**PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO**

**Analise oncológica de mamografias por machine learning**

**Guilherme Freitas de Araujo**

**Proposta Projeto Final II**

**CENTRO TÉCNICO CIENTÍFICO - CTC DEPARTAMENTO DE INFORMÁTICA**

Curso de Graduação em Engenharia da Computação

Rio de Janeiro, Junho de 2019



**Guilherme Freitas de Araujo**

**Analise oncológica de mamografias por machine learning**

Proposta de Projeto de Conclusão de Curso, apresentado ao curso de **Engenharia de Computação** da PUC-Rio como requisito parcial para a obtenção do título de Engenheiro de Computação.

**Orientadora**:

Prof. Marley Maria B. R. Vellasco

**Coorientador**:

Prof. Italo de Oliveira Matias

Rio de Janeiro  
Junho de 2019

**Agradecimentos**

À minha orientadora Marley Vellasco pelos ensinamentos, apoio e interesse ao longo do desenvolvimento desse projeto.

Ao meu co-orientador Italo Matias pela ajuda, acompanhamento e auxilio na busca de soluções, material e recursos necessários para o desenvolvimento do projeto.

Aos meus amigos e familiares que me apoiaram durante toda minha graduação.

**Resumo**

Araujo, Guilherme; Vellasco, Marley; Matias, Italo. Analise oncológica de mamografias por machine learning. Rio de Janeiro, 2019. [ALTERAR]10p. Proposta de Projeto de Conclusão de Curso - Departamento de Informática. Pontifícia Universidade Católica do Rio de Janeiro.

Este trabalho tem como objetivo desenvolver uma aplicação utilizando machine learning para análise de tumores em imagens mamograficas.

**Palavras-chave**

Machine learning; Inteligência artificial; Aprendizagem profunda;

**Abstract**

Araujo, Guilherme; Vellasco, Marley; Matias, Italo. Analysis of mammograms by machine learning. Rio de Janeiro, 2019. [ALTERAR]10p. Capstone Project Report – Department of Informatics. Pontificial Catholic University of Rio de Janeiro.

This Project aims the development of an application using machine learning for the analysis of tumors in mammographic images.

**Keywords**

Machine Learning; Artificial Intelligence; Deep Learning;

**Sumário**

[1 Introdução 7](#_Toc12219256)

[2 Situação Atual 8](#_Toc12219257)

[3 Propostas e Objetivos do trabalho 9](#_Toc12219258)

[4 Atividades Realizadas 10](#_Toc12219259)

[4.1 Estudos preliminares 10](#_Toc12219260)

[4.2 Estudos conceituais e de tecnologia 10](#_Toc12219261)

[4.3 Estudos conceituais e de tecnologia 12](#_Toc12219262)

[5 Referências 13](#_Toc12219263)

# Introdução

Câncer de mama é o segundo tipo de câncer com mais ocorrência em mulheres nos Estados Unidos da América (EUA) [1] e estima-se que por volta de 268,000 novos casos serão diagnosticados no ano de 2018 nos EUA e que mais ou menos 41,000 pessoas morrerão devido a essa patologia nessa mesma época [1]. A análise clínica de câncer de mama normalmente envolve o uso de diagnóstico por imagem e a mamografia é umas das técnicas utilizadas na detecção precoce do câncer de mama [2]. Mamografia é um exame de rastreio por imagem utilizando raios-X que permite visualmente analisar o tecido mamário [2]. Recomenda-se que mulheres acima dos 40 anos, ou consideradas dentro do grupo de alto risco de câncer mamário, façam exames de mamografia anualmente, pois a detecção precoce desse câncer permite um tratamento eficiente para cura do mesmo [3]. Diferentes características de tumores, tecidos e o erro humano podem causar diagnósticos errados dentro da área de oncologia [4].

O objetivo desse projeto é desenvolver uma aplicação para análise de câncer de mama em mulheres. Essa ferramenta utilizará redes convolucionais com aprendizagem profunda para classificar uma ou mais mamografias dentro da categoria de benigno ou maligno.

O projeto foi desenvolvido para Desktop (PC) para Windows porem pode ser facilmente portado para Linux, foi desenvolvido em Python e a principal biblioteca utilizada foi o Tensorflow da Google[ALTERAR]. A aplicação foi desenvolvida a partir de uma aplicação já existente que utiliza redes convolucionais para classificação de câncer de mama a partir de imagens histológicas[ALTERAR], a ideia foi adaptar tal aplicação para análise de mamografias.

O projeto envolveu muito estudo de tecnologias não profundamente vistas durante a graduação como machine learning, redes convolucionais e processamento digital de imagens, além do estudo multidisciplinar biológico para entendimento do domínio de mamografias e câncer de mama.

Seria muito interessante a continuidade desse trabalho para o desenvolvimento de uma ferramenta robusta que utilize o estudo e a aplicação desenvolvida para auxiliar os profissionais de medicina na análise clínica de exames de mama. Com uma interface de fácil uso e aperfeiçoamento da aplicação em questão.

# 2 Situação Atual

Não existe nenhum exame ou grupo de exames que possa assegurar que a mulher não tenha câncer de mama, os exames clínicos avaliam características teciduais diferentes que serão utilizadas pelo profissional da área para interpretação [2]. Há diversos componentes passivos de análise pelo profissional durante um exame de imagem da mama, são eles: nódulos, calcificações, distorção arquitetural, assimetria, linfonodo intramamário, lesão de pele e ducto único dilatado [2]. A análise de todos esses componentes segue uma orientação complexa e resulta em uma categorização baseada em probabilidade com diferentes condutas que podem causar estresse desnecessário para o paciente ou a prorrogação de um tratamento que já deveria ter sido realizado. Estima-se que uma a cada cinco mamografias resultam em falsos negativos [1] e que 7% a 9% de mulheres que realizam mamografias anualmente receberam um resultado falso-positivo recomendando a etapa de biopsia [5].

Categorização de imagens por meio de redes convolucionais é um processo que tem evoluído muito ao longo dos anos e possui diversas ferramentas para facilitar o desenvolvimento de soluções robustas e assertivas relacionadas a diferentes áreas. O estudo que obteve a menor taxa de erro no desafio de classificação que utiliza o CIPHAR-10 ([Canadian Institute For Advanced Research](https://en.wikipedia.org/wiki/Canadian_Institute_for_Advanced_Research)) foi o AutoAugment [6], que usou de técnicas de transformações das imagens da base de dados para aumentar a precisão de seu algoritmo.

Diversos estudos são feitos relacionando biologia e medicina com inteligência artificial, pois questões dentro dessa área comumente contém informações complexas e numerosas demais para uma análise humana ou baseada em software comum.

A utilização de ferramentas de auxílio ao profissional de medicina pode ter um impacto enorme em custos relacionados à saúde evitando tratamentos desnecessários além de proporcionar um tratamento mais rápido e assertivo ao paciente. Esse projeto estuda e desenvolve uma aplicação que poderá ser utilizada como base do desenvolvimento de tal ferramenta.

# Propostas e Objetivos do trabalho

O projeto tem como objetivo desenvolver uma aplicação que classifique diferentes imagens oriundas de mamografias no formato DICOM (Digital Imaging and Communications in Medicine) em benignas e malignas.

Será utilizada uma rede convolucional, ideal para classificação de imagens, pois dispensa boa parte do pré-processamento da base de dados, devido à inteligência do algoritmo de aprender quais filtros devem ser aplicados a imagem antes de passar para a camada de rede neural tradicional para classificação.

Algumas arquiteturas diferentes serão estudadas para se decidir qual deve ser abordada no projeto, utilizando benchmarks de classificação de imagens como CIPHAR-10, CIPHAR-100 e ILSVRC (ImageNet Large Scale Visual Recognition Challenge). Diferentes configurações dos parâmetros da rede e da base de dados utilizada para treinamento serão testadas visando aperfeiçoar a classificação resultante da rede. Dependendo do tempo necessário para se computar o treinamento da rede desenvolvido poderá ser interessante utilizar múltiplas GPUs (Graphic Processing Units) e/ou um cluster para acelerar tal etapa do projeto.

Deseja-se ter ao final do projeto uma aplicação que possa servir como base para o desenvolvimento de uma ferramenta para uso clinico real auxiliando profissionais de medicina no dia-a-dia, e ter uma rede robusta que sirva como fonte de estudo para futuro desenvolvimento relacionando oncologia e processamento de imagens com inteligência artificial.

# Atividades Realizadas

A primeira fase do projeto foi de estudo teórico da área de câncer de mama e de redes convolucionais para machine learning com imagens. A segunda fase envolveu a decisão de que linguagens, ferramentas e bibliotecas seriam utilizadas no desenvolvimento do projeto. A terceira fase envolveu testes, adaptação do projeto. A quarta e última fase envolveu a escrita deste relatório, documentando todo o projeto e suas conclusões.

# Estudos preliminares

O conhecimento e a experiência do graduando em questão em relação ao tema de machine learning, inteligência artificial e processamento digital de imagens foram obtidos de duas formas antes do desenvolvimento desse projeto.

A primeira foi durante o curso da matéria ENG1456 – Inteligência Computacional Aplicada, cuja professora é a orientadora desse projeto. Durante o desenvolvimento da disciplina o aluno teve seu primeiro contato com inteligência artificial e redes neurais, criando uma base de entendimento para idealização desse projeto.

O segundo momento de contato com inteligência artificial e processamento digital de imagens foi durante os 8 meses de iniciação cientifica realizada com o Italo Matias, co-orientador desse projeto, no laboratório de Bio-Design da PUC-Rio. Durante a iniciação foi realizado um projeto de segmentação de imagens medicas para análise de tumores pulmonares. Nesse tempo houve um acumulo de conhecimento muito grande relacionado a processamento digital de imagens.

# Estudos conceituais e de tecnologia

O início do projeto envolveu o estudo teórico da área de câncer de mama e mamografias, para isso foi utilizado o Atlas BI-RADS[2] que tem como objetivo definir, padronizar o processo de análise de imagens de mama, para melhorar a qualidade e segurança das análises clinicas.

Mamografia é uma imagem raio-x da mama, o exame para se obter tal imagem envolve colocar o peito da paciente entre duas superfícies firmes, logo em seguida as superfícies apertam o seio deixando-o mais reto o que facilita a obtenção da imagem. Após o exame as imagens são analisadas por um especialista na procura dos seguintes achados:

* Nódulos: Também chamados de massas são volumes tridimensionais classificados de acordo com sua forma, margem e densidade. A alta densidade do nódulo esta significativamente associada a malignidade do mesmo.
* Calcificações: Classificadas de acordo com a forma, tamanho e distribuição. Calcificações benignas são normalmente maiores enquanto as malignas tendem a ser muito pequenas.
* Distorção arquitetural: Diz respeito a análise das linhas do tecido da mama, podendo estar relacionada a malignidade.
* Assimetria: Significa a assimetria das fibras do tecido, pode indicar lesões e malignidade.
* Linfonodo intramamário: São nódulos circunscritos tipicamente benignos.
* Lesão na pele: Lesão que se projeta sobre a mama, tipicamente benigna.
* Ducto único dilatado: Estrutura tubular normalmente rara e benigna.

Todos esses achados devem ser analisados objetivamente seguindo o sistema de laudos, que envolve uma breve descrição da indicação do exame, descrição sucinta da composição geral da mama, descrição objetiva de quaisquer achados importantes, comparação com exames anteriores caso o radiologista julgue necessário, avaliação de acordo com a escala BI-RADS, e a conduta que deve ser tomada.

A classificação do BI-RADS é dividida em seis categorias, sendo a quarta subdividida em três subcategorias, gerando um total de nove categorias cada categoria está atrelada à uma conduta e uma probabilidade de câncer. Tentando dividir essas categorias em benignas e malignas baseado na probabilidade de câncer temos cinco categorias provavelmente benignas (de 1 a 4B) benignas e duas malignas (4C e 5).



**Figura 1**: Tabela de classificação BI-RADS

As visões padrões relacionadas as mamografias são a craniocaudal (CC) e a mediolateral-oblíqua (MLO), o motivo de ter essas duas incidências são o aumento da sensibilidade das áreas cegas, diminuição dos falsos positivos.



**Figura 2**: Imagem de visão MLO [7] **Figura 3:** Imagem de visão CC [7]

Após o estudo teórico relacionado a mamografia, foi feito um estudo mais aprofundado sobre redes neurais e machine learning na classificação de imagens.

Redes neurais artificiais surgiram com o intuito de copiar certas capacidades do cérebro humano, como classificação, generalização e reconhecimento de padores, ela é composta por unidade de processamento simples denominadas neurônios ou nós[ALTERAR]. Um neurônio possui entradas, um corpo que realiza alguma operação matemática e produzem uma saída[ALTERAR].



**Figura 4**: Modelo básico de um neurônio

Um dos primeiros modelos matemáticos desenvolvidos para formular o neurônio foi o modelo do Perceptron [ALTERAR], nesse modelo cada entrada é binaria (zero ou um) e possui um peso associado a ela, também foi adicionada uma constante como entrada denominada bias e a saída passa uma função de ativação, ****que no caso do Perceptron é a função degrau.

**Figura 5**: Perceptron

A função de ativação degrau faz com que a saída da soma das entradas associadas aos seus respectivos pesos gera um valor binário (zero ou um) que baseado em um *threshold* (limite). Esse modelo limita matematicamente as entradas e a saída devido a sua restrição binaria.

Com essa limitação outros modelos matemáticos para definir neurônios surgiram, como a *Sigmoid* e a *Rectified Linear Unit* (ReLU), ambos permitindo valores reais como entradas e gerando valores reais como saída [ALTERAR <http://cs231n.github.io/neural-networks-1/>].

Esses neurônios são conectados para formar uma rede neural, essa estrutura formada pelos neurônios permite que os sinais de entrada de rede sejam processados pelos neurônios que se comunicação entre si, alterando os pesos de cada entrada adaptando os neurônios a diferentes estímulos proporcionados pelos sinais de entrada. É comum que esses neurônios sejam distribuídos em camadas, classificadas em camada de entrada, onde os padrões são apresentados a rede, camadas intermediarias ou escondidas, onde é feita a maior parte do processamento através das conexões ponderadas pelos pesos, e a camada de saída, onde o resultado final é concluído e apresentado [ALTERAR <http://cs231n.github.io/neural-networks-1/>]. Assim é possível que a rede “aprenda”, simulando o processo de aprendizado do cérebro humano.

 **Figura 6**: Camadas da rede neural

Para que esse aprendizado ocorra três métodos de treinamento foram desenvolvidos, aprendizado supervisionado, não supervisionado e por reforço. No aprendizado supervisionado existe um agente externo que indica à rede a resposta desejada para o padrão de entrada, ou seja, a saída foi pré-definida externamente à entrada. No aprendizado não supervisionado não existe um agente externo indicando a resposta desejada aos padrões de entrada. No aprendizado por reforço a rede também não possui a informação de resposta porem recebe um reforço positivo ou negativo dependendo da resposta. Denomina-se ciclo de uma aprendizagem quando todos os pares entrada e saída do conjunto de treinamento são apresentados para a rede.

*Backpropagation* é um algoritmo de aprendizado supervisionado, em que é calculado um erro após o processamento da rede do sinal de entrada até a saída, e esse erro é propagado na direção inversa da rede, ou seja, da saída para entrada, atualizando os pesos de cada neuronio de acordo com esse erro calculado [11].

As redes perceptron multicamadas (MLP) se caracteriza por ter todos seus neurônios conectados a um neurônio da próxima camada [10], a figura 6 é um exemplo de uma MLP. Essa rede utiliza do mecanismo de *backpropagation* para realizar o seu aprendizado.

Redes convolucionais [ALTERAR] <http://cs231n.github.io/convolutional-networks/> são versões adaptadas de MLP usadas principalmente para classificação e analise de imagens, essas redes recebem esse nome devido ao uso do processo de convolução, que no sentido de redes convolucionais é uma operação linear que envolve a multiplicação de um grupamento de entradas com pesos e um filtro estabelecido, gerando na saída um mapa de ativação daquele filtro, com isso a rede aprende quais filtros são ativados quando certas características são detectadas em uma posição espacial da entrada. Assim a própria rede aprende os filtros que antes eram aplicados no pré-processamento de imagens, evitando que seja necessário um conhecimento prévio das características da imagem relevantes para sua classificação [10][12].

****

**Figura 7**: Filtro K sendo convolucionado na imagem I gerando mapa de características.

Para lidar com entradas de matrizes de grandes dimensões como imagens, é impraticável conectar neurônios com todos os neurônios da camada anterior como ocorre na MLP. Redes convolucionais utilizam da correlação da informação espacial para conectar os neurônios, ou seja, cada neurônio só está conectado à uma pequena região da entrada, aumentando desempenho e separando a imagem em frames.

Além da camada de convolução temos a camada de *pooling* que tem como objetivo diminuir progressivamente o tamanho espacial da representação para reduzir a quantidade de parâmetros para serem computados pela rede, evitando o *overfitting* que é um problema causado pela grande quantidade de parâmetros que a rede deve processar, causando uma grande variância no resultado da classificação[ALTERAR]. *Pooling* é normalmente realizado através do max pooling que obtém o maior valor de uma certa região da matriz para propagar. Essa camada de pooling normalmente é inserida logo após uma camada de convolução.

**Figura 8**: Processo de pooling



Outras camadas que podem ser adicionadas a rede são de ReLU que aplica a função de ativação removendo valores negativos do mapa de ativação, transformando-os em zeros, com isso os aspectos de decisão e não-linearidade são melhorados. E a camada de perda (*loss layer*) que determina como o treino penaliza a variação entre o que foi predito e o valor real, é comum serem utilizadas funções de *softmax* e *sigmoid cross-entropy* nessa camada.

**Figura 9**: Exemplo de arquitetura de rede convolucional

# Testes e prototipos

Python foi linguagem escolhida para o desenvolvimento da rede neural utilizada, pelo fato de ser uma linguagem que possui uma grande quantidade de ferramentas relacionadas a processamento de imagens e redes neurais. A principal biblioteca utilizada foi o TensorFlow, um framework para machine learning open source criado pela Google.

# Referências

1. SPIEGEL, Rebecca L.; MILLER, Kimberly D.; JEMAL, Ahmedin. **Cancer Statistics, 2018.** Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.3322/caac.21442>. Acesso em: 20 jun. 2018.
2. D’ORSI, C.J.; SICKLES, E.A.; MENDELSON, E.B.; MORRIS, E.A. ACR **BI-RADS® Atlas, Breast Imaging Reporting and Data System**. Reston, VA, American College of Radiology; 2013.
3. [OEFFINGER, K.C.; FONTHAM, E.T.; ETZIONI R. **Breast cancer screening for women at average risk: 2015 guideline update From the American Cancer Society**. *JAMA*. 2015.](https://www.breastcancer.org/symptoms) Disponível em: <<https://onlinelibrary.wiley.com/doi/full/10.3322/caac.21442>>. Acesso em: 20 jun. 2018.
4. [GIESS, Catherine; FROST, Elisabeth; BIRDWELL, Robyn. **Difficulties and errors in diagnosis of breast neoplasms**. 2012.](https://www.breastcancer.org/symptoms) Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/22824119>. Acesso em: 20 jun. 2018.
5. HUBBARD R.A; KERLIKOWSKI, K.; FLOWERS, C.I.; YANKASKAS, B.C.; ZHU, W.; MIGLIORETTI, D.L. **Cumulative Probability of False-Positive Recall or Biopsy Recommendation After 10 Years of Screening Mammography: A Cohort Study**. Ann Intern Med. 2011. doi: 10.7326/0003-4819-155-8-201110180-00004. Disponível em: <http://annals.org/aim/article-abstract/474984/cumulative-probability-false-positive-recall-biopsy-recommendation-after-10-years?doi=10.7326%2f0003-4819-155-8-201110180-00004>. Acesso em: 20 jun. 2018.
6. [CUBUK, Ekin D.](http://adsabs.harvard.edu/cgi-bin/author_form?author=Cubuk,+E&fullauthor=Cubuk,%20Ekin%20D.&charset=UTF-8&db_key=PRE); [ZOPH, Barret](http://adsabs.harvard.edu/cgi-bin/author_form?author=Zoph,+B&fullauthor=Zoph,%20Barret&charset=UTF-8&db_key=PRE); [MANE, Dandelion](http://adsabs.harvard.edu/cgi-bin/author_form?author=Mane,+D&fullauthor=Mane,%20Dandelion&charset=UTF-8&db_key=PRE); [VASUDEVAN, Vijay](http://adsabs.harvard.edu/cgi-bin/author_form?author=Vasudevan,+V&fullauthor=Vasudevan,%20Vijay&charset=UTF-8&db_key=PRE); [LE, Quoc V.](http://adsabs.harvard.edu/cgi-bin/author_form?author=Le,+Q&fullauthor=Le,%20Quoc%20V.&charset=UTF-8&db_key=PRE) **AutoAugment: Learning Augmentation Policies from Data.** eprint arXiv:1805.09501. 2018. Disponível em: <https://arxiv.org/abs/1805.09501>. Acesso em: 20 jun. 2018.
7. Imagens de caso do A.Prof Frank Gaillard, **Radiopedia.org**, rID: 12608
8. ROSENBLATT, Frank (1958), **The Perceptron--a perceiving and recognizing automaton**. Report 85-460-1, Cornell Aeronautical Laboratory.
9. MCCULLOCH, Warren; PITTS, Walter (1943), **A Logical Calculus of the ideas immanent in nervous activity**, Bulletin of Mathematical Biology, Vol. 52, No. 1/2.
10. NIELSEN, Michael A. **“Neural Networks and Deep Learning”**, Determination Press, 2015.
11. SATHYANARAYANA, Shashi. **“A Gentle Introduction to Backpropagation”**. July 22, 2014.
12. CIRESAN, Dan C., MEIER Ueli, MASCI Jonathan, GAMBARDELLA Luca M. SCHMIDHUBER Jurgen, “**Flexible, High Performance Convolutional Neural Networks for Image Classification**”