



UNIFOR

**UNIVERSIDADE DE FORTALEZA
CENTRO DE CIÊNCIAS TECNOLÓGICAS
CURSO DE GRADUAÇÃO EM ENGENHARIA DE CONTROLE E AUTOMAÇÃO**

PEDRO LINO AZEVÊDO LANDIM

**DIAGNÓSTICO DE LIPODISTROFIA GENERALIZADA CONGENITA ATRAVÉS
DE REDE NEURAL CONVOLUCIONAL**

**FORTALEZA – CEARÁ
2021**

PEDRO LINO AZEVÊDO LANDIM

**DIAGNÓSTICO DE LIPODISTROFIA GENERALIZADA CONGENITA ATRAVÉS DE
REDE NEURAL CONVOLUCIONAL**

Trabalho de Conclusão de Curso apresentado
ao Curso de Graduação em Engenharia de
Controle e Automação do Centro de Ciências
Tecnológicas da Universidade de Fortaleza,
como requisito parcial à obtenção do grau
de bacharel em Engenharia de Controle e
Automação.

Orientador: Joel Sotero da Cunha Neto,
Msc

FORTALEZA – CEARÁ

2021

Ficha catalográfica da obra elaborada pelo autor através do programa de geração automática da Biblioteca Central da Universidade de Fortaleza

Landim, Pedro Lino Azevedo .

DIAGNÓSTICO DE LIPODISTROFIA GENERALIZADA Congênita ATRAVÉS
DE REDE NEURAL CONVOLUCIONAL / Pedro Lino Azevedo Landim. -
2021

48 f.

Trabalho de Conclusão de Curso (Graduação) - Universidade
de Fortaleza. Curso de Eng De Controle E Automação, Fortaleza,
2021.

Orientação: Joel Sotero da Cunha Neto.

Coorientação: Paulo Cirillo Souza Barbosa.

1. DIAGNÓSTICO DE LIPODISTROFIA GENERALIZADA Congênita
ATRAVÉS . 2. Rede Neural Convolucionar. 3. Aprendizado de
máquina. I. Neto, Joel Sotero da Cunha . II. Barbosa, Paulo
Cirillo Souza . III. Título.

**DIAGNÓSTICO DE LIPODISTROFIA GENERALIZADA CONGÊNITA
ATRAVÉS DE REDE NEURAL CONVOLUCIONAL**

PEDRO LINO AZEVEDO LANDIM

PARECER: APROVADO

Data: 22 / 01 / 2021

BANCA EXAMINADORA:

Natalia Bitar da Cunha Olegario

NATALIA BITAR DA CUNHA OLEGARIO, Ma.

André Lunardi de Souza

ANDRÉ LUNARDI DE SOUZA, Me.

Paulo Cirillo

PAULO CIRILLO SOUZA BARBOSA
Coorientador(a)

Joel Sotero da Cunha Neto

JOEL SOTERO DA CUNHA NETO, Me.
Orientador(a)

Lucia Maria Barbosa Oliveira

Profª. Lucia Maria Barbosa Oliveira, Ma.
Coordenadora do Curso de Eng. de Controle e Automação

Este trabalho é dedicado aos meus pais, e a todos
que me ajudaram nesta jornada.

AGRADECIMENTOS

Aos meus pais e família, que sempre me apoiaram e incentivaram em momentos difíceis. Agradeço a meu orientador Joel Sotero da Cunha Neto e a todos os professores pelo o ensinamento e correções que me ajudaram a ter o melhor desempenho possível. Agradeço a meu amigo Paulo Cirillo pelo o ensinamento e correções que me permitiram concluir este trabalho. Agradeço a todos que contribuíram e que estiveram ao meu lado durante essa jornada de graduação.

“O fracasso é uma oportunidade de começar de novo com mais inteligência.”

(Henry Ford)

RESUMO

A dificuldade de fazer o diagnóstico de doenças raras é elevada, muitas vezes os pacientes têm um diagnóstico errado e recebem um tratamento que não é efetivo para a sua real condição. Um exemplo deste tipo de doença, é a Lipodistrofia Generalizada Congênita Rede Neural Convolucional (CNN) que acomete 1 em cada 10 milhões de nascidos vivos. Com os avanços tecnológicos, a utilização de algoritmos de aprendizagem de máquina vem se intensificando e são ferramentas promissoras para resolver problemas em diversas áreas, inclusive na medicina. Neste contexto, este trabalho apresenta um modelo de aprendizado de máquina baseado em Rede Neural Convolucional (CNN, do inglês *Convolutional Neural Network*) que tem como objetivo realizar a classificação de três tipos de pacientes, os que possuem Lipodistrofia Generalizada Congênita (LGC), os que são desnutridos e os que possuem boa nutrição. Este processo foi realizado utilizando técnica de *data augmentation*, para que a CNN fosse capaz de utilizar a maior quantidade de imagens possível. O projeto foi desenvolvido em um hardware limitado sem Unidade de Processamento Gráfico (GPU) e com apenas 4 Gb de Memória de Acesso Aleatório (RAM), o que acabou ocasionando em uma queda de desempenho nos resultados obtidos. A despadronização do banco de dado foi outro fator que levou a uma queda no desempenho dos modelos propostos, as imagens que compõem o banco de dados são tiradas em situações diferentes uma das outras, o que acaba aumentando a complexidade da classificação. Nos testes iniciais, foi possível verificar que o algoritmo obteve uma acurácia de 63% e após a identificação dos melhores parâmetros, aplicou-se a validação cruzada de quatro dobras para avaliar a taxa de acerto média ao se utilizar a CNN. Ao fim, verificou-se que a taxa de acerto média na validação foi de 62,9%.

Palavras-chave: Lipodistrofia Generalizada Congênita LGC. Rede Neural Convolucional CNN. aprendizado de máquina.

ABSTRACT

The difficulty in diagnosing rare diseases is quite high, patients are often misdiagnosed and receive treatment that is not effective for their real condition. An example of this type of disease is Congenital Generalized Lipodystrophy CNN, which affects 1 in 10 million live births. With technological advances, the use of machine learning algorithms has been intensifying and are promising tools for solving problems in several areas, including medicine. In this context, this work presents a machine learning model based on Convolutional Neural Network (CNN, from the English *Convolutional Neural Network*) that aims to classify three types of patients, those who have LGC, those who are malnourished and those with good nutrition. This process was carried out using the *data augmentation* technique, so that CNN was able to use as many images as possible. The project was developed on a limited hardware without GPU and with only 4 Gb of RAM, which ended up causing a drop in performance in the results obtained. The standardization of the database was another factor that led to a drop in the performance of the proposed models, the photos that make up the database are taken in different situations from each other, which ends up increasing the complexity of the classification. In the initial tests, it was possible to verify that the algorithm obtained an accuracy of 63 % and after the identification of the best parameters, four-fold cross validation was applied to assess the average hit rate when using CNN. In the end, it was found that the average hit rate in the validation was 62.9 %.

Keywords: Congenital Generalized Lipodystrophy LGC. Convolutional Neural Network CNN. Machine learning.

LISTA DE ILUSTRAÇÕES

Figura 1 – Herança genética de característica recessiva.	13
Figura 2 – Pacientes com LGC	15
Figura 3 – Áreas da inteligência artificial	16
Figura 4 – Modelo de um neurônio artificial.	17
Figura 5 – Esquemático de uma rede neural e de uma rede neural profunda.	17
Figura 6 – Aplicação do filtro na imagem.	18
Figura 7 – Agrupamento e máximo e medio.	19
Figura 8 – Representação da descida do gradiente.	20
Figura 9 – Taxa de aprendizado.	21
Figura 10 – Aplicação da função de ativação ReLu.	22
Figura 11 – Exemplo da técnica de <i>data augmentation</i>.	23
Figura 12 – Tela do <i>jupyter notebook</i>.	25
Figura 13 – Validação cruzada de 4 dobras.	26
Figura 14 – Grupos do banco de dado.	31
Figura 15 – Aplicação da técnica <i>data augmentation</i>.	32
Figura 16 – Fluxograma do código de tratamento das imagens.	33
Figura 17 – Fluxograma da CNN.	35
Figura 18 – Acurácia relação ao número de épocas.	37
Figura 19 – Divisão para a validação cruzada de 4 dobras.	38
Figura 20 – Erro durante o processo de treinamento.	38
Figura 21 – Imagens classificadas pelo modelo.	39
Figura 22 – Matriz de confusão.	39
Figura 23 – Matriz de confusão com dois grupos.	40

LISTA DE ABREVIATURAS E SIGLAS

ANN	<i>Artificial Neural Network</i>
CNN	Rede Neural Convolucional
CPU	Unidade Central de Processamento
DH	Doença de Huntington
GPU	Unidade de Processamento Gráfico
LGC	Lipodistrofia Generalizada Congênita
LS-SVM	<i>Least Squares Support Vector Machine</i>
MSE	Erro Quadrático Médio
RAM	Memória de Acesso Aleatório
RAS	Rede de Atenção à Saúde
ReLU	Unidades Lineares Retificadas
RF	<i>Random Forest</i>
RNA	Rede Neural Artificial
SBS	Síndrome de Berardinelli-Seip
SD	Síndrome de Down
SVM	<i>Support Vector Machine</i>
UFC	Universidade Federal do Ceará

SUMÁRIO

1	INTRODUÇÃO	10
1.1	JUSTIFICATIVA	10
1.2	OBJETIVO	11
1.3	OBJETIVOS ESPECÍFICOS	11
2	FUNDAMENTAÇÃO TEÓRICA	12
2.1	LIPODISTROFIA GENERALIZADA CONGÊNITA	12
2.2	INTELIGÊNCIA ARTIFICIAL	15
2.2.1	Aprendizado de máquina	15
2.2.2	Rede neural artificial	16
2.3	REDE NEURAL CONVOLUCIONAL (CNN)	17
2.3.1	Convolução	18
2.3.2	Mapa de características	18
2.3.3	Camadas de agrupamento	19
2.3.4	<i>Backpropagation</i>	19
2.3.5	Gradiente Descendente	20
2.3.6	Taxa de aprendizado	21
2.3.7	Função de ativação	21
2.3.8	<i>Overfitting</i>	22
2.3.9	<i>Data Augmentation</i>	22
2.4	PYTHON PARA O APRENDIZADO DE MÁQUINA	23
2.5	<i>TENSORFLOW</i>	24
2.5.1	Calculo do erro	24
2.6	<i>JUPYTER NOTEBOOK</i>	24
2.7	VALIDAÇÃO CRUZADA	25
3	TRABALHOS RELACIONADOS	27
4	METODOLOGIA	30
4.1	<i>HARDWARE E SOFTWARE</i>	30
4.2	MONTAGEM DO BANCO DE DADOS	30
4.2.1	Ampliação do banco de dados	31
4.2.2	Tratamento das imagens	32
4.3	REDE NEURAL CONVOLUCIONAL CNN	33

5	RESULTADOS	36
6	CONCLUSÃO	41
	REFERÊNCIAS	43

1 INTRODUÇÃO

Pessoas com doenças raras enfrentam diversos problemas para conseguir o diagnóstico de sua doença. Muitas vezes a doença do paciente acaba passando por complicações. As doenças raras são comumente confundidas com outras doenças o que acabam prejudicando na recuperação do paciente. Enquanto o paciente deveria estar recebendo o tratamento para uma determinada doença, acaba recebendo o tratamento para outra doença, com sintomas semelhantes e que ele não possui(CEDARO *et al.*, 2020).

A LGC é uma doença genética rara de difícil diagnóstico, assim como outras doenças raras, já que acomete 1:10.000.000 de nascidos vivos e possui muitos sintomas semelhantes a outras doenças como desnutrição e diabetes. Apesar da LGC ser uma doença de baixa prevalência, entre 300 a 500 pessoas no mundo já foram registrados com essa doença (PATNI; GARG, 2015).

A medicina tem encontrado no aprendizado computacional a solução para conseguir fazer o diagnóstico de doenças que de alguma forma apresenta complicações em seu diagnóstico, como exames invasivos ou doenças pouco conhecidas pelos médicos. Wang, Teoh e Choi (2018) desenvolveram um modelo preditivo para detectar câncer de próstata visando minimizar o número de biópsias desnecessárias. Li *et al.* (2019) proporam um procedimento não invasivo para diagnóstico de síndrome de down, visando minimizar o custo financeiro e social do diagnóstico no pré-natal. Ongsuk *et al.* (2018) apresenta um sistema de prognóstico para colangiocarcinoma, uma categoria de câncer de fígado. G. WANG, L. LI e S. ONGSUK utilizam aprendizado de máquina como principal ferramenta para o diagnóstico ou prognóstico de doenças.

1.1 JUSTIFICATIVA

Os trabalhos de Wang, Teoh e Choi (2018) e Ongsuk *et al.* (2018) na medicina tem mostrado a eficiência de modelos de aprendizado de máquina para auxiliar os médicos nos diagnósticos e na redução dos custos dos exames, facilitando e popularizando o acesso a saúde.

Implementar um modelo de aprendizado de máquina para a identificação fenotípica de uma doença rara como LGC é extremamente complicado já que o números de casos registrados são bem escassos. Devido a questões éticas e de privacidade é complicado criar grandes bancos de dados com os exames dos paciente. Uma solução para a falta de dados para fazer uma rede eficiente é a utilização da técnica *data augmentation*. A utilização da técnica *data augmentation* tem ajudado a desenvolver projetos de aprendizado de máquina para fazer o diagnóstico de doenças gastrointestinais. A técnica *data augmentation* no diagnóstico de doenças

gastrointestinais teve melhorias sensíveis da classificação em relação as outras abordagens de classificação (ASPERTI; MASTRONARDO, 2017).

Li *et al.* (2019) apresentam como é extremamente importante fazer um diagnóstico preciso e precoce para o tratamento da doença evitando possíveis complicações medicas, reduzindo os custos financeiros e aumentando a qualidade de vida do paciente que passa a ter um tratamento específico para doença melhorando assim sua recuperação.

Diversos trabalhos falam dos problemas do diagnóstico tardio como Rangel, Lima e Vargas (2015), mostram que a forma mais eficaz para o tratamento de câncer é por meio do diagnóstico e tratamento das lesões precursoras e tumorais invasores no seu estágio inicial, com isso evitando futuras complicações. Stuani *et al.* (1999) mostram os problemas do diagnóstico tardio e a importância da localização radiográfica precoce, ou seja, o diagnóstico precoce prevenindo complicações no quadro do paciente.

1.2 OBJETIVO

O trabalho tem como objetivo desenvolver um algoritmo de aprendizado de máquina para a identificação fenotípica de LGC , utilizando a linguagem python e bibliotecas próprias para o aprendizado de máquina como o *Tensorflow* .

1.3 OBJETIVOS ESPECÍFICOS

- a) Desenvolver uma rede neural convolucional para a identificação fenotípica de LGC.
- b) Analisar os parâmetros para a classificação de pessoas com LGC, anorexia e pessoas com boa nutrição por meio de aprendizado de máquina.
- c) Analisar os resultados obtido por meio da validação cruzada de 4 dobras.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os tópicos importantes para o desenvolvimento do projeto e o melhor entendimento do seu funcionamento. Serão apresentadas as características essenciais da LGC, inteligência artificial, *python* , *Tensorflow* e *jupyter notebook*.

2.1 LIPODISTROFIA GENERALIZADA CONGÊNITA

Berardinelli (1954) analisou dois pacientes brasileiros em 1954, posteriormente Seip (1959) analisou três pacientes com as mesmas características dos pacientes analisados por Berardinelli (1954). Com isso deu-se o nome da doença de Síndrome de Berardinelli-Seip (SBS) que também é conhecida como LGC. Berardinelli (1954) descreve a LGC como uma condição rara, de herança autossômica recessiva, ou seja, uma doença hereditária que afeta os genes recessivos. Na figura 1 pode-se observar como acontece a Herança genética do gene recessivo. Analisando o funcionamento da herança de genes recessivos pode-se observar que dois quartos dos filhos carrega o gene para uma nova geração, mas não possui a doença, um quarto possui a doença e passa para uma nova geração e apenas um quarto não possui a doença (BONATTO; FELTES, 2017).

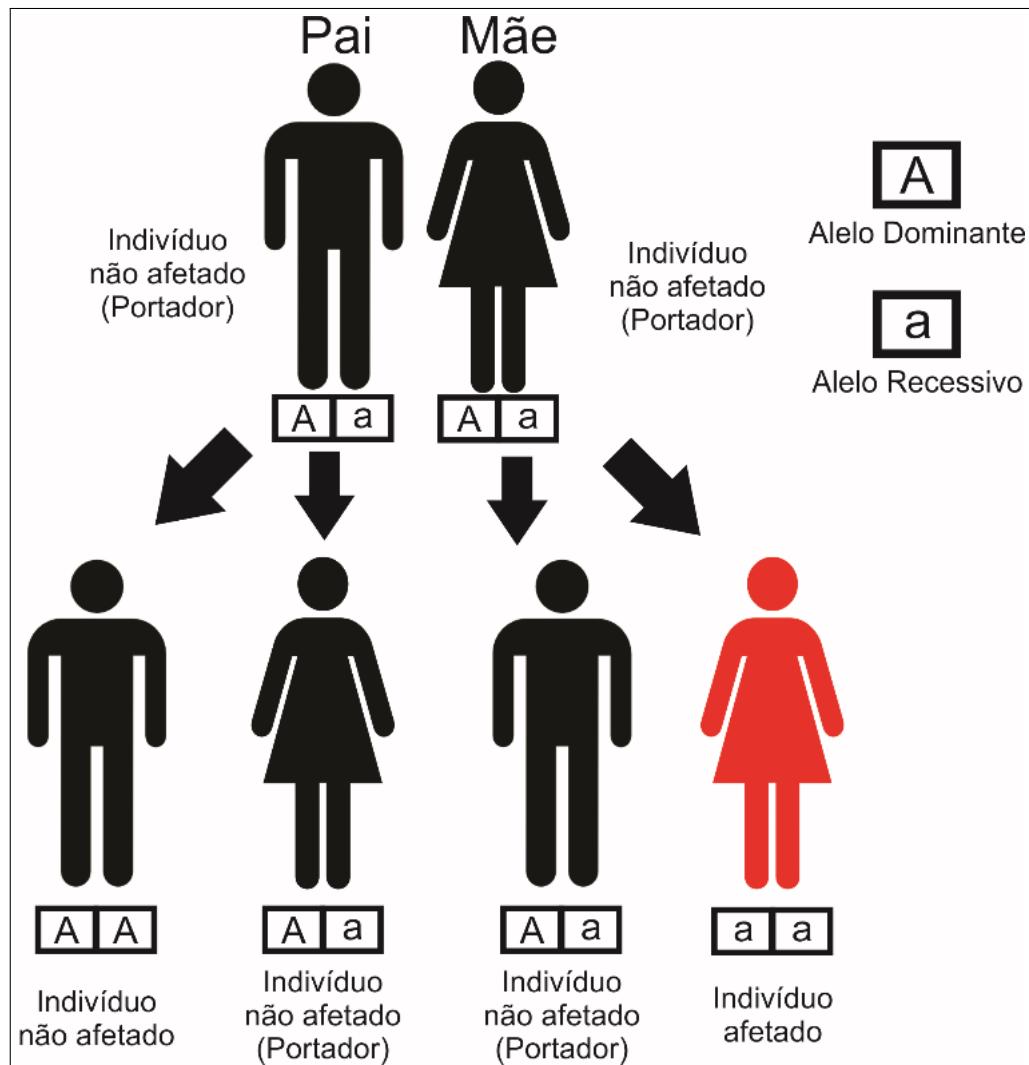
Magré *et al.* (2001) e Nolis (2014) fazem uma análise interessante, de que existe uma grande concentração de casos em países como Brasil, Portugal e Líbano ou países que tenham pessoas com ancestrais africanos. No Brasil a doença é mais comum no Nordeste (FARIA *et al.*, 2009). O elevado número de casos no Nordeste pode ser justificado pelo número de casamentos consanguíneos, que é basicamente o casamento entre parentes próximos como irmãos ou primos (MALDERGEM *et al.*, 2002).

Patni e Garg (2015) apresenta os quatro subtipos clínico-moleculares de LGC, cada um apresentando características clínicas e genótipos únicos. Pode-se observar no quadro 1 as características dos quatro tipos de LGC .

Analizando as características dos quatro tipos de LGC apresentadas por PATNI em (PATNI e GARG, 2015), pode-se fazer uma relação das características que os pacientes possuem, são elas:

- a) Hipertrofia muscular das extremidades;
- b) Veias subcutâneas superficiais proeminentes;
- c) Crescimento acelerado;
- d) Apetite voraz;

Figura 1 – Herança genética de característica recessiva.



Fonte – (BONATTO; FELTES, 2017)

- e) Proeminência do umbigo ou hérnias umbilicais;
- f) Hepatomegalia e / ou esplenomegalia;
- g) Acanthosis nigricans;
- h) Hirsutismo leve e clitoromegalia em pacientes do sexo feminino;
- i) Períodos menstruais irregulares com ovários policísticos;
- j) Idade óssea avançada;

Todos os pacientes que possuem algum tipo de LGC apresenta ausência ou redução do tecido adiposo, o que acaba prejudicando as funções metabólicas causando as manifestações clínicas. Com a capacidade de sintetizar e secretar hormônios prejudicada pela falta ou redução do tecido adiposo prejudicam diretamente na homeostase corporal. A homeostase corporal é basicamente a capacidade do organismo de ficar em equilíbrio para realizar as funções que são

Quadro 1 – Tipos de LGC.

Complicação	LGC tipo 1	LGC tipo 2	LGC tipo 3	LGC tipo 4
Tecido adiposo	Ausência de tecido adiposo metabolicamente ativo	Ausência do tecido adiposo metabolicamente ativo e do tecido adiposo mecânico	Ausência de tecido adiposo metabolicamente ativo	Ausência de tecido adiposo metabolicamente ativo
	Preservação do tecido adiposo mecânico		Preservação do tecido adiposo mecânico e da medula óssea	Preservação do tecido adiposo mecânico e da medula óssea
Complicações cardiovasculares	Sem relatos	Cardiomiotia	Sem relatos	Cardiomiotia
				Arritmias ventriculares, QT longo, Morte súbita
Ossos articulações e marcha	Lesões líticas focais em ossos longos após puberdade	Marcha espástica	Baixa estatura	Osteopenia, deformação metáfise distal, rigidez articular e instabilidade atlanto-axial
Complicações gastrointestinais	Sem relatos	Sem relatos	Megaesôfago funcional	Estenose pilórica congênita
Musculatura esquelética	Sem relatos	Sem relatos	Sem relatos	Miopatia congênita
Outras características	Características acromegalóide com aumento da mandíbula, mãos e pés	Teratozoospermia	Hipocalcemia secundária à resistência à vitamina D	Início tardia da lipodistrofia na infância

Fonte – (PATNI; GARG, 2015), modificado pelo autor.

fundamentais para o organismo de forma perfeita (GARG, 2006).

Como as manifestações clínicas da LGC surgem bem cedo nos primeiros anos de vida é importante saber os critérios para o diagnóstico, Patni e Garg (2015) apresentam alguns critérios para se fazer o diagnóstico da síndrome como a ausência do tecido adiposo subcutâneo e hipertrofia muscular, características que podem aparecer no nascimento ou um tempo depois. Na figura 2 pode-se observar as características fenotípicas clínicas visíveis da LGC como são descritas no quadro 1.

Figura 2 – Pacientes com LGC



Fonte – (BERARDINELLI, 1954)

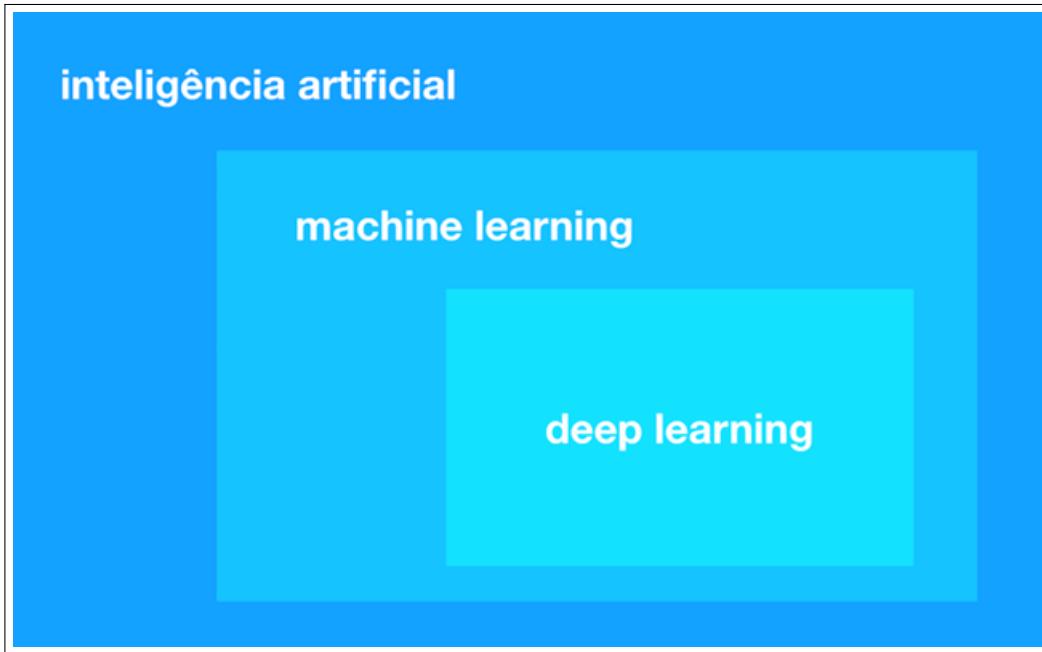
2.2 INTELIGÊNCIA ARTIFICIAL

Inteligência artificial é um termo utilizado para abranger diversas áreas de conhecimento de algoritmos que têm a capacidade de organizar dados, reconhecer padrões e fazer com que computadores possam tomar decisões inteligentes sem a necessidade de pré-programação. Dentro da inteligência artificial tem-se outras duas grandes áreas: o aprendizado de máquina *machine learning* e o aprendizado profundo *deep learning* (SALESFORCE, 2018). Estas informações podem ser ilustradas através da figura 3 , na qual é possível verificar que a inteligência artificial engloba a aprendizagem de máquina e que engloba sequencialmente o aprendizado profundo.

2.2.1 Aprendizado de máquina

Segundo Hurwitz e Kirsch (2018) o aprendizado de máquina, ou *Machine Learning*, é um modelo de algoritmo, programado com base em modelos matemáticos, que consiste em fazer com que os computadores detectam padrões com isso possam previsões que resolvem o problema. O aprendizado de máquina faz com que o computador consiga identificar padrões, se adaptando com o problema de acordo com o banco de dados que é fornecido para fazer o aprendizado de máquina. Depois que o computador aprende o padrão ele consegue classificar, organizar os dados e permite que o computador faça uma análise geral do problema, conseguindo tomar uma decisão inteligente sem a necessidade de pré-programação. Existem diversas formas para se aplicar o aprendizado de máquina como aprendizado supervisionado, aprendizado não supervisionado, aprendizado por reforço e aprendizado profundo.

Figura 3 – Áreas da inteligência artificial



Fonte – (SALESFORCE, 2018)

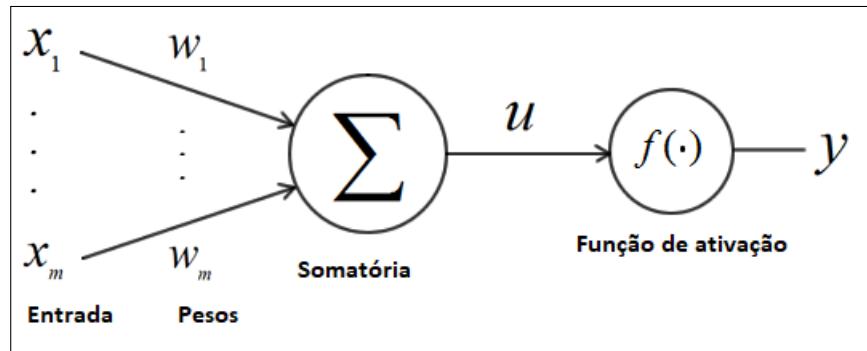
2.2.2 Rede neural artificial

Rede Neural Artificial (RNA) consiste em um modelo de aprendizado de máquina inspirado em um neurônio. A rede neural simula o funcionamento do cérebro humano. O perceptron é um tipo de RNA que representa um neurônio artificial. Na figura 4 tem-se a ilustração gráfica do *perceptron*, ele recebe os valores X_i de entrada e multiplica pelos pesos W_i e faz a somatória de todas as multiplicações da entrada em seguida é multiplicado o resultado da somatória pela função de ativação $f(.)$. Pode-se observar a equação do perceptron na equação 2.1 (MARTINIANO *et al.*, 2016).

$$y = \sum_{i=1} (X_i \cdot W_i) \cdot f(.) \quad (2.1)$$

Uma rede neural simples é composta por três camadas de neurônios artificiais interligados que processam o dado e retornam um resultado, sendo o resultado uma ação para um processo ou uma informação com base na análise dos dados fornecidos. A primeira camada é a camada de entrada na qual recebe os dados. Por exemplo em um classificador de doenças, os dados fornecidos são os sintomas e as informações do paciente como idade, sexo e peso. A segunda camada é a camada oculta, na qual é feita o processamento dos dados, os pesos são atribuídos aos dados e os neurônios artificiais executam as equações de cada nó, que é conhecido

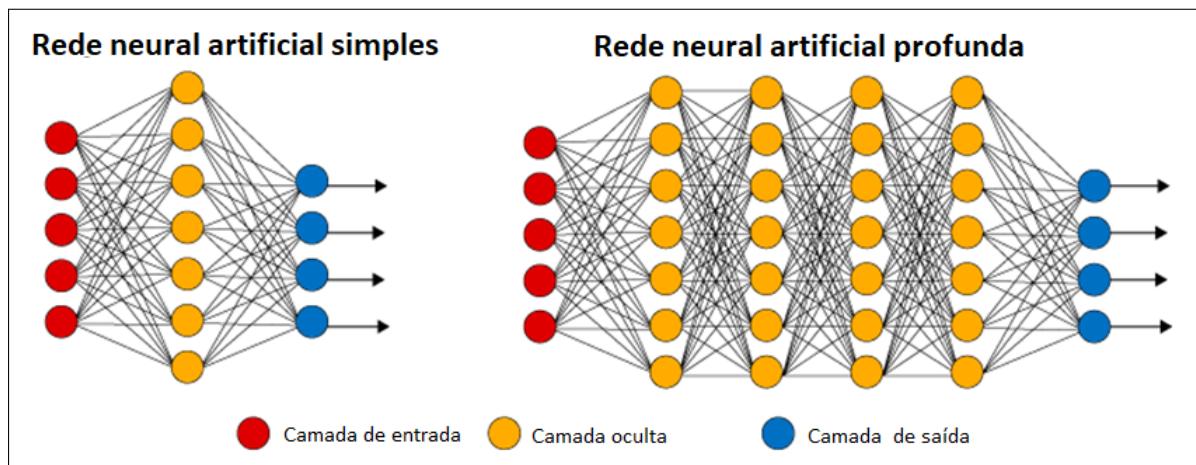
Figura 4 – Modelo de um neurônio artificial.



Fonte – (PACHECO, 2015), modificado pelo autor.

também como neurônio, para definir se os dados serão propagados ou não. A terceira camada é a camada de saída na qual o resultado é fornecido, no exemplo do classificador de doenças, seria o diagnóstico qual doença o paciente possui com base nos sintomas e informações. A diferença entre uma rede neural simples e uma rede neural profunda está na quantidade de camadas ocultas que dentro da rede neural profunda é muito maior que a rede neural simples. Na figura 5 tem-se um exemplo de uma rede neural simples e uma rede neural profunda onde é possível observar a camada de entrada, camada oculta e a camada da saída.

Figura 5 – Esquemático de uma rede neural e de uma rede neural profunda.



Fonte – (SOARES, 2019), modificado pelo autor.

2.3 REDE NEURAL CONVOLUCIONAL (CNN)

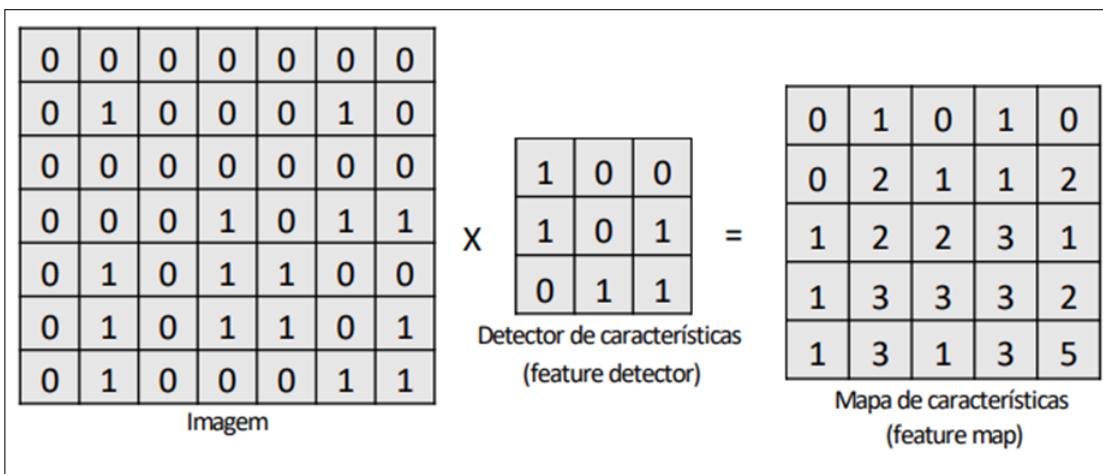
Rosa *et al.* (2018) descrevem a CNN como uma classe de rede neural artificial extremamente eficiente no processamento e análise de imagens. Vogado *et al.* (2019) aborda o conceito da rede neural convolucional que foi criada pensando no poder das redes neurais

com várias camadas que conseguem realizar abstrações extremamente precisas e com isso transferir para outros problemas, podendo construir um enorme conjunto de dados através de transformações lineares e não lineares. Por isso, as redes neurais convolucionais têm um grande percentual de acerto em suas classificações.

2.3.1 Convolução

A camada de convolução é a camada da rede neural em que é aplicado um filtro. O filtro ou *kernel* percorre toda a imagem, tendo como resultado apenas as características mais importantes da imagem. É importante constatar que as imagens e filtros são representados computacionalmente como matrizes. Na figura 6 tem-se a representação de um filtro que percorre uma matriz que representa a imagem, resultando em uma matriz com as informações mais importantes ressaltadas. O filtro é basicamente uma matriz que percorre toda a matriz que representa a imagem, tendo como resultado uma matriz com os detalhes da imagem mais ressaltadas. Como o filtro percorre toda a imagem para realçar os seus detalhes, a matriz do filtro tem a dimensão menor que a matriz da imagem analisada (MARIM, 2019).

Figura 6 – Aplicação do filtro na imagem.



Fonte – (GRANATY, 2019)

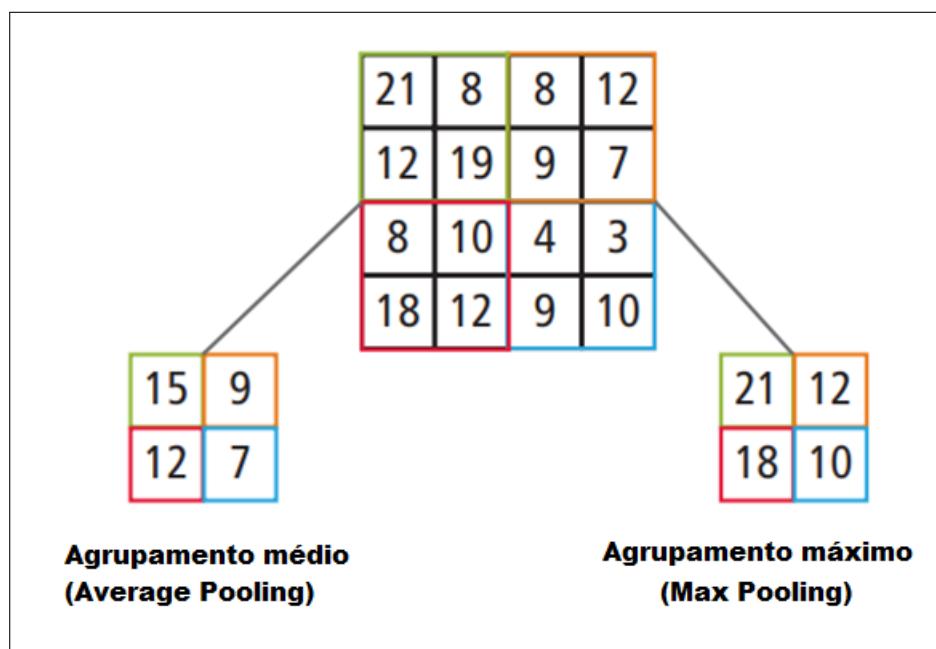
2.3.2 Mapa de características

O mapa de características é a matriz que foi criada após a aplicação do filtro na imagem durante a camada de convolução. Basicamente o mapa de características é uma matriz com os dados da imagem mais detalhados (ALVES, 2018).

2.3.3 Camadas de agrupamento

A camada de agrupamento tem como objetivo reduzir a resolução da imagem para tornar os dados mais robustos contra as distorções e ruídos, sendo possível construir essa camada de duas formas. Uma das formas é pelo agrupamento máximo e outra pelo agrupamento médio. Pode-se observar a representação desses dois modelos na figura 7, em que as duas formas têm o mesmo objetivo, tornar os dados mais fortes contra ruídos e distorções (HIJAZI; KUMAR; ROWEN, 2015).

Figura 7 – Agrupamento e máximo e medio.



Fonte – (HIJAZI; KUMAR; ROWEN, 2015), modificado pelo autor.

As camadas de agrupamento, também conhecidas como camadas de *pooling* ou *pooling layers*, fazem a combinação das saídas de grupos de neurônios em um único neurônio na próxima camada. Com isso, cada camada de *pooling* reduz a dimensão da matriz, gerando uma nova matriz com dimensão reduzida. Na figura 7 pode-se observar que uma matriz de 4x4 passando pela camada de *pooling* resulta em uma matriz de 2x2 (HIJAZI; KUMAR; ROWEN, 2015).

2.3.4 Backpropagation

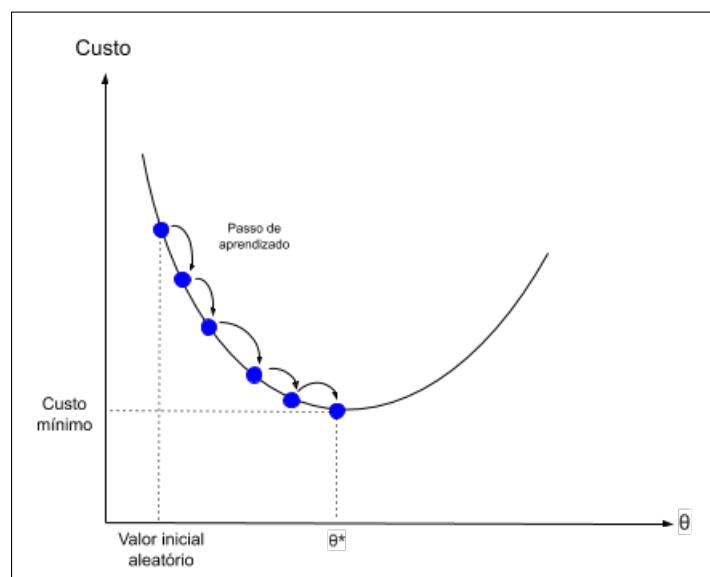
A retropropagação *backpropagation* é um algoritmo de fundamental importância para as redes neurais, sendo o algoritmo que possibilita fazer o aprendizado da rede e com isso o

aperfeiçoamento. A retropropagação ou *backpropagation* tem o objetivo de atualizar os pesos dos neurônios, que normalmente são inicializados com valores aleatórios, para a rede neural conseguir fazer previsões mais próximas do resultado esperado. Pode-se separar o algoritmo de retropropagação *backpropagation* em duas partes o passo para frente *forward pass* ou fase de propagação e o passo para trás *backward pass* (ACADEMY, 2020a). Analisando o processo de retropropagação temos as etapas que consistem em primeiro fazer a inicialização dos pesos dos neurônios com valores aleatórios, em seguida fornece um dado para entrada com isso calculamos o erro comparando o resultado da rede com o valor real do dado que fornecemos como entrada. Com erro calculado percorremos toda a rede neural, do neurônio da saída para os neurônios de entrada, atualizando os valores dos pesos de acordo com a taxa de aprendizado de modo a minimizar o erro (MOURA, 2019).

2.3.5 Gradiente Descendente

O gradiente descendente é um método de otimização que busca encontrar o mínimo de uma função. O gradiente descendente é utilizado em redes neurais no *backpropagation* para encontrar o menor erro possível, ou seja, encontrar os melhores parâmetros do modelo (HAYKIN, 2007). Na figura 8 tem-se a representação gráfica de uma equação qualquer em que é possível observar a descida do gradiente em busca do mínimo valor da equação.

Figura 8 – Representação da descida do gradiente.

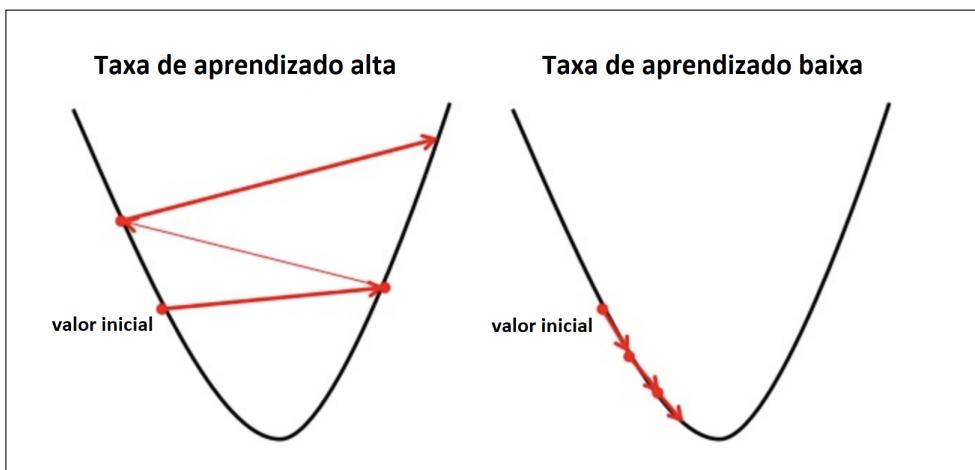


Fonte – (RANDOLFO, 2020)

2.3.6 Taxa de aprendizado

A taxa de aprendizado é um hiperparâmetro que determina o ritmo de atualização dos pesos no *backpropagation*. Uma taxa de aprendizado baixa deixa a atualização dos pesos mais suave e lenta e um valor mais alto deixa a atualização dos pesos caótica e rápida dependendo do tamanho da constante definida. Um valor de taxa de aprendizado muito alto aumenta a possibilidade do algoritmo não convergir para o mínimo da equação (PANDORFI *et al.*, 2011). Na figura 9 tem-se a representação gráfica do que acontece quando se tem uma taxa de aprendizado alta e outra com taxa de aprendizado baixa. enquanto uma converge para o mínimo da equação a outra não consegue convergir.

Figura 9 – Taxa de aprendizado.



Fonte – (JADHAV, 2019), modificado pelo autor.

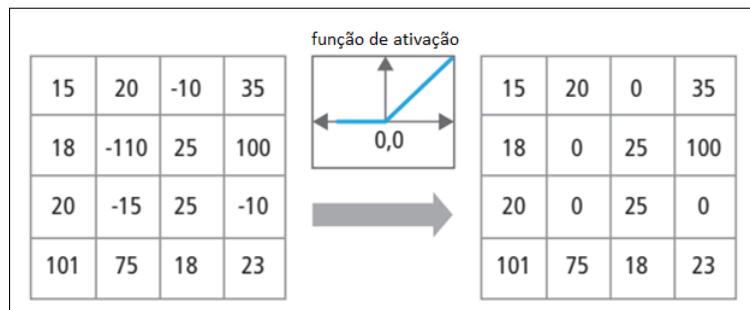
2.3.7 Função de ativação

Funções de ativação são funções matemáticas utilizadas para melhorar o resultado da saída de cada neurônio Rosa *et al.* (2018). A função de ativação Unidades Lineares Retificadas (ReLU) zera os valores de entrada que foram menores que zero. Isso acaba tornando o treinamento da rede mais rápido que as demais já que ela elimina as entradas com valores menores que zero. A função de ativação ReLu não envolve exponenciação e ela reduz o tempo de convergência dos parâmetros já que zera os valores de entrada menores que zero (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Pode-se descrever a equação da função de ativação ReLu na equação 2.2 . Na figura 10 pode-se observar a aplicação da função de ativação ReLu em uma matriz. Pode-se observar que a função de ativação zera os valores menores que zero priorizando os valores

maiores que zero.

$$f(x) = \begin{cases} 0 & \text{para } x < 0 \\ x & \text{para } x \geq 0 \end{cases} \quad (2.2)$$

Figura 10 – Aplicação da função de ativação ReLu.



Fonte – (HIJAZI; KUMAR; ROWEN, 2015), modificado pelo autor.

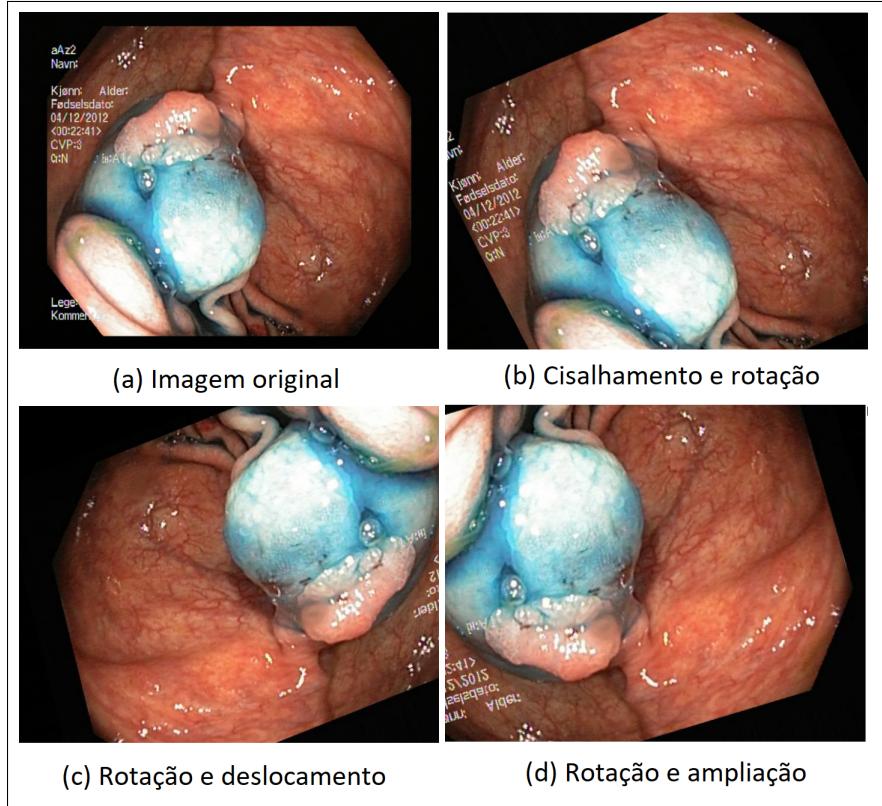
2.3.8 *Overfitting*

overfitting é o problema estatístico em que o modelo fica ajustado apenas para um conjunto de dados , basicamente a rede fica tendenciosa diminuindo a precisão da classificação, ou seja, o algoritmo se adapta aos dados de entrada tendo dificuldade para fazer a classificação de novos dados. Para minimizar o problema do *overfitting* utiliza-se a função *dropout*, que consiste em desativar alguns neurônios aleatoriamente para que a rede possa fazer a classificação correta para uma determinada imagem mesmo com esses neurônios desativados, garantindo que a rede não tenha tendência a fazer uma classificação específica (SRIVASTAVA *et al.*, 2014).

2.3.9 *Data Augmentation*

Data augmentation é uma técnica que busca aumentar a quantidade de imagens por meio de rotações, ampliações, diminuição ou qualquer outro tipo de transformação que a imagem possa passar para que se tenha uma imagem um pouco diferente da imagem original. A técnica *data augmentation* é viável para impulsionar o aprendizado de máquina de um pequeno conjunto de dados (PEREZ; WANG, 2017). Na figura 11 tem-se um exemplo da imagem de um exame gastrointestinal que passou por processos de rotação da imagem.

Figura 11 – Exemplo da técnica de *data augmentation*.



Fonte – (ASPERTI; MASTRONARDO, 2017), modificado pelo autor.

2.4 PYTHON PARA O APRENDIZADO DE MÁQUINA

Python é uma linguagem de programação bastante utilizada para desenvolver projetos de aprendizado de máquina. A linguagem *Python* é a mais popular no meio do aprendizado de máquina devido a diversos fatores como a simplicidade da linguagem, a popularidade com uma grande comunidade desenvolvendo bibliotecas e ajudando na solução de problemas (ACADEMY, 2020b). *Python* é uma linguagem ideal para prototipagem rápida já que é possível executar programas *Python* no navegador com auxílio de ferramentas como *Jupyter Notebook* (OLIPHANT, 2007).

A grande quantidade de bibliotecas de *Python* para aprendizado de máquina torna esta uma linguagem diferenciada. Na linguagem *Python* Tem-se bibliotecas para praticamente todas as funções básicas de aprendizado de máquina como a *NumPy* que utilizada para operações matemáticas, *Pandas* para a trabalhar com arquivos, *Matplotlib* que é utilizada para visualização de imagens, *TensorFlow* e *Keras* que são utilizados para a própria construção do modelo de aprendizado profundo.

2.5 TENSORFLOW

O *TensorFlow* é uma biblioteca de código aberto, desenvolvida pelo *GooGle*, focada no desenvolvimento de modelos de aprendizado de máquina. O *TensorFlow* oferece diversas ferramentas flexíveis para a computação probabilística fornecendo métodos rápidos e estáveis para gerar amostras e estatísticas (DILLON *et al.*, 2017). O trabalho de Taqi *et al.* (2018) apresenta uma característica do *TensorFlow* que é a habilidade da própria biblioteca fazer a estimativa dos gradientes necessários para a otimização das variáveis e melhorar o desempenho do modelo. De acordo com Abadi *et al.* (2016) o *TensorFlow* é um sistema de aprendizado de máquina que opera em locais heterogêneos.

2.5.1 Calculo do erro

O *TensorFlow* oferece diversas formas de calcular o erro da CNN. Como por exemplo:

- Erro médio absoluto ou *Mean Absolute Error*: Faz o cálculo da média da diferença absoluta entre os dados para verificação e as previsões. Faz o somatório de todos os valores previstos menos os valores de verificação e tira a média. O erro médio absoluto é descrito pela equação 2.3 (TENSORFLOW, 2020).

$$mae = \frac{\sum_{i=1}^n abs(y_i - \lambda(x_i))}{n} \quad (2.3)$$

- Erro Quadrático Médio ou *Mean Squared Error*: Faz o cálculo da média entre a diferença entre os dados para verificação ao quadrado e as previsões ao quadrado. Faz o somatório de todos os valores previstos menos os valores de verificação, eleva o resultado ao quadrado e tira a média. O erro Quadrático Médio é descrito pela equação 2.4 (TENSORFLOW, 2020).

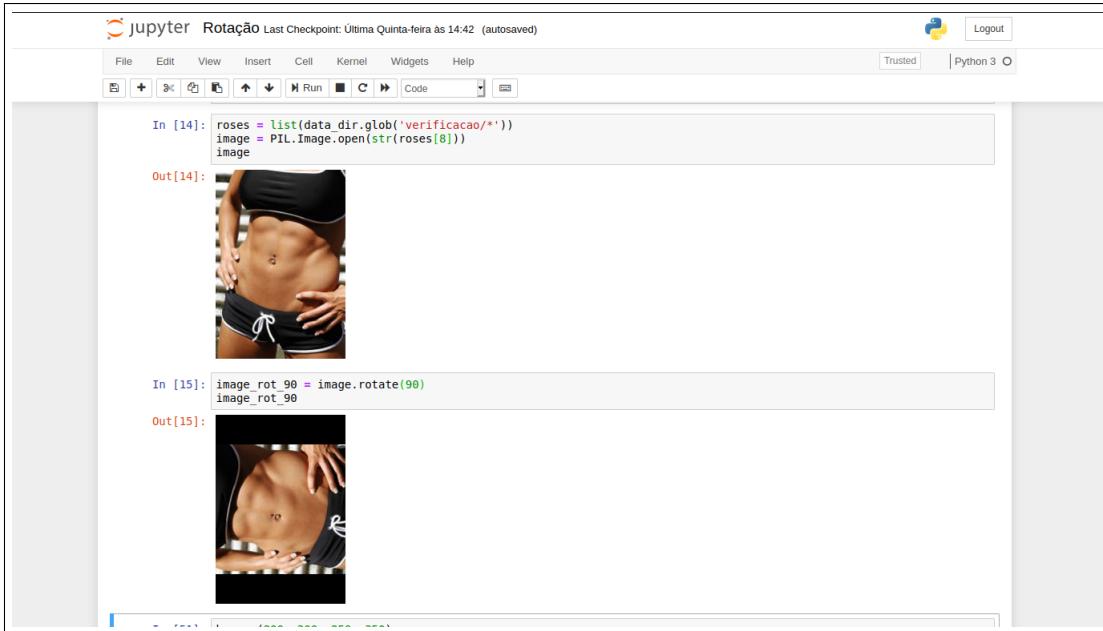
$$MSE = \frac{1}{n} \sum (y - \hat{y})^2 \quad (2.4)$$

2.6 JUPYTER NOTEBOOK

Jupyter notebook é um ambiente de programação *open-source*, em que o usuário pode programar em uma página *web* com interface gráfica. O *jupyter notebook* possui blocos de

programação que permite que apenas uma parte do código seja executada por vez, de uma forma mais didática e mostrado os resultados logo em seguida (RANDLES *et al.*, 2017). Na figura 5 tem-se ilustrado o exemplo de aplicação do *jupyter notebook*, em que é executado um comando para exibir uma figura e em seguida um comando para a imagem em 90°.

Figura 12 – Tela do *jupyter notebook*.



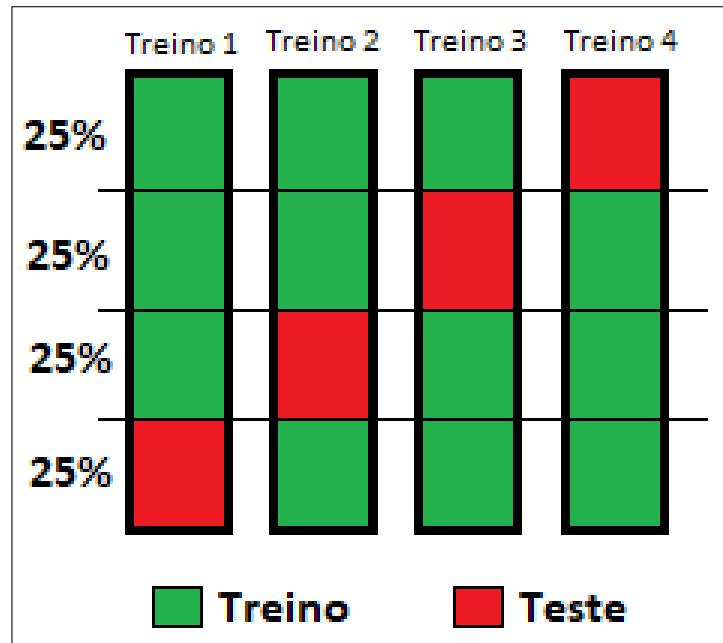
Fonte – O autor.

2.7 VALIDAÇÃO CRUZADA

A validação cruzada é um procedimento estatístico livre de distribuição teórica que vem ganhando aplicação recente. A validação cruzada é utilizada para avaliar a capacidade de generalização em problemas de aprendizado de máquina . A validação cruzada consiste em dividir por um valor arbitrário um grupo de dados. Dividindo o banco de dados em 4 partes iguais tem-se a validação cruzada de 4 dobras, dividindo em 10 tem-se a validação cruzada de 10 dobras e assim por diante. O intuito de fazer a divisão é utilizar uma parte diferente do banco de dados em cada treino (PIOVESAN; ARAÚJO; DIAS, 2009).

Na figura 13 é ilustrado como funciona a validação cruzada de 4 dobras, em que cada treino o algoritmo de aprendizado de máquina utiliza uma parte diferente do banco de dados para treinar e testar.

Figura 13 – Validação cruzada de 4 dobras.



Fonte – O autor.

3 TRABALHOS RELACIONADOS

Wang, Teoh e Choi (2018) abordam a importância do diagnóstico precoce de câncer de próstata para o sucesso do tratamento. A biopsia invasiva apresenta o diagnóstico definitivo, entretanto existe a possibilidade de efeitos colaterais e riscos. Pensando em reduzir o custo financeiro e os riscos do exame Wang, Teoh e Choi (2018) propõem alguns modelos preditivos utilizando aprendizado de máquina para fazer o diagnóstico. Os modelos preditivos utilizaram quatro tipos de aprendizado de máquina *Support Vector Machine* (SVM), *Least Squares Support Vector Machine* (LS-SVM), *Artificial Neural Network* (ANN) e *Random Forest* (RF). Os modelos utilizam os dados obtidos na pré-biópsia. Wang, Teoh e Choi (2018) avaliaram os modelos preditivos utilizando o pré-biópsia de 1.625 chineses com câncer de próstata no hospital de Hong Kong. Os quatro modelos preditivos tiveram bons resultados o modelo ANN obteve os melhores resultados com uma precisão de 95,27% com um valor de acurácia de 97,55%.

Li *et al.* (2019) abordam como é importante fazer a triagem pré-natal para a Síndrome de Down (SD) e como é recomendado para todas as mulheres independentemente de idade ou histórico familiar. A SD é causada pela cópia de um terceiro cromossomo 21 que pode aparecer de forma parcial ou completamente. Li *et al.* (2019) apresentam um trabalho pensando em reduzir os custos financeiros e sociais causados pelos exames necessários para o diagnóstico no pré-natal de SD. A utilização de um modelo preditivo com aprendizado de máquina para SD é prejudicado devido a dificuldade de lidar com os dados da triagem que muitas vezes são desequilibrados e relacionados a recursos. Pensado nos problemas da utilização de aprendizado de máquina para o pré-natal de SD , Li *et al.* (2019) propõem uma estrutura de aprendizado de máquina projetada para o diagnóstico no pré-natal de SD , com três estágios complementares. No primeiro estágio faz um julgamento com a técnica de floresta de isolamento. No segundo estágio faz o conjunto de modelos por estratégia de votação. No terceiro estágio faz o julgamento final utilizando regressão logística. O resultado da estrutura nos dados de triagem é superior ao de alguns métodos de aprendizado de máquina. Os melhores resultados foram obtidos utilizando como dados de entrada a alfa-fetoproteína, gonadotrofina coriônica humana, estriol não conjugado e idade materna. O método de Li *et al.* (2019) consegue fazer previsões mais precisas para dados desequilibrados e correlacionados a características, resultando uma abordagem nova e eficaz para análise de outros tipos de doenças.

Ongsuk *et al.* (2018) discutem como o colangiocarcinoma é o subconjunto de câncer de fígado. Um dos cinco principais tipos de câncer na Tailândia, uma das causas de morte

na Tailândia que vem crescendo desde 2014. Ongsuk *et al.* (2018) apresentam uma solução para reduzir os riscos de colangiocarcinoma é preciso encontrar os fatores que causam câncer e prever a probabilidade de câncer o mais cedo possível. Mas um problema para prever o colangiocarcinoma é que a probabilidade do paciente possuir ele é de 1%. Pensando nesses problemas Ongsuk *et al.* (2018) propõem um modelo de aprendizado de máquina adaptativo para o prognóstico do câncer de colangiocarcinoma. Utilizando a técnica *CanWiser* para aprender automaticamente com o conjunto de dados do paciente. A estrutura proposta por Ongsuk *et al.* (2018) conseguem gerar um modelo de predição com uma sensibilidade de 75%, especificidade de 83,41% e precisão de 83,34%. O trabalho de Ongsuk *et al.* (2018) conseguem gerar de forma adaptativa novos modelos que são ajustados de acordo com um novo conjunto de dados.

Rangel, Lima e Vargas (2015) mostra os fatores que contribuem para o diagnóstico tardio de câncer de colo de útero em pacientes do Instituto Nacional do Câncer no Rio de Janeiro. Fatores como os relacionados à acessibilidade, incluindo os problemas como integração e ação dos serviços, sentimentos pessoais, valores e costumes que distanciam das práticas preventivas. Outro fator importante que distancia as mulheres das práticas de preventivas é a fragilidade entre o vínculo entre as pacientes e os profissionais de saúde, o que acaba diminuindo a resposta ao serviço e dificultando a continuidade do cuidado. Esses fatores acabam se relacionando e comprometendo a busca por orientação e tratamento das pacientes, prejudicando a saúde das pacientes. Rangel, Lima e Vargas (2015) entrevistou 9 mulheres maiores de 18 anos e que não portavam doenças mentais, com idade variando de 31 a 76 anos. O grau de escolaridade das entrevistadas foi de 2 com ensino médio, 1 com ensino médio e técnico, 3 apenas com o ensino fundamental, 1 com ensino fundamental incompleto, 1 apenas alfabetizada e 1 sem nenhum estudo. O trabalho de Rangel, Lima e Vargas (2015) contribui para a ampliação e compreensão do problema da relação dos diagnósticos tardios de câncer de colo de útero, visado diminuir o quadro de morbidade da doença no País.

Cedaro *et al.* (2020) fizeram um estudo para conhecer como as pessoas portadoras da Doença de Huntington (DH) percorrem o sistema de saúde em busca do diagnóstico e tratamento. Cedaro *et al.* (2020) entrevistaram 5 pacientes com DH. Os pacientes com DH começam na atenção primária em saúde, realizando os procedimentos incluindo exames laboratoriais passando a serem encaminhados para outros níveis de complexidade. Após passar pela atenção primária os pacientes são encaminhados para a atenção secundária e terciária para realizar exames mais complexos, como tomografias, ressonância magnética e atendimentos especializados. Após esses procedimentos os pacientes vão para as instituições privadas e os centros de pesquisa acadêmica

para que eles possam prestar exames e apoio terapêutico. A pessoa com doenças raras enfrentam grandes dificuldades em busca de seu diagnóstico, muitos dos pacientes acabam ficando perdidos dentro da Rede de Atenção à Saúde (RAS), indo por caminhos sem a orientação correta.

4 METODOLOGIA

Neste capítulo é apresentado todas as etapas necessárias para o desenvolvimento da CNN para o diagnóstico de LGC . Por isso este capítulo é dividido entre cada uma das etapas para o desenvolvimento do projeto, que são:

- *Hardware e Software* utilizados.
- Montagem do banco de dados:
 1. Ampliação do banco de dados ou *data augmentation*.
 2. Modificar as imagens para trabalhar na CNN.
- Configuração da CNN.

4.1 HARDWARE E SOFTWARE

Antes de iniciar as etapas práticas é importante apresentar o *Hardware e Software* em que o projeto foi construído.

O projeto foi desenvolvido em um notebook sem GPU, com 4 GB de RAM e uma Unidade Central de Processamento (CPU) com o processador intel i5-4210 U 1.70 GHz 2.40 GHz.

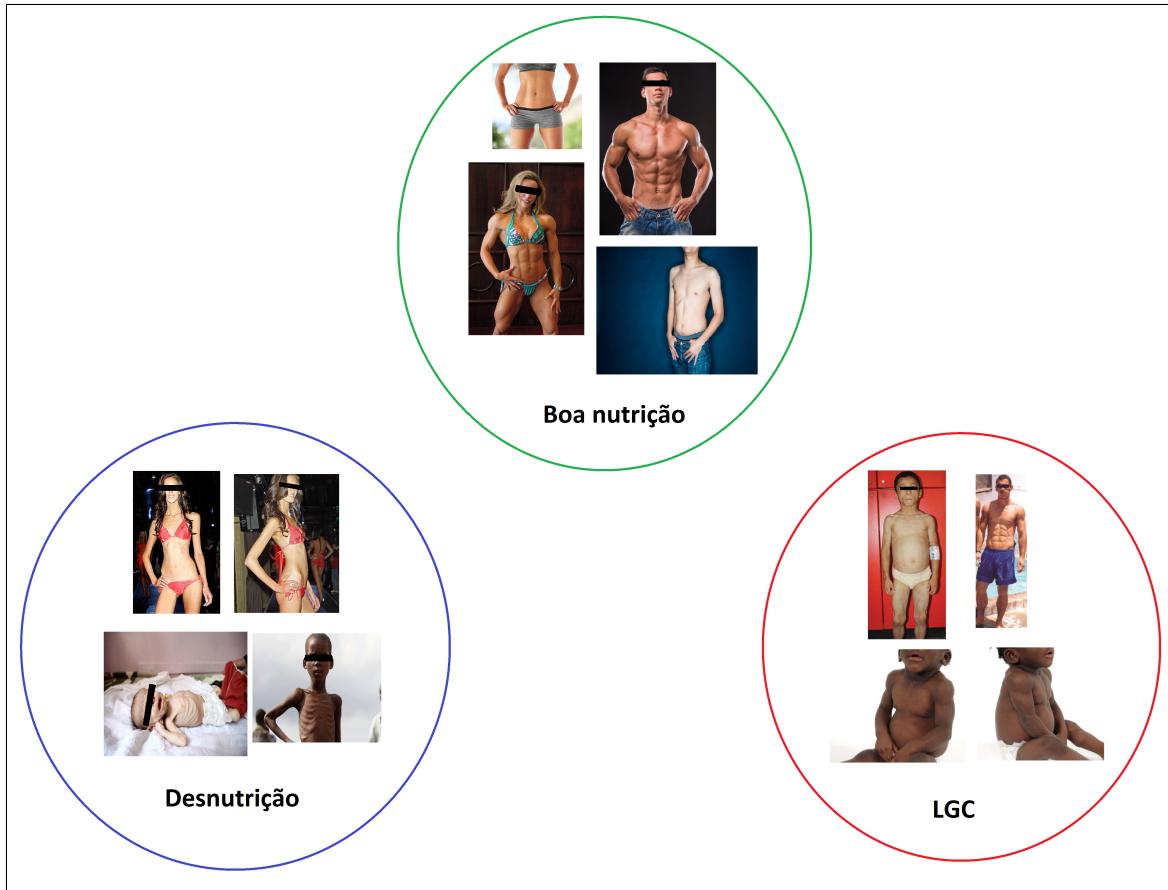
Todo o código do projeto é desenvolvido em *Python 3*. As importantes bibliotecas de software utilizadas são *Numpy* versão 1.17.4, *PIL* versão 6.2 e *Tensorflow* versão 1.15.

4.2 MONTAGEM DO BANCO DE DADOS

Para o desenvolvimento do projeto utilizou-se um banco de dados montado em parceria com alunos de doutorado em saúde coletiva da Universidade Federal do Ceará (UFC). Como a LGC está relacionada a um baixo percentual de gordura corporal, o banco de dado é composto por imagens de pessoas com boa nutrição, com desnutrição e com a LGC .O banco de dados é dividido em três grupos, no primeiro grupo tem-se imagens de pacientes desnutridos. No segundo grupo tem-se imagens de pacientes eutróficos, que são pessoas com uma boa nutrição. No terceiro grupo tem-se imagens com pessoas com a LGC. Estas informações estão ilustradas na figura 14 , na qual é possível verificar algumas imagens que estão dentro dos grupos.

O banco de dado é composto por um total de 337 imagens, no qual 32 imagens são de pessoa com desnutrição, 48 imagens de pessoas com boa nutrição e 257 imagens de pessoas com LGC.

Figura 14 – Grupos do banco de dado.



Fonte – O autor

4.2.1 Ampliação do banco de dados

Como no grupo de pessoas com desnutrição e boa nutrição possui uma quantidade relativamente baixa de imagens, torna-se importante ampliar a quantidade de imagens. Para aumentar o número de imagens utiliza-se a técnica *data augmentation*, que consiste em fazer uma imagem passar por processos de rotação, ampliação, redução ou qualquer outro processo que resulte em uma imagem um pouco diferente da original. Na figura 15 é apresentando uma imagem utilizado no trabalho que passou pelo processo de *data augmentation*. Para realizar os processos de rotação, ampliação e redução utilizou-se comandos próprios do *python* para essa tarefa.

Para o trabalho, cada imagem do grupo de pessoas com desnutrição e boa nutrição passou por 8 processos escolhidos arbitrariamente, resultando em uma nova imagem. São eles:

- rotacionar a imagem em 45°.
- rotacionar a imagem em 90°.
- rotacionar a imagem em 180°.

Figura 15 – Aplicação da técnica *data augmentation*.



Fonte – O autor

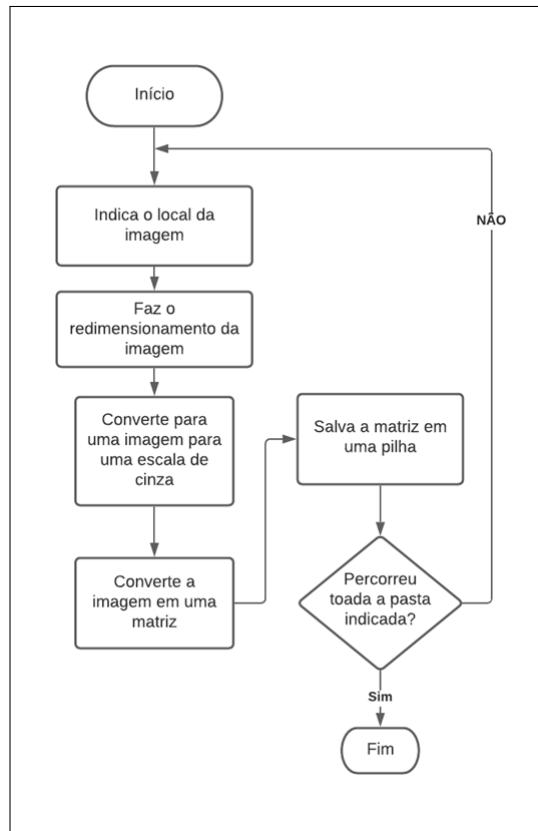
- ampliar e rotacionar a imagem em 18° .
- ampliar e rotacionar a imagem em 114° .
- rotacionar a imagem em -90° .
- rotacionar a imagem em -50° .
- rotacionar a imagem em -45° .

Com o procedimento de *data augmentation* conseguiu-se aumentar o número de imagens de 80 para 640 imagens.

4.2.2 Tratamento das imagens

Para utilizar o algoritmo da CNN é preciso transformar a imagem em uma matriz. Para fazer a transformação programou-se em python um algoritmo para fazer esta tarefa. Primeiro é informado em qual pasta a imagem está, em seguida a imagem a imagem tem o tamanho redimensionado, depois a imagem é transformada em uma escala de cinza. Em seguida, utiliza um comando em python que converte a imagem em uma matriz. Para finalizar adiciona-se a matriz que representa a imagem em uma pilha para utilizar na CNN. Na figura 16 tem-se o fluxograma do código utilizado para fazer os processos de redimensionamento do tamanho da imagem, de converter a imagem em uma escala de cinza, de converter a imagem para uma matriz e salvar essa matriz em uma pilha.

Figura 16 – Fluxograma do código de tratamento das imagens.



Fonte – O autor

4.3 REDE NEURAL CONVOLUCIONAL CNN

Pensando em desenvolver um modelo de CNN capaz de classificar uma imagem em 3 grupos, o de pessoas com desnutrição, com boa alimentação e pessoas com LGC. Em que as pessoas com desnutrição são do grupo 0, pessoas com boa nutrição do grupo 1 e pessoas com LGC do grupo 2. Escreveu-se um algoritmo de aprendizado de máquina utilizando os comandos do *TensorFlow* na linguagem *python*. Boa parte do processo já é pré programado, a dificuldade está em ajustar os parâmetros da rede CNN para fazer a classificação. Os principais parâmetros de uma CNN padrão que devem ser ajustados são:

- Número de camadas de convolução.
- Tamanho do filtro de características *kernel size*.
- Número de camadas de agrupamento *pooling*.
- Tamanho do filtro de agrupamento.
- Número de camadas ocultas.
- Número de neurônios por camada.
- Percentual de neurônios que são desligados para reduzir o *overfitting*.

- Função de ativação.
- Taxa de aprendizagem.
- Número máximo de épocas.

Observando os parâmetros da CNN fixou-se os parâmetros que são apresentados na tabela 2. Escolheu-se fixar os parâmetros da tabela 2 de acordo com a documentação do *TensorFlow*. Que são uma das opções de valores recomendados para iniciar os testes e a partir deles ir mudando de acordo com a necessidade da CNN.

Tabela 2 – Parâmetros definidos

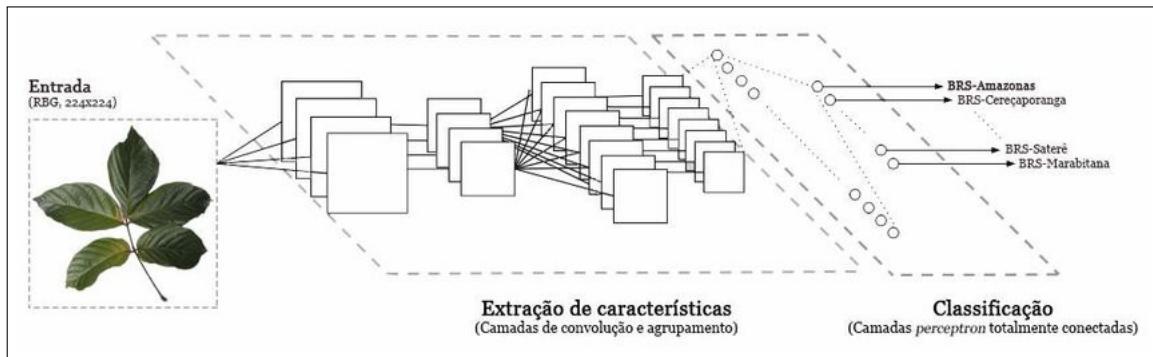
Número camadas de convolução	2
Ordem do filtro de características	[2x2]
Número de camadas de agrupamento	2
Taxa de aprendizagem	0,001
Quantidade de neurônios por camada	1024
Percentual de neurônios à desligar	20%
Função de ativação	ReLU

Fonte – O autor

Com os parâmetros definidos basta utilizar os comandos da biblioteca *TensorFlow* e seguir a estrutura da CNN. Primeiro monta-se a camada de entrada passando a dimensão da imagem que a CNN vai trabalhar e se a rede vai trabalhar com imagens coloridas ou em escala de cinza. Em seguida define-se a camada de convolução em que é definido o tamanho do filtro kernel size. Com isso feito monta-se as camadas ocultas, nela é definido o número de neurônios e a função de ativação. Em seguida monta-se a camada de saída em que é definido o número de saídas da CNN. Com a estrutura da rede construída, define-se a taxa de aprendizado da rede e escolhe-se a equação do Erro Quadrático Médio (MSE), que é uma das equações para calcular o erro já programadas no *TensorFlow*.

Na figura 17 tem-se a representação da estrutura de uma CNN , nela é possível observar a ordem das camadas que é preciso seguir para a configuração da CNN.

Figura 17 – Fluxograma da CNN.



Fonte – (LACERDA, 2019)

5 RESULTADOS

Para realizar o treinamento da CNN utilizou-se 75% do banco de dados para fazer o treino e 25% para fazer o teste. Com o percentual de imagens para treino e teste definido, determinou os parâmetros para fazer o treinamento da rede. Realizou-se treinos em que a diferença estava na quantidade de camadas ocultas por isso separou-se os treinos em 2 grupos. No primeiro grupo, a CNN foi configurada com apenas 1 camada oculta. No segundo grupo, a CNN foi configurada com 5 camadas ocultas. Os parâmetros que são fixos e iguais para os dois grupos são apresentados na tabela 2.

Em cada treinamento do grupo A e B, repetiu-se 5 vezes para verificar o percentual de acerto médio dos 5 treinamentos e a média do Erro Quadrático que foi calculado pelo *tensorflow*. Somando o tempo gasto para fazer o treinamento dos modelos do grupo A, grupo B e os treinos da validação cruzada gastou-se aproximadamente 22 horas de treinamento. Na tabela 3 pode-se observar os resultados médios das 5 repetições para cada uma das 3 configurações dos modelos do grupo A. Na tabela 4 pode-se observar os resultados médios das 5 repetições para cada uma das 3 configurações dos modelos do grupo B.

Tabela 3 – Parâmetros do grupo A

camadas ocultas	numero máx. de épocas	acurácia média	Tempo médio (minutos)
1	200	58,93%	8,56
1	600	62,32%	26,20
1	1000	63,10%	43,70

Fonte – O autor

Tabela 4 – Parâmetros do grupo B

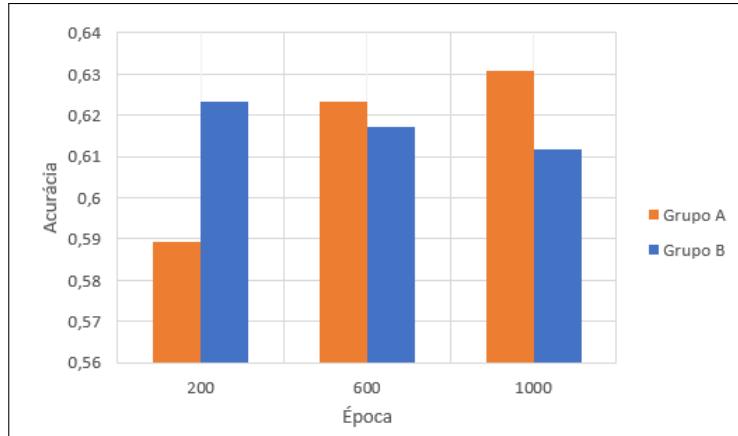
camadas ocultas	numero máx. de épocas	acurácia média	Tempo médio (minutos)
5	200	62,32%	9,58
5	600	61,72%	29,00
5	1000	61,16%	48,00

Fonte – O autor

Na figura 18 pode-se observar a representação da relação entre a acurácia e o número de épocas. Em que a acurácia de um modelo com 5 camadas ocultas consegue ser elevado mesmo com um baixo número de épocas para a realização do treinamento. Entretanto o desempenho do modelo tende a cair quando se aumenta o número de épocas já que acaba sendo necessário mais tempo para fazer o treinamento e com aumento no número de épocas é possível verificar

uma queda na acurácia do modelo. Esses fatores acabam favorecendo os modelos com poucas camadas ocultas, que têm um desempenho um pouco maior já que leva-se menos tempo para se realizar os treinamentos devido ao baixo número de camadas ocultas, também é possível verificar que a acurácia tende a aumentar junto com o número de épocas.

Figura 18 – Acurácia relação ao número de épocas.



Fonte – O autor

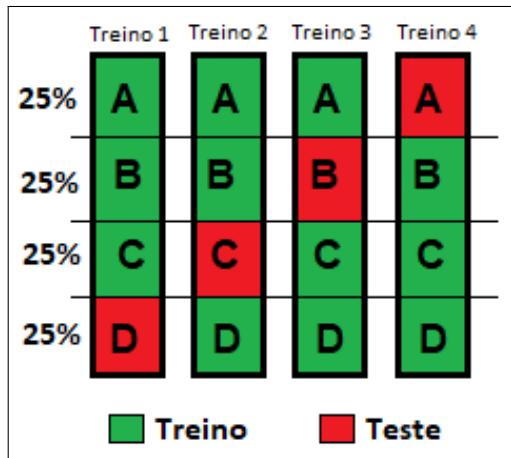
Em seguida escolheu a CNN do grupo A com 600 épocas para aplicar a validação cruzada de 4 dobras. Pois obteve o maior valor de acurácia e é mais leve que a rede com 1000 épocas, gastando menos tempo para realizar os treinamentos. A tabela 5 apresenta os resultados da validação cruzada de 4 dobras. Na figura 19 tem-se a representação de como o banco de dados foi dividido para fazer a validação cruzada de 4 dobras. Na tabela 5 é possível verificar que o modelo mantém estável as alterações feitas no conjunto de dados para treino e teste, desde que os hiperparâmetros sejam fixados. A maior variação é entre a acurácia de 66,52% utilizando os dados de A,B,D para treino e a acurácia de 60,49% utilizando os dados de B,C,D para treino.

Tabela 5 – Resultado da validação cruzada de 4 dobras.

Treino	Teste	Acurácia Média	Tempo médio (minutos)
A,B,C	D	62,32%	26,20
A,B,D	C	66,52%	25,40
A,C,D	B	62,50%	27,00
B,C,D	A	60,49%	26,20
	Média	62,41%	26,20

Fonte – O autor

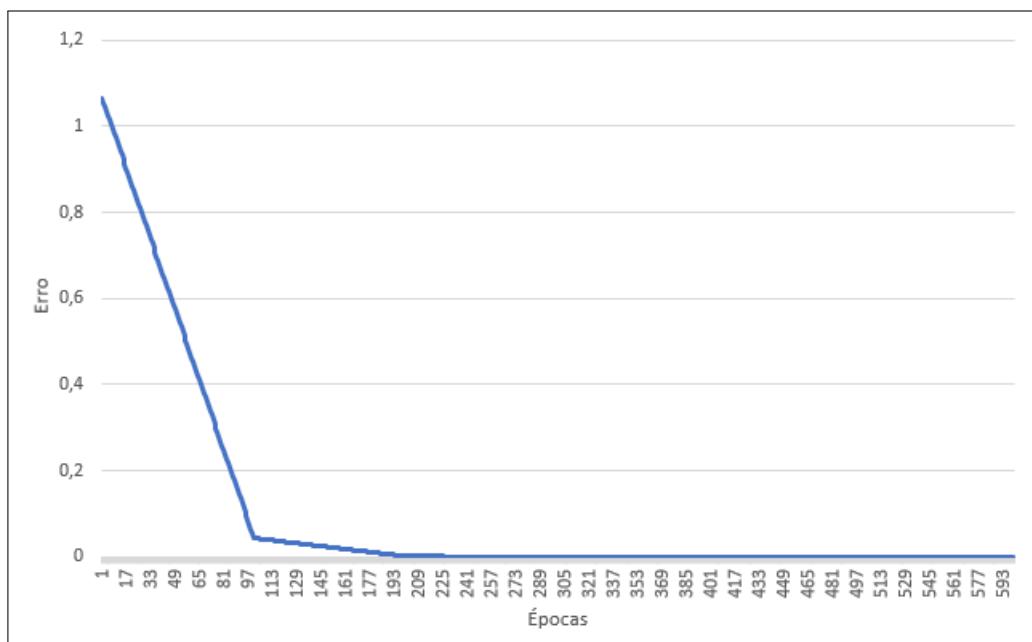
Figura 19 – Divisão para a validação cruzada de 4 dobras.



Fonte – O autor

Com o decorrer do tempo o modelo tende a convergir o resultado calculado para o valor que é esperado, minimizando o erro à medida que as épocas vão aumentando. Representou-se graficamente na figura 20 o erro calculado durante o processo de treinamento, para o modelo com 1 camada oculta e 600 épocas. Em que pode-se observar que o erro inicia com um valor um pouco acima de 1 e em seguida passa a cair chegando bem perto de um erro 0 a partir da época 225, mantendo o erro bem próximo de 0 até o final na época 600.

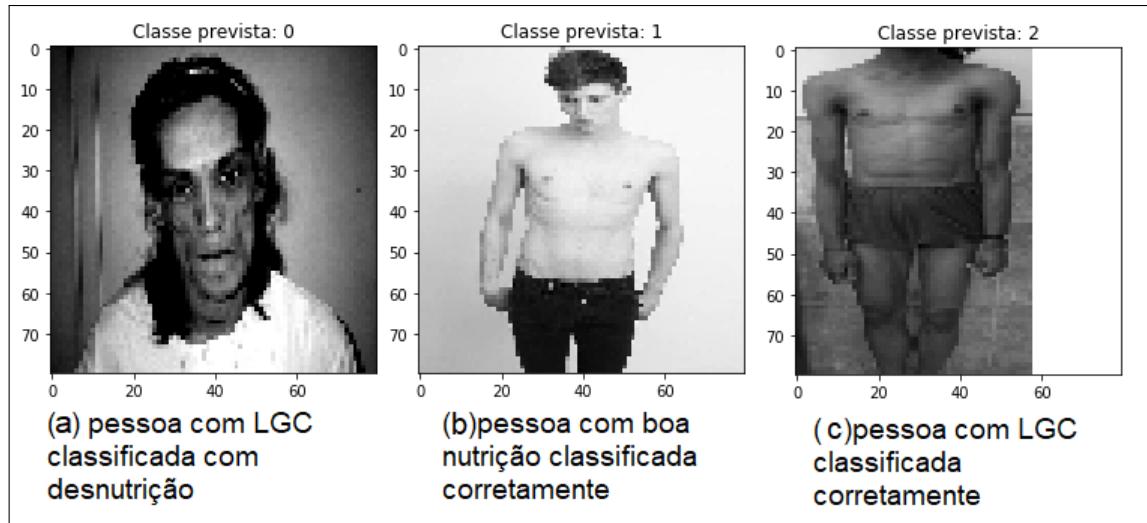
Figura 20 – Erro durante o processo de treinamento.



Fonte – O autor

Na figura 21 pode-se observar 3 imagens que passaram pelo processo de classificação do modelo. Nesta etapa, obteve-se um erro de classificação da figura 21(a) e (c), acertando apenas a figura 21(b). A classe 0 representa as pessoas com desnutrição, a classe 1 pessoas com boa nutrição, a classe 2 pessoas com LGC.

Figura 21 – Imagens classificadas pelo modelo.



Fonte – O autor

Na figura 22 é apresentado a matriz de confusão para este modelo com 1 camada oculta e 600 épocas, em que na diagonal com os quadrados azuis estão os números de classificações corretas e nos quadrados vermelhos estão os números de classificações erradas os falsos positivos. O modelo apresentou 4 classificações corretas para desnutridos, 85 classificações corretas para boa nutrição e 53 classificações corretas para LGC, o modelo apresentou um total de 69 falsos positivos e 13 falsos negativos. Com isso pode-se observar que o modelo acerta 63,39%.

Figura 22 – Matriz de confusão.

		Classe real		
		Desnutrição	Boa nutrição	LGC
Classificação	Desnutrição	4	2	0
	Boa nutrição	59	85	11
	LGC	1	9	53

Fonte – O autor

Ainda da matriz de confusão ilustrada na figura 22, analisando apenas os casos de LGC pode-se observar que das 64 imagens de LGC, 53 foram classificadas corretamente, o que significa que 82% das imagens que são do grupo LGC foram classificadas corretamente. A classificação da LGC apresentou 10 falsos positivos, em que as imagens eram dos outros grupos e foram classificadas como LGC, 1 imagem de pessoa com desnutrição foi classificada como LGC e 9 imagens de pessoas com boa nutrição foram classificadas como LGC. Os casos de LGC apresentaram 11 falsos negativos, em que eram imagens de pessoas com LGC e foram classificadas como pessoas com boa nutrição.

Analizando a matriz de confusão pode-se observar que se o classificador fosse um classificador binário para dizer apenas se a pessoa apresenta ou não as características da LGC o percentual de acurácia subiria de 63,39% para 90,62%. O modelo não está com problemas para identificar as características de LGC e sim para diferenciar as características de desnutrição. Na figura 23 tem-se a representação da matriz de confusão fazendo o agrupamento das classes de desnutrição com as pessoas com boa nutrição, em que nos quadrados vermelhos e azuis dentro do quadrado verde seriam classificações corretas.

Figura 23 – Matriz de confusão com dois grupos.

		Classe real		
		Desnutrição	Boa nutrição	LGC
Classificação	Desnutrição	4	2	0
	Boa nutrição	59	85	11
	LGC	1	9	53

Fonte – O autor

6 CONCLUSÃO

O desenvolvimento de um modelo de aprendizado de máquina para diagnóstico de LGC foi alcançado através de ajustes finos e vários testes. O desempenho deste modelo não conseguiu atingir um percentual muito elevado de precisão, fazendo previsões variando de 58% e 63%.

O desenvolvimento de um modelo de aprendizado de máquina para diagnóstico de LGC foi alcançado através de ajustes finos e vários testes. O desempenho deste modelo não conseguiu atingir um percentual muito elevado de taxa de acerto, conseguindo fazer a classificação das imagens nos 3 grupos com uma precisão variando entre 58% e 63%. Analisando os parâmetros estabelecidos verificou-se que aumentar o número de camadas ocultas de 1 para 5 não significou aumento na acurácia do modelo, pelo contrário teve uma queda na performance já que o tempo para fazer o treinamento aumentou em média 11%. Os modelos com apenas 1 camada oculta acaba tendo um desempenho maior, já que leva menos tempo para treinar. Por meio da validação cruzada de 4 dobras consegue-se verificar a eficácia do modelo, fixando os hiperparâmetros e alterando apenas a ordem dos dados utilizados no treino e teste, mantendo a acurácia com variação de no máximo 4%. Além das limitações de *hardware*, já citadas, o projeto teve problemas com a padronização do banco de dados. As imagens que formam o banco de dados são bem diferentes entre si, tiradas em ângulos diferentes, situações diferentes dentre outras diferenças que prejudicam o modelo de aprendizado de máquina convergir para o resultado. Em um grupo de dados para teste composto de 64 imagens de pessoas com LGC, 53 foram classificadas corretamente apresentando um percentual de acerto de 82% mesmo com a limitação de hardware e com um banco de dados despadronizado. Focando a análise dos resultados apenas no grupo de teste de pessoas com LGC conseguiu-se um bom percentual de acerto chegando a 82%.

A importância do trabalho está diretamente relacionado a analisar formas de auxiliar os profissionais da saúde para fazer diagnóstico de doenças que são facilmente confundidas com outras doenças por causa de sintomas semelhantes. As características de uma doença rara, como o baixo número de casos, acaba levando os profissionais da saúde a descartar a possibilidade de ser uma doença rara.

Para trabalhos futuros é importante estudar formas de fazer a padronização do banco de dados utilizado, retirando o fundo das imagens e deixar apenas a pessoa que está sendo analisada. A aplicação de GPU para o desenvolvimento do algoritmo de aprendizado de máquina

deve ser considerada para melhorar a performance e com isso passar a ser possível realizar treinos com hiperparâmetros mais altos como número de épocas e a resolução da imagem.

REFERÊNCIAS

ABADI, M.; BARHAM, P.; CHEN, J.; CHEN, Z.; DAVIS, A.; DEAN, J.; DEVIN, M.; GHE-MAWAT, S.; IRVING, G.; ISARD, M. *et al.* Tensorflow: A system for large-scale machine learning. In: **12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)**. [S.l.: s.n.], 2016. p. 265–283.

ACADEMY, D. S. **Capítulo 15 – Algoritmo Backpropagation Parte 2.** 2020. Disponível em: <<http://deeplearningbook.com.br/algoritmo-backpropagation-parte-2-treinamento-de-redes-neurais/>>. Acesso em: 11 novembro 2020.

ACADEMY, D. S. **POR QUE A LINGUAGEM PYTHON É TÃO POPULAR EM MACHINE LEARNING E INTELIGÊNCIA ARTIFICIAL?** 2020. Disponível em: <<http://deeplearningbook.com.br/algoritmo-backpropagation-parte-2-treinamento-de-redes-neurais/>>. Acesso em: 10 novembro 2020.

ALVES, G. **Entendendo Redes Convolucionais (CNNs).** 2018. Disponível em: <<https://medium.com/neuronio-br/entendendo-redes-convolucionais-cnns-d10359f21184#:~:text=O%20reconhecimento%20de%20imagem%20%C3%A9,entre%20convolu%C3%A7%C3%A1%C3%B5es%20e%20fully%20connected.>> Acesso em: 30 de novembro de 2020.

ASPERTI, A.; MASTRONARDO, C. The effectiveness of data augmentation for detection of gastrointestinal diseases from endoscopical images. **arXiv preprint arXiv:1712.03689**, 2017.

BERARDINELLI, W. An undiagnosed endocrinometabolic syndrome: report of 2 cases. **The Journal of Clinical Endocrinology & Metabolism**, Oxford University Press, v. 14, n. 2, p. 193–204, 1954.

BONATTO, D.; FELTES, B. C. **Conformational study of DDB2- DDB1 protein complex and its mutant variants in Xeroderma Pigmentosum disease (Portuguese).** Tese (Doutorado), 08 2017.

CEDARO, J. J.; CANIZARES, V. S. de A.; RAMOS, N. O.; FRANÇA, A. K. de; XAVIER, J. do N.; CAMPELO, T. N. C.; GONÇALVES, T. L. P.; MEDEIROS, J. G. A. de. Doença neurodegenerativa rara: itinerário de portadores de doença de huntington em busca de diagnóstico e tratamento. **Brazilian Journal of Health Review**, v. 3, n. 5, p. 13182–13197, 2020.

DILLON, J. V.; LANGMORE, I.; TRAN, D.; BREVDO, E.; VASUDEVAN, S.; MOORE, D.; PATTON, B.; ALEMI, A.; HOFFMAN, M.; SAUROUS, R. A. Tensorflow distributions. **arXiv preprint arXiv:1711.10604**, 2017.

FARIA, C. A.; MORAES, R. S.; SOBRAL-FILHO, D. C.; REGO, A. G.; BARACHO, M. F.; EGITO, E. S.; BRANDÃO-NETO, J. Autonomic modulation in patients with congenital generalized lipodystrophy (berardinelli-seip syndrome). **Europace**, Oxford University Press, v. 11, n. 6, p. 763–769, 2009.

GARG, A. Adipose tissue dysfunction in obesity and lipodystrophy. **Clinical cornerstone**, Elsevier, v. 8, p. S7–S13, 2006.

GRANATY, J. **TensorFlow: Machine Learning e Deep Learning com Python.** [S.l.]: udemy, 2019.

- HAYKIN, S. **Redes neurais: princípios e prática.** [S.I.]: Bookman Editora, 2007.
- HIJAZI, S.; KUMAR, R.; ROWEN, C. Using convolutional neural networks for image recognition. **Cadence Design Systems Inc.: San Jose, CA, USA**, p. 1–12, 2015.
- HURWITZ, J.; KIRSCH, D. Machine learning for dummies. **IBM Limited Edition**, John Wiley & Sons, Inc, v. 75, 2018.
- JADHAV, M. **Gradient descent: Why and How ?** 2019. Disponível em: <<https://medium.com/analytic...-e369950ae7d3>>. Acesso em: 30 de novembro de 2020.
- KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F.; BURGES, C. J. C.; BOTTOU, L.; WEINBERGER, K. Q. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2012. v. 25, p. 1097–1105. Disponível em: <<https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>>.
- LACERDA, L. **Deep Learning & Visão Computacional — REDES NEURAIS CONVOLUCIONAIS.** 2019. Disponível em: <<https://medium.com/@lucaaslb/deep-learning-vis%C3%A3o-computacional-redes-neurais-convolucionais-c21f19f5ec34>>. Acesso em: 12 de novembro de 2020.
- LI, L.; LIU, W.; ZHANG, H.; JIANG, Y.; HU, X.; LIU, R. Down syndrome prediction using a cascaded machine learning framework designed for imbalanced and feature-correlated data. **IEEE Access**, IEEE, v. 7, p. 97582–97593, 2019.
- MAGRÉ, J.; DELÉPINE, M.; KHALLOUF, E.; GEDDE-DAHL, T.; MALDERGEM, L. V.; SOBEL, E.; PAPP, J.; MEIER, M.; MÉGARBANÉ, A.; LATHROP, M. *et al.* Identification of the gene altered in berardinelli–seip congenital lipodystrophy on chromosome 11q13. **Nature genetics**, Nature Publishing Group, v. 28, n. 4, p. 365–370, 2001.
- MALDERGEM, L. V.; MAGRE, J.; KHALLOUF, T.; GEDDE-DAHL, T.; DELEPINE, M.; TRYGSTAD, O.; SEEMANOVA, E.; STEPHENSON, T.; ALBOTT, C.; BONNICI, F. *et al.* Genotype-phenotype relationships in berardinelli-seip congenital lipodystrophy. **Journal of medical genetics**, BMJ Publishing Group Ltd, v. 39, n. 10, p. 722–733, 2002.
- MARIM, Y. V. R. Detecção de objetos: Estudo e aplicação da arquitetura r-cnn. 2019.
- MARTINIANO, A.; FERREIRA, R. P.; FERREIRA, A.; FERREIRA, A.; SASSI, R. J. Utilizando uma rede neural artificial para aproximação da função de evolução do sistema de lorentz. **Revista Produção e Desenvolvimento**, v. 2, n. 1, p. 26–38, 2016.
- MOURA, A. A. D. H. E. A. D. A machine learning approach to semantic segmentation of surface defects in steel plates. 2019.
- NOLIS, T. Exploring the pathophysiology behind the more common genetic and acquired lipodystrophies. **Journal of human genetics**, Nature Publishing Group, v. 59, n. 1, p. 16–23, 2014.
- OLIPHANT, T. E. Python for scientific computing. **Computing in Science & Engineering**, IEEE, v. 9, n. 3, p. 10–20, 2007.

- ONGSUK, S.; KOMOLVATIN, S.; KUNAKORNTUM, I.; PHUNCHONGHARN, P.; AMONYINGCHAROEN, S.; HINTHONG, W. An adaptive cancer prognosis framework for cholangiocarcinoma based on machine learning techniques. In: IEEE. **2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII)**. [S.l.], 2018. p. 82–85.
- PACHECO, A. **Introdução a Redes Neurais Artificiais**. 2015. Disponível em: <<http://computacao-inteligente.com.br/artigos/redes-neurais-artificiais/>>. Acesso em: 30 de novembro de 2020.
- PANDORFI, H.; SILVA, I. J. O.; SARNIGHAUSEN, V. C. R.; VIEIRA, F. M. C.; NASCIMENTO, S. T.; GUISELINI, C. Uso de redes neurais artificiais para predição de índices zootécnicos nas fases de gestação e maternidade na suinocultura. **Revista Brasileira de Zootecnia**, SciELO Brasil, v. 40, n. 3, p. 676–681, 2011.
- PATNI, N.; GARG, A. Congenital generalized lipodystrophies—new insights into metabolic dysfunction. **Nature Reviews Endocrinology**, Nature Publishing Group, v. 11, n. 9, p. 522, 2015.
- PEREZ, L.; WANG, J. The effectiveness of data augmentation in image classification using deep learning. **arXiv preprint arXiv:1712.04621**, 2017.
- PIOVESAN, P.; ARAÚJO, L. B. d.; DIAS, C. T. d. S. Validação cruzada com correção de autovalores e regressão isotônica nos modelos de efeitos principais aditivos e interação multiplicativa. **Ciência Rural**, SciELO Brasil, v. 39, n. 4, p. 1018–1023, 2009.
- RANDLES, B. M.; PASQUETTO, I. V.; GOLSHAN, M. S.; BORGMAN, C. L. Using the jupyter notebook as a tool for open science: An empirical study. In: IEEE. **2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)**. [S.l.], 2017. p. 1–2.
- RANDOLFO, M. **Método de Ensemble: vantagens da combinação de diferentes estimadores**. 2020. Disponível em: <<https://sigmoidal.ai/metodo-de-ensemble-vantagens-da-combinacao-de-diferentes-estimadores/>>. Acesso em: 30 de novembro de 2020.
- RANGEL, G.; LIMA, L. D. d.; VARGAS, E. P. Condicionantes do diagnóstico tardio do câncer cervical na ótica das mulheres atendidas no inca. **Saúde em Debate**, SciELO Public Health, v. 39, p. 1065–1078, 2015.
- ROSA, R. d. P. *et al.* Método de classificação de pragas por meio de rede neural convolucional profunda. Universidade Estadual de Ponta Grossa, 2018.
- SALESFORCE. **Machine Learning e Deep Learning: aprenda as diferenças**. 2018. Disponível em: <<https://www.salesforce.com/br/blog/2018/4/Machine-Learning-e-Deep-Learning-aprenda-as-diferencias.html>>. Acesso em: 16 abril 2020.
- SEIP, M. Lipodystrophy and gigantism with associated endocrine manifestations. a new dien-cephalic syndrome? **Acta paediatrica**, v. 48, p. 555, 1959.
- SOARES, V. “**Nobel da Computação” vai para os pais do Deep Learning**. 2019. Disponível em: <<https://www.institutodeengenharia.org.br/site/2019/04/01/nobel-da-computacao-%E2%80%8B-vai-para-os-pais-do-deep-learning/>>. Acesso em: 16 abril 2020.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: a simple way to prevent neural networks from overfitting. **The journal of machine learning research**, JMLR. org, v. 15, n. 1, p. 1929–1958, 2014.

STUANI, A. S.; STUANI, A. S.; STUANI, M. B. S.; MATSUMOTO, M. A. N. As complicações do diagnóstico tardio do mesiodens: revista de literatura e relato de caso clínico. **Rev. fac. odontol. Univ. Fed. Bahia**, p. 61–7, 1999.

TAQI, A. M.; AWAD, A.; AL-AZZO, F.; MILANOVA, M. The impact of multi-optimizers and data augmentation on tensorflow convolutional neural network performance. In: IEEE. **2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)**. [S.l.], 2018. p. 140–145.

TENSORFLOW, d. **Module: tf.compat.v1.losses**. 2020. Disponível em: <https://www.tensorflow.org/api_docs/python/tf/compat/v1/losses>. Acesso em: 30 de novembro de 2020.

VOGADO, L. H.; VERAS, R. M.; ARAUJO, F. H.; SILVA, R. R.; AIRES, K. R. Rede neural convolucional para o diagnóstico de leucemia. In: SBC. **Anais Principais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde**. [S.l.], 2019. p. 46–57.

WANG, G.; TEOH, J. Y.-C.; CHOI, K.-S. Diagnosis of prostate cancer in a chinese population by using machine learning methods. In: IEEE. **2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**. [S.l.], 2018. p. 1–4.