas	30
3.2.1	Google Colab	30
3.2.2	Keras	30
3.2.3	Matplotlib	31
3.2.4	NumPy	31
3.2.5	OpenCV	31
3.2.6	Pandas	31
3.2.7	Python	31
3.2.8	Scikit Learn	31
3.2.9	Seaborn	31
3.2.10	TensorFlow	32

3.2.11	Train Test Split	32
3.3	Cronograma	32
4	DESENVOLVIMENTO	33
4.1	Pré Processamento	34
4.2	Análise Exploratória	34
4.3	Modelo de Rede Neural	38
5	CONCLUSÃO	42
5.1	Trabalhos Futuros	42
	REFERÊNCIAS	43

1 Introdução

Não é de hoje que o câncer é uma das doenças que mais assola a humanidade. De acordo com o Ministério da Saúde, "câncer (ou tumor maligno) é o nome dado a um conjunto de mais de 100 doenças que têm em comum o crescimento desordenado de células"(MINISTÉRIO DA SAÚDE, 2020). Segundo Fayed (2020) o caso mais antigo relacionado a câncer, foi documentado no Egito, em 1500 antes de Cristo. Documentando casos de tumores de mama.

Em uma publicação realizada pelo (HOSPITAL CÂNCER BARRETOS, 2015) os tipos mais comuns da doença são:

Câncer de pele; Câncer de próstata; Câncer de cólon e reto; Câncer de pulmão; Câncer de mama; Câncer de estômago.

O Ministério da Saúde afirma que o tratamento do câncer pode ser feito por meio de uma ou de várias modalidades de tratamento combinadas. A principal delas é a cirurgia oncológica, que pode ser realizada em conjunto com radioterapia, quimioterapia ou transplante de medula óssea, conforme cada caso (MINISTÉRIO DA SAÚDE, 2020). É claro que o diagnóstico precoce auxilia e facilita muito o tratamento, uma vez que ele visa identificar sinais e sintomas iniciais da doença, conduzindo a terapias mais simples e menos invasivas. No Brasil, entre os anos de 2015 a 2018 ocorreu um total de mortalidade de 862.493 mil pessoas, sendo 454.964 homens e 407.464 mulheres, dados os quais retirados a partir do sistema [Atlas Online de Mortalidade \(2020\)](#). Desse número total de casos obtidos, representados pelo sexo feminino, 66.280 referem-se ao câncer de mama. Segundo o Inca, a estimativa para o ano de 2020 é de um total de 626.030 novos casos de câncer (INCA, 2020a).

Atualmente, a Inteligência Artificial (IA) tem ganhado cada vez mais espaço em todas as áreas, pois é um ramo de pesquisa da ciência da computação que busca, através de símbolos computacionais, construir mecanismos e/ou dispositivos que simulem a capacidade do ser humano de pensar, resolver problemas, além de ser também um campo de estudo acadêmico, auxiliando também em interpretações e reconhecimentos de laudos, sejam eles, imagens de radiografia, ressonâncias, tomografias, etc.

"A maior vantagem da Inteligência Artificial na medicina é, sem dúvidas, o auxílio na diagnose de patologias. Sabemos que nem todo caso é diagnosticado com facilidade, mas com essa tecnologia o processo se tornará consideravelmente mais tranquilo com análises muito mais seguras". ([CALDEIRA, 2017](#))

A proposta deste trabalho consiste em aplicar os recursos da IA com o uso de Redes Neurais (RNA) para o diagnóstico de células cancerígenas. No capítulo 2, encontra-se o referencial teórico, no capítulo 3, os materiais e métodos desse trabalho, no capítulo 4, o

desenvolvimento e no capítulo 5, a conclusão e finalmente as referências deste trabalho.

1.1 Objetivo

Partindo disso, o objetivo desse trabalho será focado em um tipo específico de câncer e um dos mais comuns, que é o câncer de mama. Tido isto, a ideia é construir uma topologia de RNA que obtenha a melhor precisão possível, em seguida realizar o teste com outras células e ver realmente se essa precisão continua fora do treinamento do modelo.

1.2 Problema

Quando falamos de problemas nessa tese, surgem e existem diversos, visto que este é um trabalho de certa forma ambicioso, uma vez que toca em um assunto muito delicado, o diagnóstico de tratamento do câncer. Para definir o melhor procedimento, o paciente passa por diversos doutores e especialistas da área, pois cada caso é muito específico e varia de paciente pra paciente.

Porém, diante disso como seria possível auxiliar na classificação das células, de modo a facilitar o tratamento precoce?

1.3 Hipótese

A partir de um banco de dados (*dataset*) bem estruturado e com informações pertinentes, um modelo de inteligência artificial com uma precisão e acurácia alta pode resolver diversos desses possíveis problemas, por exemplo. A partir de uma entrada, a IA pode classificar o tumor como maligno ou benigno, tendo como referência as características de outros tratamentos que tenham tido sucesso ou não, observando as semelhanças nos pacientes, de modo a poder vir a decidir qual o melhor procedimento a ser seguido para determinado paciente específico.

1.4 Justificativa

A junção de modelos de IA definidos por especialistas na matéria e a análise de uma volumetria de dados gigantesca, são capazes de propor soluções para problemas médicos ([LOBO, 2017](#)).

De acordo com [TOTVS \(2020\)](#), "através da inteligência artificial, a tecnologia genômica oferece avanços no tratamento de doenças como o câncer, possibilitando tratamentos mais

eficazes para a doença", através de diagnósticos mais precisos e eficientes, num período bem menor, de modo a acarretar o início do procedimento logo no começo da doença, uma vez que o diagnóstico precoce é uma estratégia que possibilita tratamento e terapias mais simples e efetivas, contribuindo para a redução do estágio de apresentação do câncer ([MINISTÉRIO DA SAÚDE, 2010](#)).

Mediante ao tratamento precoce, pode-se obter uma porcentagem maior de casos estáveis (sem agravamentos), facilitando também o tomar de decisões sobre o tratamento, levando a uma vida mais saudável seja o procedimento para curar o câncer, controlar ou tratar os problemas provocados pela doença.

Justifica-se também a proposta deste trabalho pelo fato de solidificar e ampliar conceitos estudados no decorrer do curso de Ciência da Computação.

2 Fundamentação Teórica

2.1 Câncer

Pode-se entender como "câncer", segundo publicação no site do [MINISTÉRIO DA SAÚDE \(2020\)](#):

"Câncer (ou tumor maligno) é o nome dado a um conjunto de mais de 100 doenças que têm em comum o crescimento desordenado de células. Dividindo-se rapidamente, estas células agrupam-se formando tumores, que invadem tecidos e podem invadir órgãos vizinhos e até distantes da origem do tumor (metástases). O câncer é causado por mutações, que são alterações da estrutura genética (DNA) das células. Cada célula sadias possui instruções de como devem crescer e se dividir. Na presença de qualquer erro nestas instruções (mutação), pode surgir uma célula doente que, ao se proliferar, causará um câncer. O câncer pode surgir em qualquer parte do corpo. Entretanto, alguns órgãos são mais afetados do que outros; e cada órgão, por sua vez, pode ser acometido por tipos diferenciados de tumor, mais ou menos agressivos".

2.1.1 Tipos Mais Comuns

Dando continuidade ao que foi abordado na introdução. O [HOSPITAL CÂNCER BARRETOS \(2015\)](#) realizou uma publicação no site comentando sobre os tipos mais comuns de câncer, sendo eles:

- Câncer de pele: É um proliferação incontrolável de células cutâneas anormais;
- Câncer de próstata: Esse é o mais comum câncer do tipo sólido entre os homens. As chances dele aparecer ficam maiores conforme a idade vai avançando;
- Câncer de colôn e reto: A doença começa na camada superficial do revestimento intestinal e com o tempo vai atingindo as camadas mais profundas;
- Câncer de pulmão: Não apresenta sintomas em suas fases iniciais, tornando o diagnóstico mais difícil de ser feito e, por isso, a maioria descobre quando o câncer já está avançado;
- Câncer de mama: As células perdem sua função normal e passam a desenvolver atividades anormais, como um crescimento desorganizado, formando um tumor;
- Câncer de estômago: Conforme a evolução da doença, as células cancerígenas vão substituindo o tecido normal do órgão gradativamente.

2.1.2 No Brasil

Sobre a proliferação da doença, segundo [Teixeira \(2012\)](#):

"Apesar dos avanços conceitual e normativo, e embora as evidências científicas demonstrem que nos países em desenvolvimento cerca de um terço dos cânceres possam ser prevenidos e outro terço evitado, o câncer é a segunda maior causa de morte no Brasil".

Pode-se afirmar que para o enfrentamento do câncer, não basta um país ser apenas desenvolvido tecnologicamente, uma vez que a tecnologia por si só auxilia na descoberta e tratamento da doença, porém não é capaz de resolver problemas como esse, se utilizada sem os recursos necessários.

Ao analisar os dados retirados do [Atlas Online de Mortalidade \(2020\)](#) obteve-se entre os anos de 2000 e 2009 o seguinte total, conforme a Figura 1:

Figura 1 – Mortalidade Entre 2000 e 2009.

Faixa Etária	Homens		Mulheres	
	Número de Óbito	Taxa Específica	Número de Óbito	Taxa Específica
00 a 04	4.058	4,74	3.391	4,12
05 a 09	3.652	4,18	2.843	3,38
10 a 14	3.557	4,05	2.907	3,42
15 a 19	5.397	6,05	3.740	4,31
20 a 29	12.517	7,45	11.906	7,17
30 a 39	21.530	15,79	32.359	23,4
40 a 49	66.971	61,2	78.278	68,72
50 a 59	138.260	190,52	120.789	153,73
60 a 69	193.683	459,94	144.944	294,93
70 a 79	201.797	911,96	149.944	521,8
80 ou mais	123.602	1.512,61	110.826	854,45
Idade ignorada	445	0	271	0
Total			1.440.845	

Fonte: Adaptada [Atlas Online de Mortalidade \(2020\)](#).

Já no intervalo entre os anos de 2010 a 2018 temos um aumento de mais de 400 mil óbitos, sendo possível ver a diferença na Figura 2:

Figura 2 – Mortalidade Entre 2010 e 2018.

Faixa Etária	Homens		Mulheres	
	Número de Obito	Taxa Específica	Número de Obito	Taxa Específica
00 a 04	3192	4,67	2821	4,31
05 a 09	2965	4,05	2308	3,29
10 a 14	3134	4,01	2493	3,32
15 a 19	4883	6,24	3203	4,22
20 a 29	12185	7,81	11531	7,48
30 a 39	22194	15,32	36214	24,74
40 a 49	62563	54,19	83364	69,52
50 a 59	165856	187,27	155438	162,19
60 a 69	245459	460,87	192038	311,42
70 a 79	247806	956,51	187976	551,46
80 ou mais	186180	1751	167578	959,84
Idade ignorada	258	0	73	0
Total		1.805.164		

Fonte: Adaptada [Atlas Online de Mortalidade \(2020\)](#).

Por mais que a tecnologia tenha evoluído, os números continuaram aumentando e cabe a nós desenvolvedores, criar soluções que realmente tenham impacto positivo nessa crescente estatística, que aumenta exponencialmente a cada ano.

2.1.3 Câncer de Mama

Este trabalho abordará o câncer de mama, sendo este, um dos tipos mais comuns no Brasil, considerando também a disponibilidade do *dataset*, onde o mesmo é composto por milhares de imagens com subtipo Carcinoma Ductal Invasivo (*IDC*).

O câncer de mama é uma doença heterogênea com múltiplos prognósticos. O Carcinoma Ductal Invasivo (*IDC*) e o Câncer Lobular Invasivo (*ILC*), que são classificados por suas diferentes estruturas histológicas e histórias de progressão, sendo os dois tipos principais de câncer de mama ([MARTINEZ, 2017](#)).

“Uma das primeiras representações do câncer de mama aparece na pintura A Noite, de Michele di Ridolfo del Ghirlandaio, provavelmente pintada entre 1553 e 1555, inspirada em uma escultura de Michelangelo. Na imagem, a mulher nua, que está reclinada e dormindo, em um mundo de sonhos, tem o seio esquerdo menor do que o direito, e o mamilo retráido, sinais típicos de câncer.” ([SILVESTRE, 2020](#))

Figura 3 – La Notte.



Fonte: Michele di Ridolfo del Ghirlandaio.

[Silvestre \(2020\)](#) afirma que por mais que o autoexame não prove uma queda na mortalidade, a mulher deve ser estimulada a conhecer seu corpo e perceber alterações suspeitas, por meio da observação e palpação ocasionais de suas mamas, sem periodicidade e técnica padronizadas.

Em publicação [INCA \(2020b\)](#) comenta que, "quando a doença é diagnosticada no início, o tratamento tem maior potencial curativo. Quando há evidências de metástases, o tratamento tem por objetivos principais prolongar a sobrevida e melhorar a qualidade de vida", tendo as modalidades de tratamento divididas em duas formas:

- Tratamento Local: Cirurgia e radioterapia (além de reconstrução mamária);
- Tratamento Sistêmico: Quimioterapia, hormonioterapia e terapia biológica.

O prognóstico está relacionado ao estadiamento, o qual possui 3 níveis:

- Estadios I e II: Consiste de cirurgia, que pode ser conservadora, com retirada apenas do tumor ou mastectomia, com retirada da mama e reconstrução mamária;
- Estadio III: Tratamento sistêmico (na maioria das vezes, com quimioterapia) é a modalidade terapêutica inicial;
- Estadio IV: Nesse estadio, é fundamental a busca do equilíbrio entre a resposta tumoral e o possível prolongamento da sobrevida, tendo por modalidade principal sistêmica.

2.2 Tecnologia

A seguir vamos dissertar um pouco sobre a história da computação, passando por alguns marcos importantes para a mesma, deixando de lado os mais conhecidos como Alan Turing, Steve Jobs e Bill Gates, que foram fundamentais em toda a evolução como um todo, e logo em seguida, comentar sobre o atual momento e a utilização da tecnologia na saúde.

2.2.1 História

Para começar nesse túnel do tempo, vamos para 1901 onde foi realizado a descoberta do Mecanismo de Anticítera, que tratava-se de uma calculadora astronômica, criada no século I a.C ([BRUDERER, 2020](#)).

Figura 4 – Mecanismo de Anticítera.



Fonte: [Bruderer \(2020\)](#)

No ano de 1815, nasceu Augusta Ada Byron, mais conhecida como Ada Lovelace criadora do primeiro algoritmo utilizado para cálculos matemáticos ([GNIPPER, 2016](#)).

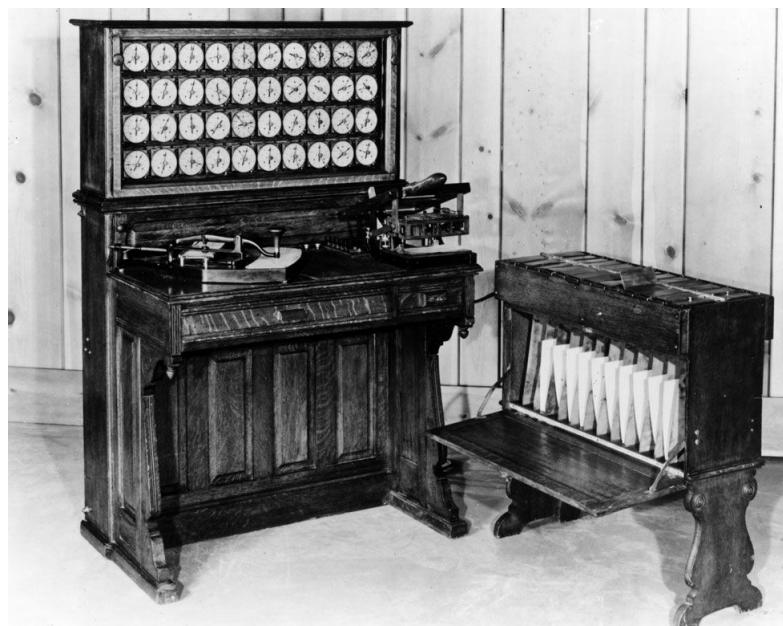
Figura 5 – Ada Lovelace.



Fonte: Adaptado [Gnipper \(2016\)](#)

Em 1860, nascia o engenheiro norte-americano Herman Hollerith, que ficou conhecido por desenvolver um sistema baseado em cartões perfurados. Sendo o primeiro a realizar processamento de dados, deu origem anos depois a tão conhecida IBM sob influência de Thomas J. Watson ([INFOPÉDIA, 2020](#)), empresa a qual é responsável por diversos marcos na tecnologia como os fundamentos de banco de dados relacionais.

Figura 6 – Tabulador e classificador de Herman Hollerith.



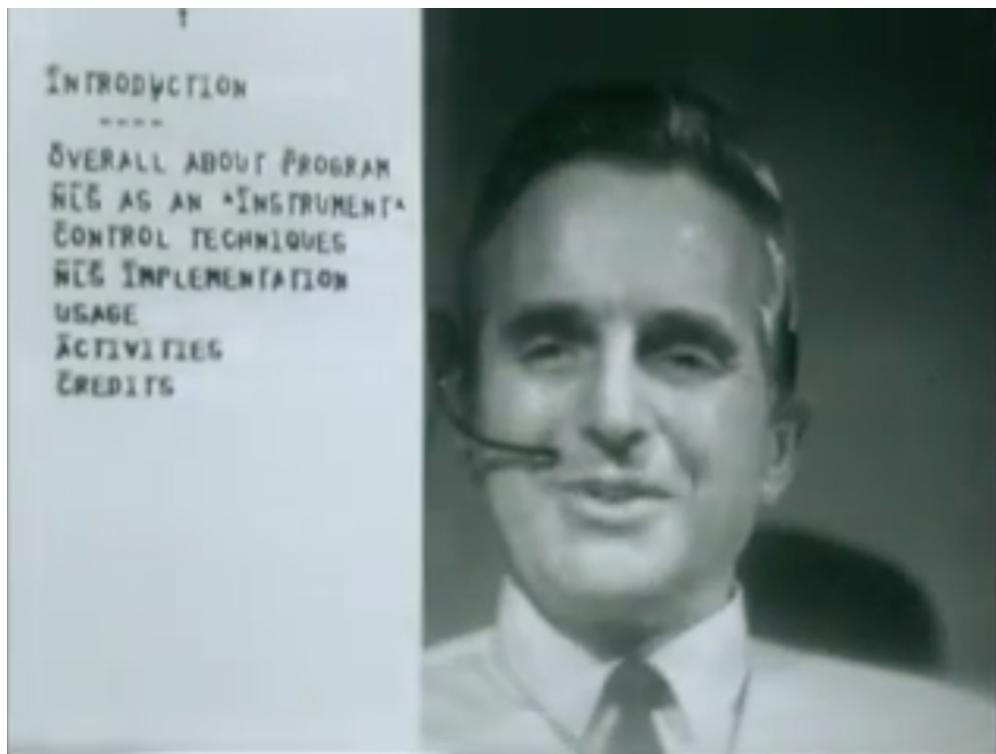
Fonte: Autor desconhecido.

Em 1959, foi ano de lançamento de uma das linguagens que se perpetua até hoje,

denominada COBOL, desenvolvida pela analista de sistema da marinha Grace Hopper ([STRAWN, 2015](#)).

Conhecida como "*Mother of All Demos*" foi uma apresentação de Douglas Engelbart em uma conferência, no ano de 1968. Tratava-se de um sistema de computação interativa, onde ocorreu a primeira aparição do tão conhecido mouse, também foi demonstrado uma videoconferência em tempo real, edição de texto, hipertexto e o funcionamento de janelas ([MARKOFF, 2013](#)).

Figura 7 – "*Mother of All Demos*".

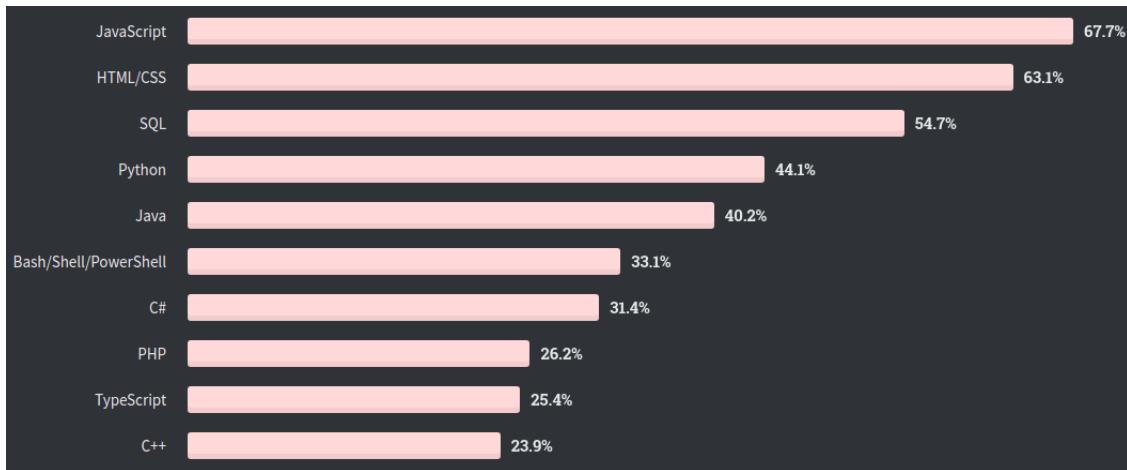


Fonte: [Markoff \(2013\)](#).

Com o passar dos anos, diversas dessas tecnologias criadas anteriormente foram sendo aprimoradas, dando origem a novos nichos, computadores portáteis, *smartphones* e novas linguagens de programação.

A Figura 8 a seguir mostra o top 10 das linguagens de programação, *script* e marcação mais populares em uma pesquisa realizada pelo [STACKOVERFLOW \(2020\)](#):

Figura 8 – Tecnologias Mais Populares.



Fonte: [STACKOVERFLOW \(2020\)](#).

É possível notar a grande explosão de novos desenvolvedores utilizando JavaScript e Python nos últimos anos. Tal utilização se deve pela facilidade do JavaScript em atuar em mais de uma plataforma, utilizando a mesma linguagem. Já o Python está muito ligado ao aumento considerável da utilização de Inteligência artificial e aprendizado de maquina por grandes empresas como Google, Facebook, Spotify, etc. Dissertarei melhor sobre inteligência artificial em um tópico específico.

2.2.2 Na Saúde

No que tange a tecnologia na área da saúde, é possível perceber a utilização da mesma em hospitais de médio a grande porte com a utilização de equipamentos de ultima geração. É possível ver a utilização de dispositivos móveis, para auxiliar o serviço dos profissionais da saúde, como no caso de um enfermeiro que utiliza de um *tablet* para realizar as consultas, checagens, entre outros procedimentos. Ao falar dos sistemas utilizados que englobam diferentes postos, quase sempre manuseando o mesmo banco de dados, em diferentes cidades e lugares, com diversas tecnologias, até mesmo de inteligência artificial. Por exemplo, um hemocentro ao utilizar do aprendizado de maquina para prever o período de maior risco, ou seja, um período que possa ficar sem estoque de algum tipo sanguíneo específico. Sendo bem grande a quantidade de possibilidades da utilização de IA na área.

O sistema que marcou o uso de IA na medicina foi o Sistema Especialista MYCIN, tendo o inicio de seu desenvolvimento em 1972, na Universidade de Standfor. O sistema consiste em realizar diagnósticos a partir de entradas incertas ou incompletas. Foi desenvolvido com o intuído de identificar bactérias causadoras de infecções graves, recomendando o tratamento específico, com doses ajustadas conforme o peso do paciente ([GUARIZI, 2014](#)). "Para o seu desenvolvimento, foram necessárias diversas e extensas entrevistas com especialistas na área. Consultas a

livros didáticos e estudos de casos anteriores também foram necessários"(GOLDSCHMIDT, 2010).

2.3 Inteligência Artificial

É nítido que a Inteligência Artificial está cada vez mais popular, grandes empresas estão mudando o *core* de suas aplicações, migrando para sistemas com aprendizado de máquina.

Segundo Goldschmidt (2010) a IA vai além da perspectiva de compreensão do pensamento humano, pois ela também busca construir entidades artificiais inteligentes.

Algumas habilidades que necessariamente envolvem inteligência, podem ser citadas:

- Capacidade de raciocínio, dedução e inferência;
- Capacidade de aprendizado;
- Capacidade de percepção;
- Capacidade de evolução e adaptação.

Entre as habilidades envolvendo ideias, pontos de vista, conceitos e técnicas de diversas áreas, dentre as quais é possível citar:

Filosofia; Matemática; Economia; Neurociência; Psicologia; Ciência da computação e Linguística.

2.4 Aprendizado de Máquina

Khanna e Awad (2015) diz que o aprendizado de máquina é um dos ramos da IA, que busca sintetizar relações implícitas em um conjunto de dados, aplicadas em diversas áreas de interesse, como buscas na internet, análise de crédito, saúde, esportes, entre diversas outras.

Problemas de aprendizado de máquina podem ser categorizados em (SCIKIT LEARN, 2019):

- Aprendizado Supervisionado: Caso em que os dados possuem atributos adicionais associados às entradas e que desejam ser preditos. Este tipo de problema se subdivide em:
 - Classificação: Há amostras de dados já rotuladas pertencentes a duas ou mais classes e deseja-se prever amostras não rotuladas;

- Regressão: Caso onde a saída desejada consiste de uma ou mais variáveis contínuas;
- Aprendizado Não-supervisionado: É o tipo de aprendizado em que se é fornecido uma série de vetores sem um valor correspondente. O objetivo neste tipo de problema, pode ser descobrir grupos com amostras similares, redução de dimensionalidade, etc.

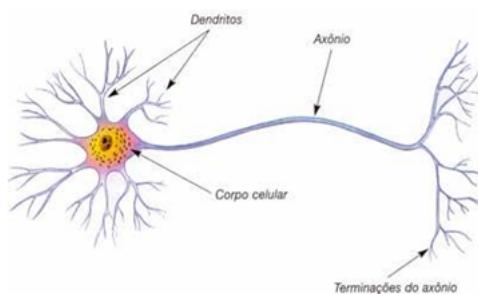
2.5 Redes Neurais

Segundo [Hecht-Nielsen \(1988\)](#), uma RNA pode ser formalmente definida como:

“Uma estrutura que processa informação de forma paralela e distribuída e que consiste de unidades computacionais (as quais podem possuir memória local e podem executar operações locais) interconectadas por canais unidirecionais chamados de conexões. Cada unidade possui uma única conexão de saída, que pode ser dividida em quantas conexões laterais se fizer necessário, sendo que cada uma destas conexões transporta o mesmo sinal (sinal de saída da unidade). Esse sinal de saída pode ser contínuo ou discreto. O processamento executado por cada unidade pode ser definido 74 arbitrariamente, com a restrição de que ele deve ser completamente local, isto é, deve depender somente dos valores atuais dos sinais de entrada que chegam até a unidade via as conexões e dos valores armazenados na memória local da unidade computacional.”

Mas como é o funcionamento de um neurônio? A Figura 9 a seguir, exemplifica:

Figura 9 – Neurônio.



Fonte: Autor Desconhecido.

- Dendritos: São responsáveis por captar informações;
- Corpo celular: Responsáveis por processar as informações;
- Axônio: É responsável por distribuir as informações processadas para outros neurônios ou células do corpo.

[Goldschmidt \(2010\)](#) diz que, em relação a topologia, existem algumas classificações utilizadas na caracterização da arquitetura ou topologia das redes neurais artificiais. Sendo:

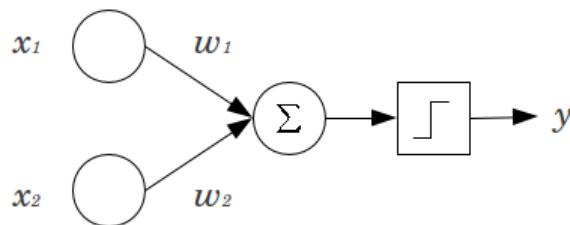
- Redes de camada única (*perceptron*);
- Redes de múltiplas camadas (*perceptron* multicamadas);
- Redes *feedforward*;
- Redes *feedback*;
- Redes com recorrência auto associativa.

2.5.1 Perceptron

O modelo de rede neural mais antigo é o perceptron. Tem como objetivo classificar corretamente o conjunto de estímulos aplicados externamente à rede (GOLDSCHMIDT, 2010). Sendo a unidade mais básica de uma rede neural, tem apenas a capacidade de classificar sim ou não, a partir de uma regressão que recebe seus valores e calcula uma saída.

Como é possível ver na Figura 10 abaixo:

Figura 10 – *Perceptron*.

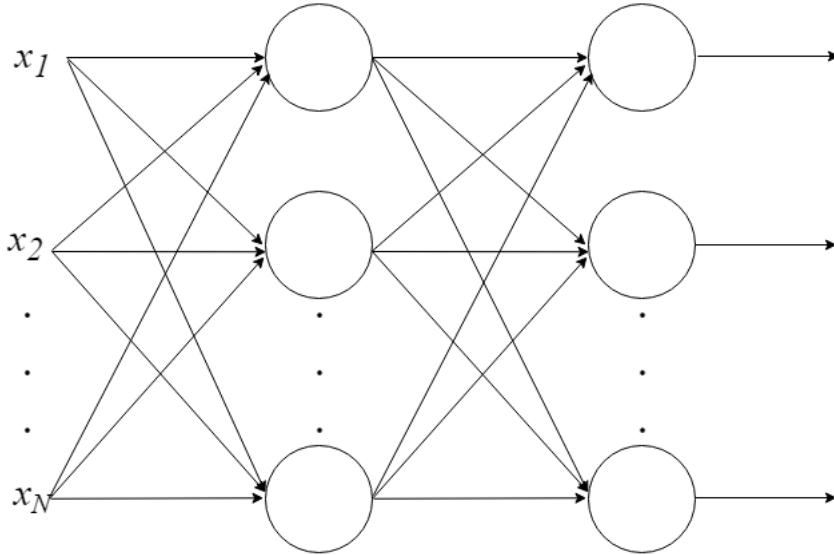


Fonte: Autor Desconhecido.

2.5.2 Perceptron Multicamadas

Diferente da Perceptron, a Perceptron Multicamadas pode ter mais de um neurônio na camada de saída, sendo um modelo mais complexo, contendo vários neurônios e várias camadas neurais, resolvendo não só problemas lineares, mas não lineares também, onde cada neurônio possui uma saída independente.

Figura 11 – *Perceptron Multicamadas.*



Fonte: Autor Desconhecido.

2.5.3 Redes Neurais Recorrentes

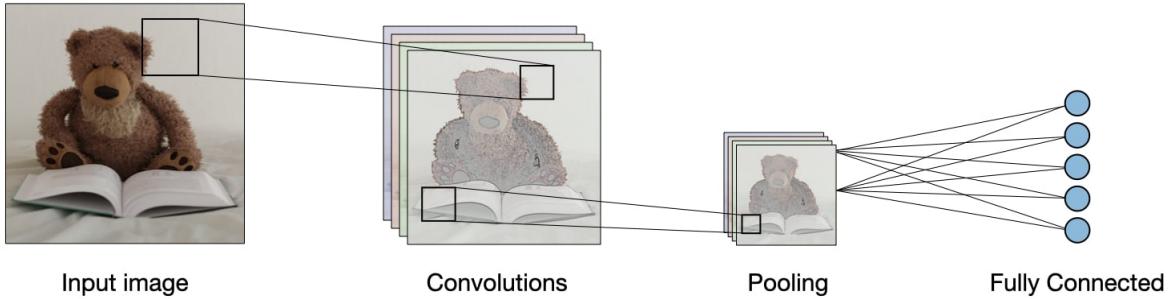
[Géron \(2019\)](#) diz que uma classe de redes podem prever o futuro até certo ponto, de modo a poder analisar dados de séries temporais (como preços de ações), e dizer quando comprar ou vender. No caso de sistemas de direção autônoma, pode antecipar as trajetórias dos carros, de modo a ajudar a evitar acidentes. Em outras palavras, podem trabalhar em sequências de comprimentos arbitrários, em vez de em entradas de tamanho fixo.

2.5.4 Redes Neurais Convolucionais

Redes convolucionais usam uma arquitetura especial, que é particularmente bem adaptada para classificar imagens. Usar essa arquitetura torna as redes convolucionais mais rápidas para treinar. Isso, por sua vez, nos ajuda a treinar redes profundas de muitas camadas, que são muito boas para classificar imagens. Hoje, as redes convolucionais profundas ou alguma variante próxima são usadas na maioria das redes neurais para reconhecimento de imagem ([NIELSEN, 2017](#)).

Segundo [Amidi \(2020\)](#) os tipos de camadas que geralmente compõem uma *CNN* são:

Figura 12 – Arquitetura de uma CNN Tradicional.



Fonte: [Amidi \(2020\)](#)

- Camada de Convolução: A camada de convolução usa filtros que realizam operações de convolução, enquanto verifica a entrada, no que diz respeito às suas dimensões. Seus hiperparâmetros incluem o tamanho do filtro e caminhar. A saída resultante é chamada de mapa de recursos ou mapa de ativação;
- *Pooling*: A camada de *pooling* é uma operação geralmente aplicada após uma camada de convolução, que faz alguma invariância espacial. Em particular, o agrupamento máximo e médio são tipos especiais de agrupamento em que o valor "máximo e médio" é obtido, respectivamente;
- *Fully Connected*: A camada totalmente conectada, opera em uma entrada plana, onde cada entrada é conectada a todos os neurônios. Se presentes, as camadas FC são geralmente encontradas no final das arquiteturas CNN e podem ser usadas para otimizar objetivos, como pontuações de classe.

Em relação aos hiperparâmetros, serão utilizados os seguintes disponíveis no módulo de [TensorFlow \(2020a\)](#):

- *Conv2D*: Essa camada cria um *kernel* de convolução que é convolvido com a entrada da camada para produzir um tensor de saídas:
 - *filters*: Inteiro, a dimensionalidade do espaço de saída (ou seja, o número de filtros de saída na convolução);
 - *kernel_size*: Um inteiro ou tupla / lista de 2 inteiros, especificando a altura e a largura da janela de convolução 2D. Pode ser um único inteiro para especificar o mesmo valor para todas as dimensões espaciais;
 - *activation*: Função de ativação a ser usada. Se você não especificar nada, nenhuma ativação será aplicada;

- *input_shape*: A primeira camada em um modelo deve receber o *shape* das imagens do treinamento.
- *MaxPooling2D*: Reduz a resolução da representação de entrada tomando o valor máximo sobre a janela definida por *pool_size* para cada dimensão ao longo do eixo dos recursos;
- *Flatten*: Nivela a entrada. Não afeta o tamanho do lote;
- *Dropout*: Define aleatoriamente as unidades de entrada para 0, com uma frequência de ‘rate’ em cada etapa durante o tempo de treinamento, o que ajuda a evitar *overfitting*;
- *Dense*: Implementa a operação $output = activation(dot(input, kernel) + bias)$ onde *activation* é a função de ativação elemento a elemento passada como o *activation* argumento, *kernel* é uma matriz de pesos criada pela camada e ‘bias’ é um vetor de polarização criado pela camada (aplicável apenas se *use_bias* for *True*).

Este trabalho consistirá em uma classificação a partir de imagens, portanto, será utilizada um aprendizado supervisionado, utilizando a topologia de uma rede neural convolucional.

2.5.4.1 Métricas

Após a criação da topologia, inicia-se o treinamento do modelo. Ao finalizar, é importante que se faça uma avaliação do algoritmo para testar se seu comportamento está de acordo com o esperado. Ou seja, é necessário métricas que quantifiquem e qualifiquem a eficácia das previsões. Uma métrica bastante utilizada para realizar essas mensurações é a matriz de confusão, que compara as classes preditas com as classes reais, utilizando as seguintes nomenclaturas ([KOHAVI, 1998](#)):

- *True Positive*: Ocorre quando no conjunto real, a classe que estamos buscando foi prevista corretamente;
- *True Negative*: Ocorre quando no conjunto real, a classe que estamos buscando prever foi prevista incorretamente;
- *False Positive*: Ocorre quando no conjunto real, a classe que não estamos buscando prever foi prevista corretamente;
- *False Negative*: Ocorre quando no conjunto real, a classe que não estamos buscando prever foi prevista incorretamente.

A partir disso, são derivadas medidas que dão uma noção do desempenho do modelo. Algumas delas são ([SCIKIT LEARN, 2019](#)):

- *Accuracy*: $\frac{TP+TN}{TP+TN+FP+FN}$
- *Precision* (P): $\frac{TP}{TP+FP}$
- *Recall* (R): $\frac{TP}{TP+FN}$
- *F1-Score*: $\frac{2.P.R}{P+R}$

As métricas acima, são apresentadas com números entre 0 e 1, podendo variar dependendo da implementação, sendo que, quanto maior este valor, melhor seu desempenho.

Uma técnica para avaliar a capacidade de generalização de um modelo treinado é a validação cruzada. Esta técnica prevê a separação de um conjunto de testes para a etapa de treinamento do modelo, de forma que se avalie seu desempenho, com um conjunto não visto antes pelo algoritmo ([KHANNA; AWAD, 2015](#)).

3 Metodologia

3.1 Etapas

A pesquisa desenvolvida neste trabalho, consiste em duas etapas, a primeira trata-se de uma pesquisa bibliográfica, a qual visa reunir as informações e dados, que servirão de base para a construção da investigação proposta.

O levantamento bibliográfico foi realizado a partir da análise das seguintes fontes: livros, artigos, documentos monográficos, periódicos (jornais, revistas, etc.), sites, entre outros locais que apresentam um conteúdo documentado.

Após a seleção do material, foi realizada uma leitura do mesmo, analisando e interpretando, para o correto entendimento de todo o assunto envolvido no tema deste trabalho.

A segunda etapa, trata-se da parte prática da proposta, sendo o primeiro passo, um pré processamento nas imagens, pois o *dataset* que será utilizado, possui algumas imagens fora do padrão, impactando no treinamento do modelo. Em seguida será realizada uma análise exploratória, para apresentar indicadores e pontos importantes do conjunto de dados, como por exemplo o balanceamento das amostras. Por fim, o desenvolvimento de um modelo de aprendizado de máquina, utilizando redes convolucionais para identificar se o tumor é benigno ou maligno.

3.2 Ferramentas

Não é possível lincar todas as bibliotecas que foram utilizadas, uma vez que esta, possui outras dependências, porém, as mais utilizadas e com maior importância estarão referenciadas.

3.2.1 Google Colab

Colaboratory, ou *Colab*, permite que você escreva e execute *Python* no seu navegador ([GOOGLE COLAB, 2020](#)).

3.2.2 Keras

Keras segue as práticas recomendadas para reduzir a carga cognitiva: oferece *API* consistente e simples, minimiza o número de ações do usuário, necessárias para casos de uso comuns e fornece mensagens de erro claras e açãoáveis ([KERAS, 2020](#)).

3.2.3 Matplotlib

Matplotlib é uma biblioteca abrangente, para a criação de visualizações estáticas, animadas e interativas em *Python*. ([MATPLOTLIB, 2020](#)).

3.2.4 NumPy

NumPy é um projeto de código aberto, com o objetivo de permitir a computação numérica com *Python*. Foi criado em 2005, com base no trabalho inicial das bibliotecas *Numerical* e *Numarray* ([NUMPY, 2005](#)).

3.2.5 OpenCV

OpenCV (Open Source Computer Vision Library) é uma biblioteca de software de visão computacional e aprendizado de máquina de código aberto ([OPENCV, 2020](#)).

3.2.6 Pandas

Pandas é uma ferramenta de análise e manipulação de dados de código aberto, rápida, poderosa, flexível e fácil de usar, construída sobre a linguagem de programação *Python* ([PANDAS, 2008](#)).

3.2.7 Python

Python é uma linguagem de programação que permite trabalhar rapidamente e integrar sistemas de forma mais eficaz ([PYTHON, 2020](#)).

3.2.8 Scikit Learn

Ferramentas simples e eficientes para análise preditiva de dados, acessível a todos e reutilizável em vários contextos ([SCIKIT LEARN, 2020](#)).

3.2.9 Seaborn

É uma biblioteca de visualização de dados *Python*, baseada em *Matplotlib* [3.2.3](#) . Ele fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e informativos ([SEABORN, 2020](#)).

3.2.10 TensorFlow

"A principal biblioteca de código aberto para desenvolver e criar modelos de ML. Execute notebooks do *Colab* diretamente no navegador para começar rapidamente"(TENSORFLOW, 2020b).

3.2.11 Train Test Split

Divide listas ou matrizes em treinamento aleatório e subconjuntos de teste (TRAIN TEST SPLIT, 2020).

3.3 Cronograma

A Figura 13 a seguir mostra o cronograma seguido durante o desenvolvimento do trabalho.

Figura 13 – Cronograma.

Atividades	2020											
	FEV	MAR	ABR	MAI	JUN	JUL	AGO	SET	OUT	NOV	DEZ	
Levantamento bibliográfico	X	X	X	X								
Estudo material		X	X	X								
Elaboração do projeto de pesquisa			X	X								
Relatório parcial			X	X								
Apresentação pré-banca			X	X	X							
Desenvolvimento do trabalho [...]						X	X	X	X			
[...]						X	X	X	X	X		
Relatório final (monografia)												

Fonte: Elaborada pelo autor.

4 Desenvolvimento

Foi utilizado um *dataset* disponível em uma competição no [Kaggle \(2020\)](#). O conjunto de dados original consistia em 162 imagens de slides inteiros de espécimes de câncer de mama, escaneados a 40x. A partir daí, 277.524 imagens de tamanho 50x50 foram extraídos,

- 198.738 *IDC* negativos;
- 78.786 *IDC* positivos.

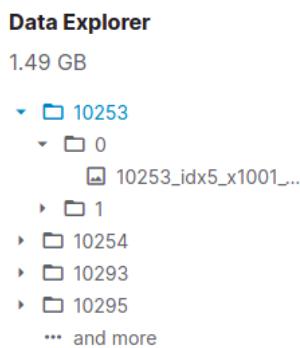
O nome de arquivo de cada imagem tem o formato:

- uxXyYclassC.png;;
- 10253idx5x1351y1101class0.png.

Onde u é a ID do paciente (10253idx5), X é a coordenada x de onde esta imagem foi cortada, Y é a coordenada y de onde esta imagem foi cortada e, C indica a classe onde 0 é *IDC* negativo e 1 é *IDC* positivo.

O *dataset* é estruturado da seguinte forma, conforme Figura 14:

Figura 14 – Estrutura do *dataset*



Fonte: [Kaggle \(2020\)](#)

O desenvolvimento consiste em 3 passos. Começando por um pré processamento no *dataset*, especificamente nas imagens, pois as mesmas não estão totalmente balanceadas e padronizadas. Após, será realizado uma análise exploratória das imagens, buscando indicadores das amostras, a serem utilizados posteriormente. A finalização consta com o desenvolvimento do modelo de rede convolucional para classificar o *IDC* da célula.

4.1 Pré Processamento

A função abaixo recebe três parâmetros, sendo *imagePath*, *width* e *height*. Ao realizar a leitura da imagem a partir do primeiro parâmetro, deve-se em seguida redimensionar a imagem para 50x50, pois algumas imagens do *dataset* estão em outras dimensões, podendo impactar no treinamento do modelo. Como demonstrado na Figura 15 abaixo:

Figura 15 – Função Pré Processamento

```
1 def getImage(imagePath, width=50, height=50):
2     image = cv2.imread(imagePath, cv2.IMREAD_COLOR)
3     imageColor = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
4     imageNormalized = cv2.resize(imageColor, (width, height), interpolation=cv2.INTER_AREA)
5     return imageNormalized
```

Após a normalização das imagens é salvo um arquivo binário contendo as informações do *dataframe*, podendo recuperar o mesmo, e não precisar realizar o passo do pré processamento novamente.

4.2 Análise Exploratória

Esta sessão está relacionada ao levantamento realizado durante a análise do *dataset*, que consiste em comparar os dados, identificar os mais predominantes e visualizar algumas imagens, para facilitar na elaboração da topologia do modelo.

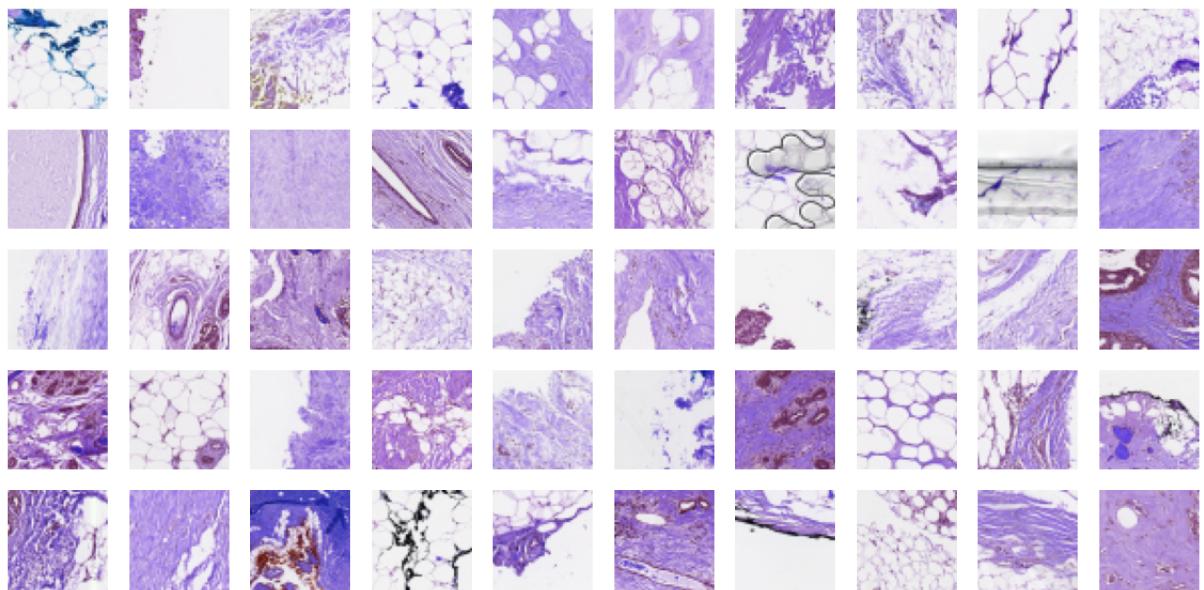
A Figura 16 abaixo, representa como os dados estão dispostos no *dataframe*, onde *path* é o caminho lógico de onde a imagem está localizada, *label* é o rótulo para classificar se a célula contém ou não o *IDC* e, *image* é composta pelo *array* de *pixels* da imagem.

Figura 16 – Dataframe.

Shape do dataframe (44006, 3). Estrutura do dataframe:			
	path	label	image
0	/content/drive/My Drive/UNIP/TCC/MODELO/datafr...	1	[[[213, 162, 190], [220, 169, 191], [228, 175,...
1	/content/drive/My Drive/UNIP/TCC/MODELO/datafr...	1	[[[210, 155, 183], [224, 169, 192], [187, 137,...
2	/content/drive/My Drive/UNIP/TCC/MODELO/datafr...	1	[[[158, 122, 164], [167, 136, 175], [219, 166,...
3	/content/drive/My Drive/UNIP/TCC/MODELO/datafr...	1	[[[228, 123, 157], [229, 103, 139], [229, 99, ...
4	/content/drive/My Drive/UNIP/TCC/MODELO/datafr...	1	[[[226, 146, 169], [230, 164, 186], [232, 166,...

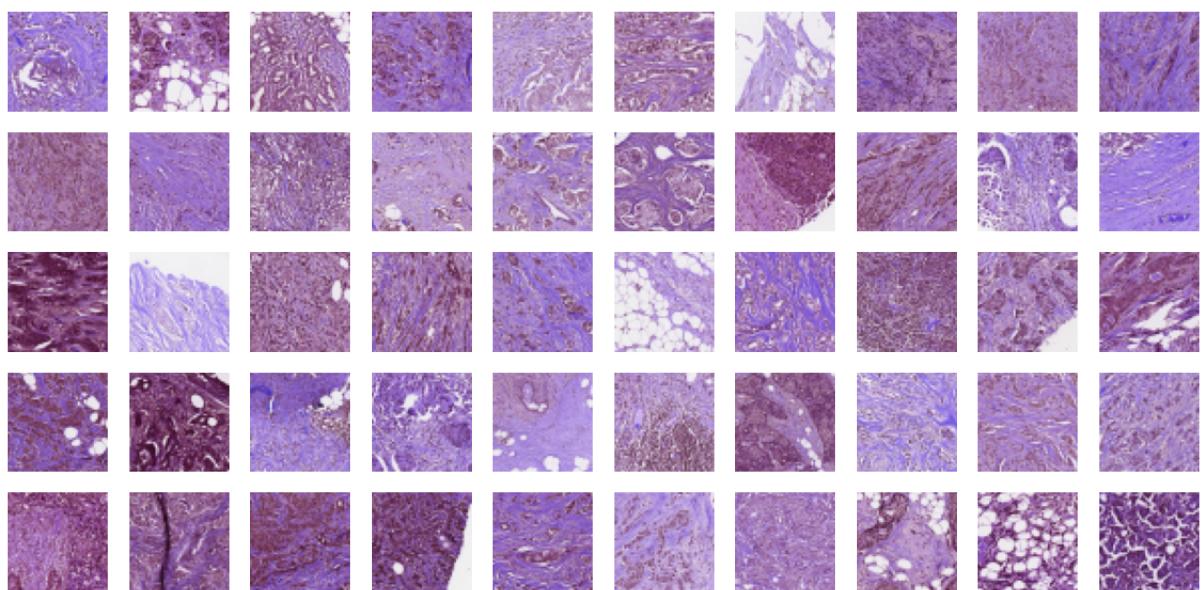
A Figura 18 exibe imagens com células que possuem *IDC* negativo.

Figura 17 – Células *IDC* Negativo.



Já a Figura 17 exibe imagens com células que possuem *IDC* positivo.

Figura 18 – Células *IDC* Positivo.



Foi criada uma função para analisar o mapa de cor das imagens, com o intuito de buscar *insights*, como demonstra as duas Figuras abaixo:

Figura 19 – Célula *IDC* Negativo.

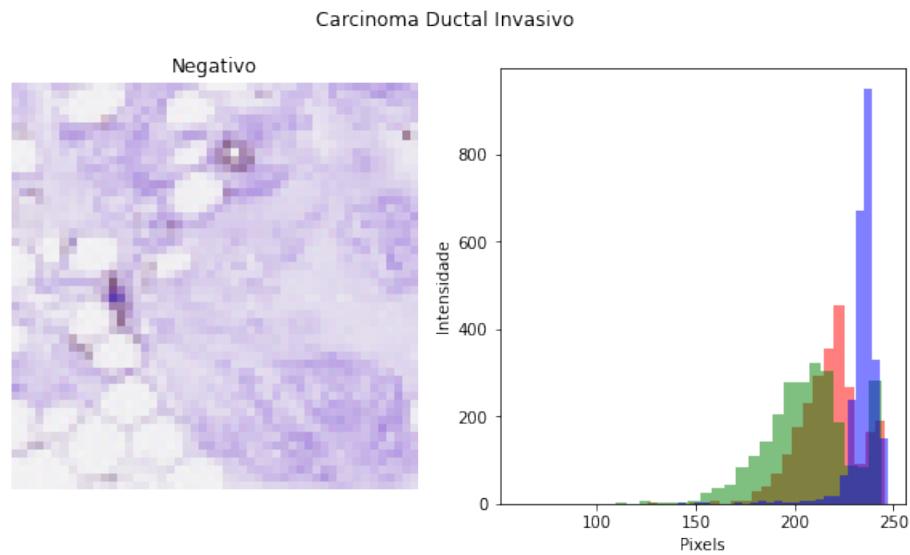
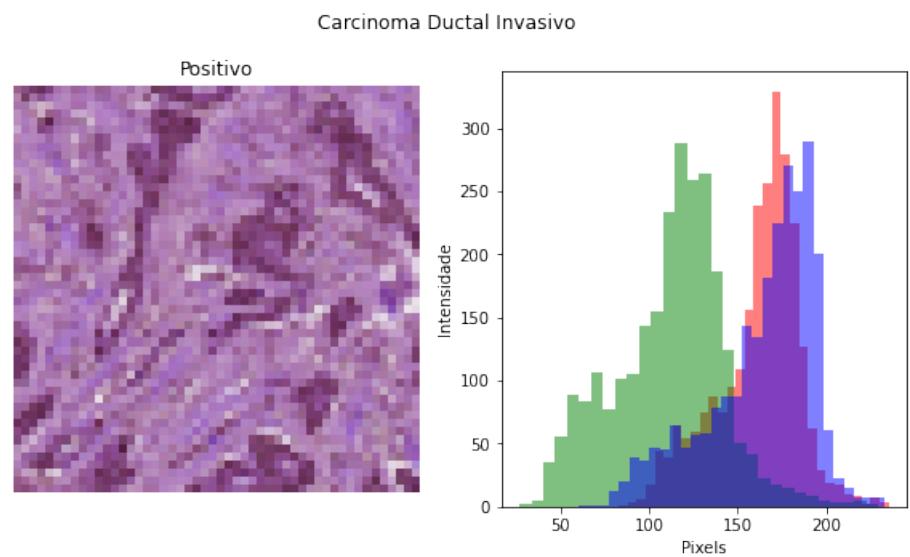


Figura 20 – Célula *IDC* Positivo.

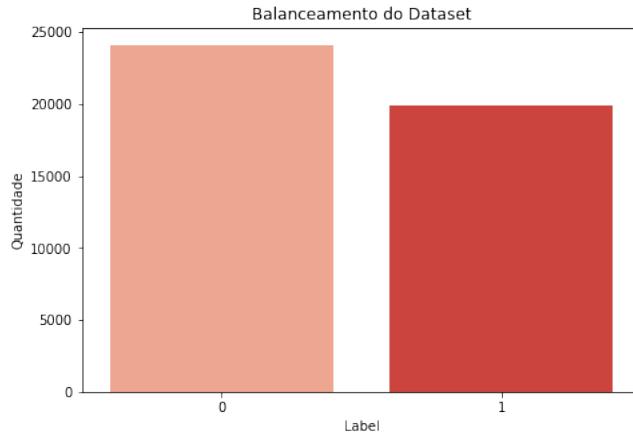


Pode-se notar que:

- Para células positivas, tanto a intensidade, quanto os *pixels* são menores;
- Para células negativas, todos os canais de cores RGB começam a partir de 100 *pixel*.

Ao analisar a Figura 21 nota-se que o *dataset* não está balanceado, ou seja, a porcentagem de dados é maior em uma *label* e menor em outra.

Figura 21 – Balanceamento do *dataset*.



Para realizar a divisão de conjunto de treinamento e teste, de forma automática e randômica, será utilizada a biblioteca *train test split* citada no tópico 3.2.11, para auxiliar. A nível de código, a função responsável por realizar o *split* para dados, em treinamento e teste, funciona da seguinte forma:

Figura 22 – Separação do *dataset*.

```
1 trainDataset, testDataset, trainLabel, testLabel = train_test_split(
2     DATASET['image'], DATASET['label'], test_size=0.2, random_state=15
3 )
```

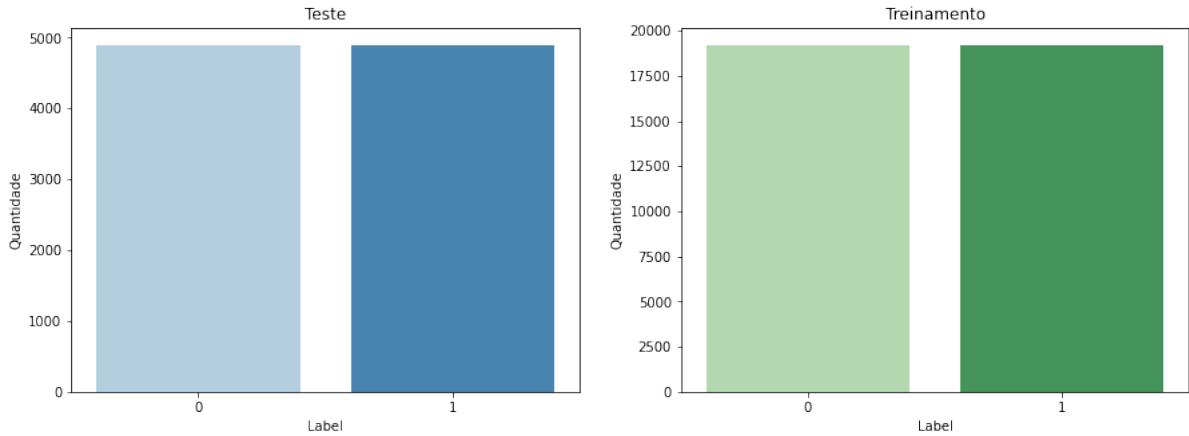
Onde *trainDataset* e *testDataset* são *arrays* e cada posição é uma imagem e, *trainLabel* e *testLabel* é o rótulo que caracteriza essa imagem.

Para realizar o balanceamento dos dados foram testadas duas bibliotecas, sendo:

- [Random Under Sampler \(2014\)](#): Cria novas amostras a partir da classe majoritária;
- [Random Over Sampler \(2014\)](#): Cria novas amostras a partir da classe minoritária.

Ambas escolhem amostras aleatoriamente, para que contenha a mesma quantidade em todos os subconjuntos, dando origem a novos exemplares que seguem o mesmo comportamento, garantindo a normalização, levando o modelo a não enviesar para nenhum dos rótulos e, melhorando a busca por padrões. Após o balanceamento dos dados do *dataset*, tem-se os seguintes resultados, conforme Figura 23:

Figura 23 – Normalização do *dataset*.



4.3 Modelo de Rede Neural

Como abordado no tópico de *CNN 2.5.4* a topologia da rede utilizou a *API* do módulo [TensorFlow \(2020a\)](#), passando os seguintes parâmetros:

Figura 24 – Topologia *CNN*.

```

1 model = tf.keras.Sequential()
2 model.add(layers.Conv2D(32, (3, 3), activation='relu', input_shape=inputShape))
3 model.add(layers.MaxPooling2D((2, 2)))
4 model.add(layers.Dropout(0.2))
5
6 model.add(layers.Conv2D(64, (3, 3), activation='relu'))
7 model.add(layers.MaxPooling2D((2, 2)))
8 model.add(layers.Dropout(0.2))
9
10 model.add(layers.Conv2D(128, (3, 3), activation='relu'))
11 model.add(layers.MaxPooling2D((2, 2)))
12 model.add(layers.Dropout(0.2))
13
14 model.add(layers.Flatten())
15 model.add(layers.Dense(128, activation='relu'))
16 model.add(layers.Dropout(0.2))
17 model.add(layers.Dense(2, activation='softmax'))
```

O modelo possui 3 camadas de convolução, onde a quantidade de neurônios de cada camada varia entre 32, 64 e 128, é utilizado o mesmo *kernel* e todas as funções de ativações são *relu*. Em seguida é realizado o filtro de *pooling*, pegando os maiores valores da convolução, por fim é adicionada a camada de *dropout* para ajudar a evitar *overfitting*.

Após esses passos de convoluções, *pooling* e *dropout*, é realizado o *flatten* para nivelar os dados, sem afetar o tamanho. Em seguida, é adicionado as camadas densas fortemente conectadas, sendo que a última camada densa recebe o valor da quantidade de possibilidades.

No caso deste trabalho, são apenas duas opções.

A compilação do modelo recebe os seguintes parâmetros:

- *optimizer*: É um método de gradiente descendente estocástico, que se baseia na estimativa adaptativa de momentos de primeira e segunda ordem;
- *loss*: Calcula a perda de entropia cruzada, entre os rótulos e as previsões;
- *metrics*: Lista de métricas a serem avaliadas pelo modelo durante o treinamento e teste.

É possível ver a compilação na Figura 25 abaixo:

Figura 25 – Compilação do Modelo.

```
1 model.compile(  
2     optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])  
3 )
```

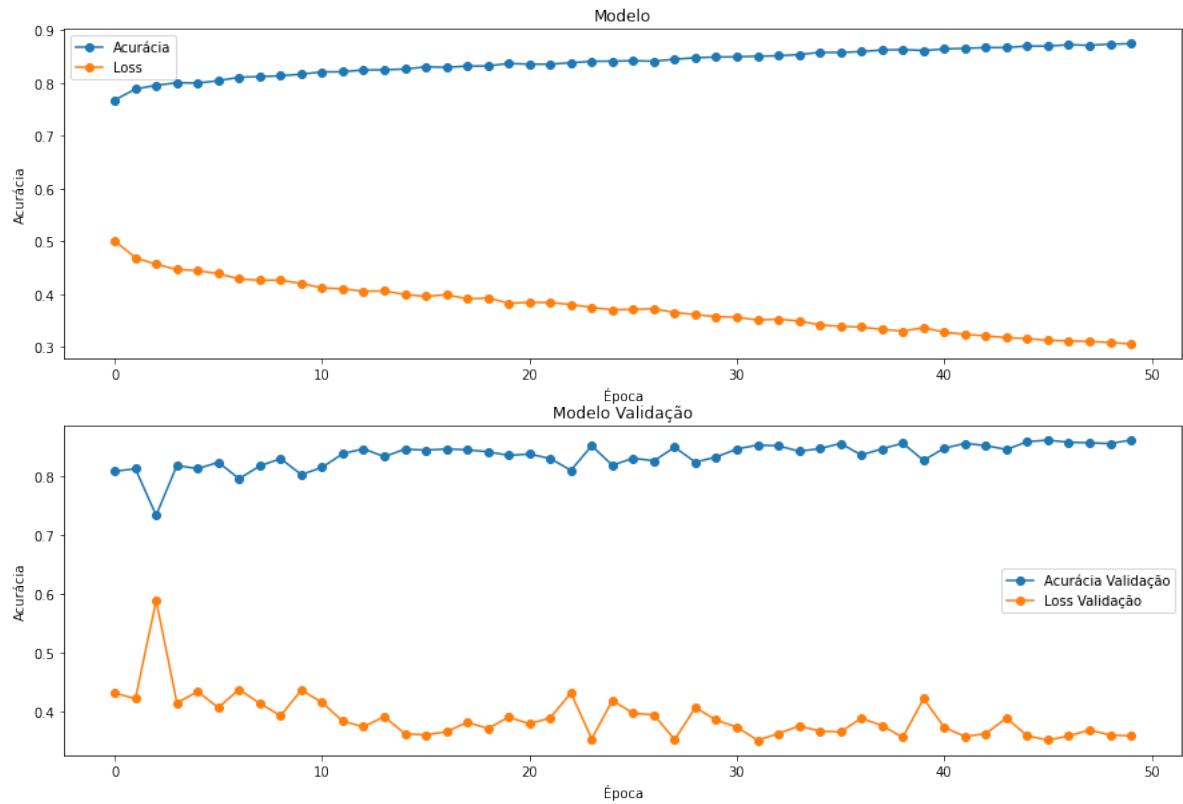
Para o treinamento, é passado o *trainDatasetReshaped* e *trainLabelOver*, capturados durante o processo de balanceamento dos dados, informando a quantidade de épocas para realizar o treinamento e um valor para realizar a validação na época treinada, sendo invocada da seguinte forma:

Figura 26 – Treinamento do Modelo.

```
1 history = model.fit(trainDatasetReshaped, trainLabelOver, epochs=50, validation_split=0.2)
```

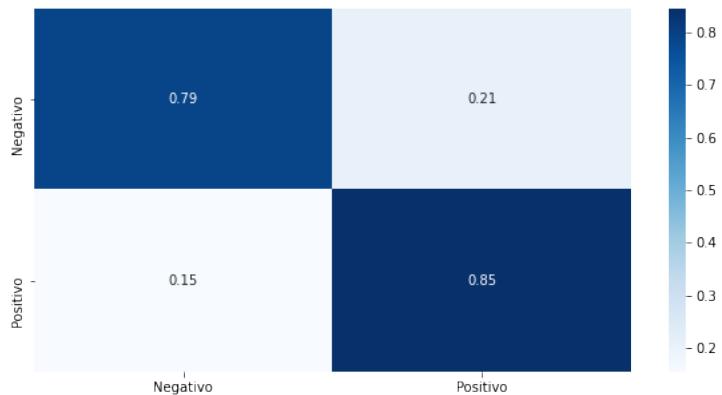
O treinamento retorna os valores de acurácia e *loss*, tanto para treinamento e validação. A Figura 27 representa o valor de acurácia contra o *loss*, onde acurácia é o nível de confiança em que a IA entende que aquela informação prevista está correta, e o *loss* representa a função de custo.

Figura 27 – Acurácia x Loss.



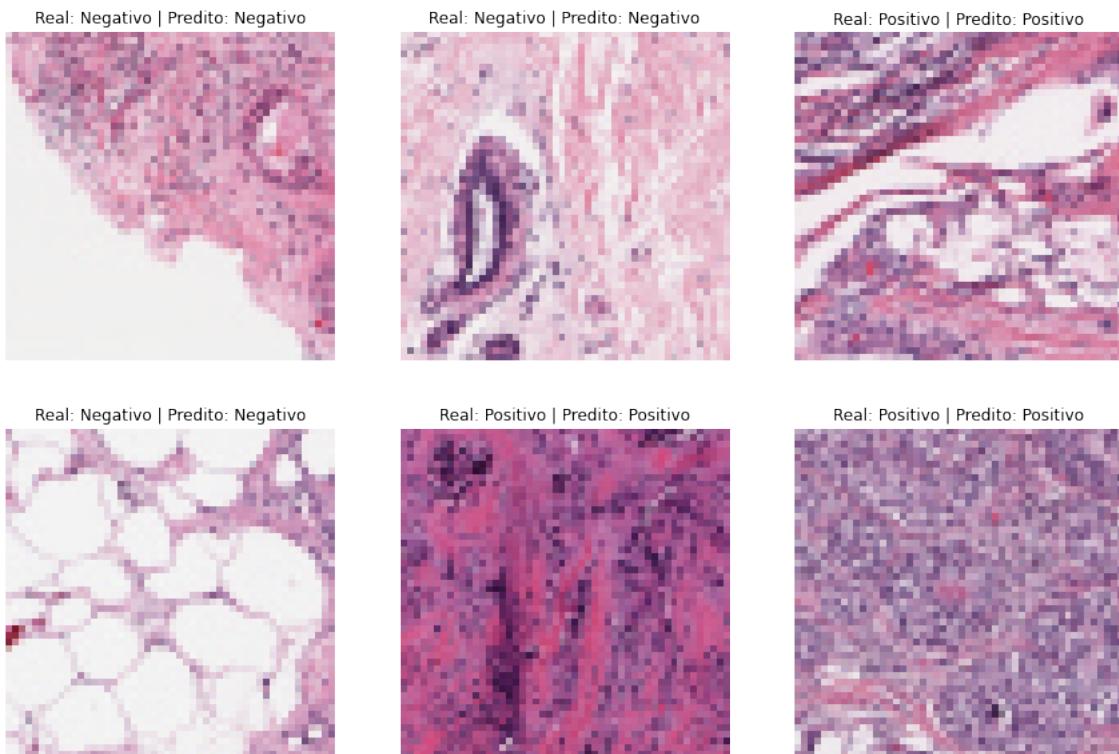
A matriz de confusão a seguir mostra a frequência de classificação para cada classe (*label*) do modelo.

Figura 28 – Matriz de Confusão.



Por fim, a predição se baseia em pegar valores aleatórios do *testDataset* e exibir a imagem referente a esse valor, conforme Figura 29:

Figura 29 – Predizendo Amostra.



O título de cada imagem representa o resultado da predição, onde:

- Real: Valor real da célula;
- Predito: Valor que a IA retornou após a predição.

Conclui-se que, para a primeira imagem selecionada randomicamente, tanto a carcinoma da célula, quanto a carcinoma que foi predita pelo modelo, são rotuladas como 0, ou seja, carcinoma negativo.

5 Conclusão

No decorrer deste trabalho, desenvolveu-se um modelo de inteligencia artificial que conseguiu discernir células saudáveis e células cancerosas a partir de imagens. Tal fato se deu, devido a tecnologia poder e dever auxiliar a área da saúde, facilitando o tratamento precoce e aumentando a qualidade de vida dos pacientes.

Durante o desenvolvimento, foi estudado sobre alguns tipos câncer e seus diversos tipos de tratamento, técnicas de balanceamento de dados, filtros, redimensionamento e estruturação de topologias relacionadas a modelos de IA, atingindo um resultado bastante satisfatório, com acuráncias acima de 0.80, logo nos primeiros treinamentos do modelo.

5.1 Trabalhos Futuros

Como sugestão para trabalhos futuros, considera-se a melhoria dos resultados obtidos neste trabalho. Para tal, pode-se verificar o uso de outros modelos de aprendizado de máquina como por exemplo *U-NET Architecture* e *Efficient Net*, também a utilização das técnicas de *CycleGAN* para realizar a *Data Augmentation*.

Outro ponto a ser melhorado refere-se ao tipo de câncer, ou seja, não deixar apenas o de mama, mas montar uma topologia que realize o pré processamento necessário, que balanceie os dados e consiga, a partir disso classificar qualquer outro tipo e estágio da doença.

Referências

- AMIDI, S. *Convolutional Neural Networks cheatsheet*. 2020. Disponível em: <<https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks>>. Acesso em: 16 Out. 2020.
- ATLAS ONLINE DE MORTALIDADE. *Taxas de mortalidade por câncer, brutas e ajustadas por idade pelas populações mundial e brasileira, por 100.000, segundo sexo, faixa etária, localidade e por período selecionado*. 2020. Disponível em: <<https://mortalidade.inca.gov.br/MortalidadeWeb/pages/Modelo03/consultar.xhtml>>. Acesso em: 18 Mai. 2020.
- BRUDERER, H. The antikythera mechanism. Abril 2020. Disponível em: <<https://dl.acm.org/doi/pdf/10.1145/3368855>>. Acesso em: 19 Mai. 2020.
- CALDEIRA, H. A inteligência artificial aplicada na medicina. Fevereiro 2017. Disponível em: <<https://cmtecnologia.com.br/blog/inteligencia-artificial/>>. Acesso em: 25 Mai. 2020.
- FAYED, L. *How Cancer Was First Discovered and Treated*. 2020. Disponível em: <<https://www.verywellhealth.com/the-history-of-cancer-514101>>. Acesso em: 18 Mai. 2020.
- GNIPPER, P. Mulheres históricas: Ada lovelace, a primeira programadora de todos os tempos. Junho 2016. Disponível em: <<https://canaltech.com.br/curiosidades/mulheres-historicas-ada-lovelace-a-primeira-programadora-de-todos-os-tempos-71395>>. Acesso em: 19 Mai. 2020.
- GOLDSCHMIDT, R. R. *Uma Introdução à Inteligência Computacional: fundamentos, ferramentas e aplicações*. [S.l.: s.n.], 2010. 6-13, 82-91 p.
- GOOGLE COLAB. *Google Colab*. 2020. Disponível em: <<https://colab.research.google.com>>. Acesso em: 20 Mai. 2020.
- GUARIZI, D. D. Estudo da inteligÊncia artificial aplicada na ÁREA da saÚde. Outubro 2014. Disponível em: <<http://www.unoeste.br/site/enepe/2014/suplementos/area/Exactarum/Computa%C3%A7%C3%A3o/ESTUDO%20DA%20INTELIG%C3%8ANCIA%20ARTIFICIAL%20APLIACADA%20NA%20%C3%81REA%20DA%20SA%C3%9ADE.pdf>>. Acesso em: 23 Mai. 2020.
- GÉRON, A. *Neural Networks and Deep Learning*. [s.n.], 2019. 167-170 p. Disponível em: <<http://neuralnetworksanddeeplearning.com/>>.
- HECHT-NIELSEN, R. *Applications of counterpropagation networks*. [S.l.]: Elsevier, 1988. 131-139 p.
- HOSPITAL CÂNCER BARRETOS. *Informação: saiba quais são os tipos de câncer mais comuns no Brasil*. 2015. Disponível em: <<https://www.hcancerbarretos.com.br/82-institucional/noticias-institucional/1300-informacao-saiba-quais-sao-os-tipos-de-cancer-mais-comuns-no-brasil>>. Acesso em: 18 Mai. 2020.

- INCA. *Estatísticas de câncer*. 2020. Disponível em: <<https://www.inca.gov.br/numeros-de-cancer>>. Acesso em: 18 Mai. 2020.
- INCA. *Tratamento para o câncer de mama*. 2020. Disponível em: <<https://www.inca.gov.br/controle-do-cancer-de-mama/acoes-de-controle/tratamento>>. Acesso em: 14 Out. 2020.
- INFOPÉDIA. Herman hollerith. Março 2020. Disponível em: <[https://www.infopedia.pt/\\$herman-hollerith](https://www.infopedia.pt/$herman-hollerith)>. Acesso em: 19 Mai. 2020.
- KAGGLE. *Breast Histopathology Images*. 2020. Disponível em: <<https://www.kaggle.com/paultimothymooney/breast-histopathology-images>>. Acesso em: 14 Out. 2020.
- KERAS. *Keras*. 2020. Disponível em: <<https://keras.io/>>. Acesso em: 15 Out. 2020.
- KHANNA, R.; AWAD, M. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. [S.l.: s.n.], 2015.
- KOHAVI, R. *Glossary of Terms*. 1998. Disponível em: <<http://ai.stanford.edu/~ronnyk/glossary.html>>. Acesso em: 17 Out. 2020.
- LOBO, L. C. Inteligência artificial e medicina. Junho 2017. Disponível em: <https://www.scielo.br/scielo.php?pid=S0100-55022017000200185&script=sci_arttext>. Acesso em: 23 Mai. 2020.
- MARKOFF, J. The mouse inventor's vision of computing. Julho 2013. Disponível em: <<https://bits.blogs.nytimes.com/2013/07/03/the-mouse-inventors-vision-of-computing>>. Acesso em: 19 Mai. 2020.
- MARTINEZ, V. Invasive lobular carcinoma of the breast: A special histological type compared with invasive ductal carcinoma. Novembro 2017. Disponível em: <<https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2559.1979.tb03029.x?sid=nlm%3Apubmed>>. Acesso em: 15 Out. 2020.
- MATPLOTLIB. *Matplotlib*. 2020. Disponível em: <<https://matplotlib.org/>>. Acesso em: 03 Set. 2020.
- MINISTÉRIO DA SAÚDE. Série A. *Normas e Manuais Técnicos Cadernos de Atenção Primária*, n. 29. 2010. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/caderno_atencao_primaria_29_rastreamento.pdf>. Acesso em: 23 Mai. 2020.
- MINISTÉRIO DA SAÚDE. Câncer: sintomas, causas, tipos e tratamentos. 2020. Disponível em: <<https://saude.gov.br/saude-de-a-z/cancer>>. Acesso em: 18 Mai. 2020.
- NIELSEN, M. *Hands-On Machine Learning with Scikit-Learn TensorFlow*. [S.l.: s.n.], 2017. 379 p.
- NUMPY. *NumPy*. 2005. Disponível em: <<https://numpy.org/>>. Acesso em: 15 Out. 2020.
- OPENCV. *OpenCV*. 2020. Disponível em: <<https://opencv.org/>>. Acesso em: 15 Out. 2020.
- PANDAS. *Pandas*. 2008. Disponível em: <<https://pandas.pydata.org/>>. Acesso em: 15 Out. 2020.
- PYTHON. *Python*. 2020. Disponível em: <<https://www.python.org>>. Acesso em: 22 Mai. 2020.

RANDOM OVER SAMPLER. *Random Over Sampler*. 2014. Disponível em: <https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.RandomOverSampler.html>. Acesso em: 17 Out. 2020.

RANDOM UNDER SAMPLER. *Random Under Sampler*. 2014. Disponível em: <https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.under_sampling.RandomUnderSampler.html>. Acesso em: 17 Out. 2020.

SCIKIT LEARN. *Scikit Learn*. 2019. Disponível em: <<https://scikit-learn.org/stable/>>. Acesso em: 14 Out. 2020.

SCIKIT LEARN. *Scikit Learn*. 2020. Disponível em: <<https://scikit-learn.org/>>. Acesso em: 03 Set. 2020.

SEABORN. *Seaborn*. 2020. Disponível em: <<https://seaborn.pydata.org/>>. Acesso em: 13 Set. 2020.

SILVESTRE, G. S. F. P. A. *A história do câncer, enfermidade que ainda desafia a ciência; epidemiologia e o papel do screening no tratamento da doença; a evolução e personalização dos cuidados que permitiram melhora da eficácia e maior segurança; e a inovadora terapia de células CAR-T, que apresenta resultados surpreendentes*. [S.l.: s.n.], 2020.

STACKOVERFLOW. *2020 Developer Survey*. 2020. Disponível em: <<https://insights.stackoverflow.com/survey/2020>>. Acesso em: 13 Set. 2020.

STRAWN, G. Grace hopper: Compilers and cobol. Fevereiro 2015. Disponível em: <<https://ieeexplore.ieee.org/abstract/document/7030179>>. Acesso em: 19 Mai. 2020.

TEIXEIRA, L. A. Políticas públicas de controle de câncer no brasil: elementos de uma trajetória. 2012. Disponível em: <https://www.researchgate.net/profile/Marco_Porto2/publication/274953427_Public_policies_for_cancer_control_in_Brazil_elements_of_a_trajectory/links/552cf4b60cf2e089a3ad0e57/Public-policies-for-cancer-control-in-Brazil-elements-of-a-trajectory.pdf>. Acesso em: 22 Mai. 2020.

TENSORFLOW. *Module: tf.keras*. 2020. Disponível em: <https://www.tensorflow.org/api_docs/python/tf/keras>. Acesso em: 15 Out. 2020.

TENSORFLOW. *TensorFlow*. 2020. Disponível em: <<https://www.tensorflow.org/>>. Acesso em: 03 Set. 2020.

TOTVS. *Os benefícios da inteligência artificial na saúde*. 2020. Disponível em: <<https://www.totvs.com/blog/instituicoes-de-saude/inteligencia-artificial-na-saude/>>. Acesso em: 23 Mai. 2020.

TRAIN TEST SPLIT. *Train Test Split*. 2020. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>. Acesso em: 16 Out. 2020.