APLICAÇÃO DE TÉCNICAS DE MACHINE LEARNING NA DETECÇÃO DO CÂNCER DE PULMÃO

APPLICATION OF MACHINE LEARNING TECHNIQUES IN LUNG CANCER DETECTION

Jonas Natan Manoel Messias Soares
ORCID: 0000-0001-7232-7449
Centro Universitário Adventista de São Paulo – Hortolândia/SP
jonas.natan@hotmail.com

Vítor Gutierrez Trindade
ORCID: 0000-0003-0233-3725
Centro Universitário Adventista de São Paulo – Hortolândia/SP
vitorgutierrez@live.com

Orientador: Mestre Ackley Dias Will
ORCID: 0000-0001-7828-1007
Centro Universitário Adventista de São Paulo – Hortolândia/SP
ackleywill@gmail.com

Co-orientador: Doutor Bassiro Só
ORCID: 0000-0002-7661-6923
Centro Universitário Adventista de São Paulo – Hortolândia/SP
bassiroso@gmail.com

RESUMO: O câncer foi responsável por cerca de 9.6 milhões de mortes no mundo no ano de 2018. O câncer de pulmão é o mais comum, com 2.09 milhões de casos no mesmo ano. No Brasil, o câncer de pulmão é o segundo mais comum entre os homens e o quinto entre as mulheres. Devido a isso, nos últimos anos houve um crescimento expressivo nas pesquisas em relação a doença. Novos métodos passaram a ser desenvolvidos para que a doença fosse descoberta nos seus estágios iniciais, antes que apresentassem sintomas, possibilitando assim tratamentos antecipados e mais eficientes. Com os avanços tecnológicos na área da medicina, grandes quantidades de dados têm sido coletadas e disponibilizadas para as pesquisas médicas. A questão é como podemos aplicar a tecnologia de modo a auxiliar os médicos a obterem resultados mais rápidos e precisos? Nos últimos anos o uso do machine learning tem se mostrou bastante eficaz e trazendo bons resultados em predições diagnósticas de doenças como o próprio câncer. Diante disso o objetivo desse estudo é analisar como a aplicação dessas técnicas podem auxiliar na detecção do câncer de pulmão. Através da visão computacional e de redes neurais convolucionais, propriamente tratando-se da arquitetura U-Net, desenvolveu-se um modelo escrito em Python que utiliza uma biblioteca de rede neural conhecida como Keras, capaz de receber as imagens de tomografias, previamente preparadas, possibilitando a aplicação de uma sequência de etapas de convoluções convertendo essas imagens em matrizes computacionais para análise e segmentação dos nódulos nas imagens. Criando assim um modelo computacional com uma acurácia de cerca de 70% o que é considerada extremamente baixa tratando-se de diagnósticos médicos. Isso acontece devido a duas grandes dificuldades encontradas na abordagem desse tipo assunto, a primeira se dá com a dificuldade de se obter as imagens das tomografias pulmonares que posam ser utilizadas

no treinamento do modelo. A segunda limitação envolve os requisitos computacionais necessários para se trabalhar com essas imagens, sua preparação, treinamento e criação do modelo, já que todas estas etapas exigem grande poder computacional. Com um bom conjunto de imagens, uma escolha correta de heurística e uma boa capacidade computacional pode-se criar modelos com precisão muito maior e que realmente possam fazer a diferença na entrega de algum diagnóstico, além da possibilidade de estudos de outras arquiteturas que possam contribuir com melhores resultados.

ABSTRACT: Cancer accounted for about 9.6 million deaths worldwide in 2018. Lung cancer is the most common, with 2.09 million cases in the same year. In Brazil, lung cancer is the second most common among men and the fifth among women. Because of this, in recent years there has been a significant growth in research regarding the disease. New methods were developed so that the disease was discovered in its early stages, before they presented symptoms, thus enabling early and more efficient treatments. With technological advances in the field of medicine, large amounts of data have been collected and made available for medical research. The question is how can we apply technology to help doctors get faster and more accurate results? In recent years the use of machine learning has been shown to be very effective and bringing good results in diagnostic predictions of diseases such as cancer itself. Therefore, the aim of this study is to analyze how the application of these techniques can help in the detection of lung cancer. Through computational vision and convolutional neural networks, specifically the U-Net architecture, a model written in Python was developed that uses a neural network library known as Keras, capable of receiving previously prepared tomography images. enabling the application of a sequence of convolution steps converting these images into computational matrices for analysis and segmentation of nodules in the images. So, creating a computational model with an accuracy of about 70% which is considered extremely low in the case of medical diagnostics. This is due to two major difficulties encountered in addressing this type of issue, the first being the difficulty in obtaining the images of pulmonary tomography's that can be used in the training of the model. The second limitation involves the computational requirements necessary to work with these images, their preparation, training and model creation, since all these steps require great computational power. With a good set of images, a correct choice of heuristics and a good computational capacity one can create models with much higher precision that can really make a difference in the delivery of some diagnosis, besides the possibility of studies of other architectures that could contribute. with better results.

PALAVRAS-CHAVE: Redes neurais convolucionais; tecnologia; predições; U-Net;

KEYWORDS: Neural networks; technology; prediction; convolutional;

1 INTRODUÇÃO

De acordo com a Organização Mundial da Saúde (OMS) o câncer foi responsável por cerca de 9.6 milhões de mortes no mundo no ano de 2018. O câncer de pulmão é o mais comum, com 2.09 milhões de casos no mesmo ano (GLOBAL CANCER OBSERVATORY, 2019). Devido a isso, nos últimos anos houve um crescimento expressivo

nas pesquisas em relação a doença. Novos métodos passaram a ser desenvolvidos para que a doença fosse descoberta nos seus estágios iniciais, antes que apresentassem sintomas, possibilitando assim tratamentos antecipados e mais eficientes (PAINS, 2018).

Apenas no Brasil o câncer causa um grande impacto todos os anos, são mais de 225 mil mortes no Brasil a cada ano sem contar no prejuízo econômico que acaba sendo gerado. Segundo a BBC, estima-se que que o país sofra um prejuízo de cerca de R$ 15 bilhões por ano o que corresponde a 0,21% de toda a riqueza gerada. No Brasil a maior parte das mortes são ocasionadas pelo câncer de pulmão equivalente a R$ 1,3 bilhão de reais ao ano. Em média, cada vida perdida por câncer no Brasil na população economicamente ativa gera uma perda de R$ 176 mil (WENTZEL, 2018).

Diante disso o objetivo desse estudo é analisar como a aplicação de técnicas de classificação, que é uma sub-área de machine learning (aprendizado de máquina), podem auxiliar na detecção do câncer de pulmão. Através dessa técnica, é realizado um treinamento através de características e classificações. Após o treinamento é então possível realizar a classificação de dados de entrada ainda não conhecidos (SILVEIRA; BULLOCK, 2017) (GÉRON, 2019).

Através do estudo de algoritmos e técnicas de aprendizado de máquina que tem a capacidade de analisar imagens médicas de tomografias computadorizadas e assim prover uma probabilidade de existência de um câncer em um paciente. O trabalho visa incentivar a procura por novos meios tecnológicos, que sejam acessíveis e que possam auxiliar não só no diagnóstico de doenças como o câncer, mas que possam ser aplicadas em outras áreas da saúde e promovam melhorias para os pacientes e para os médicos.

2 METODOLOGIA

No presente trabalho se dará o estudo e a análise de algumas técnicas e algoritmos de aprendizado de máquina para detecção do câncer de pulmão. Atualmente, este tipo de câncer é diagnosticado através da inspeção médica nas imagens de tomografias computadorizadas de um paciente, procurando pequenas bolhas nos pulmões chamadas nódulos. Um nódulo por si só não representa o câncer, mas possui algumas características que podem ajudar nesse diagnóstico.

Através de técnicas de visão computacional, machine learning e deep learning, pode-se criar um modelo capaz de receber as imagens de tomografias, previamente preparadas,

e dizer o percentual de chance de o paciente da imagem possuir ou não um nódulo cancerígeno.

Por se tratar de técnicas matemáticas extremamente complexas utilizou-se da linguem de programação Python e de algumas bibliotecas que facilitam e comportam muitas das funções necessárias para o processo de desenvolvimento do modelo computacional. Dentre elas: o numpy, pacote fundamental para a computação científica, para recursos de álgebra e trabalho com matrizes; o scikit-learn, para análise de dados; o Keras, biblioteca para se trabalhar com criações de redes neurais de forma simplificada e mais rápida e a própria arquitetura U-Net, para auxilio com a segmentação dos nódulos da imagens radiográficas dos pulmões.

Para criação desse modelo faz-se o uso um dataset ou banco de dados imagens, para o presente estudo utilizou-se o LUNA16. Um dataset que foi cedido publicamente para uso em um desafio de 2016, o Lung Nodule Analysis, o qual tinha o objetivo semelhante ao estudo do trabalho. O The Lung Image Database Consortium Image Collection (LIDC-IDRI) é um banco de dados de imagem público, disponível através do Cancer Imaging Archive, o qual forma a base de imagens do LUNA16.

O dataset contém 888 tomografias computadorizadas de tamanho 512 x 512, de onde selecionou-se apenas algumas imagens para treinamento do modelo já que, para o processamento de uma grande quantidade de informação como essa faz-se necessário uma grande capacidade computacional.

O processo inicia-se com a preparação das imagens para o treinamento do modelo, onde elas precisam ser digitalizadas e classificadas para assim poderem ser usadas para treinar o modelo. A tarefa de classificação sempre que dará a base para o treinamento do modelo deve ser realizada por alguém da área de negócio que tenha o devido conhecimento. O dataset selecionado já contém imagens preparadas e classificadas, disponibilizadas publicamente na internet, o LUNA 2016 (BRAM; COLIN, 2016). Além da radiografia em si, o conjunto com localizações de nódulos correspondentes anotadas por quatro radiologistas.

Para detectar os nódulos nas imagens utilizou-se de uma arquitetura conhecida como U-Net. Por se tratar de imagens médicas, o processo de segmentação acaba sendo um pouco mais complexo, pois dificilmente essas imagens irão possuir características lineares simples. Devido a essa dificuldade, vários algoritmos e diversas técnicas vêm sendo aplicadas e desenvolvidas como modelos de Deep Learning, onde usa-se uma rede neural com várias camadas. É aí que a U-Net entra em cena. Essa arquitetura foi

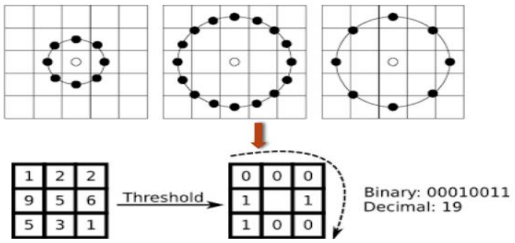
inicialmente projetada especialmente para a segmentação de imagens médicas e devido seu sucesso, começou a ser usado em outras áreas também (SANKESARA, 2018).

A arquitetura aplica várias etapas de convolução, ou seja, matrizes numéricas na imagem que funcionam como um filtro (ABDULKADIR, 2016). Desse modo extrai-se algumas características da imagem, como o contorno do pulmão e dos nódulos de uma forma que o algoritmo compreenda essas informações gerando uma saída.

Contudo, ainda se tem vários retornos de falsos nódulos, chamados de falsos positivos. Para reduzir o número de falsos positivos retornados pela U-Net, é necessário a aplicação da seguinte regra: para qualquer nódulo em qualquer fatia, se não for detectado um nódulo próximo na fatia na vizinha daquele nódulo, ele é descartado. A ideia é que os nódulos se estendam por mais de uma fatia, um nódulo que aparece em apenas uma fatia, é assumido como um falso positivo. Extrai-se então fatias de 48 x 48 em torno dos nódulos candidatos e aplica-se a análise entre as fatias vizinhas.

Estas fatias precisam ser transformadas em vetores de recursos, que representam nas nossas imagens os possíveis nódulos. Para isso aplica-se a técnica de Padrões Binários Locais ou Local Binary Patterns (LBP), usado para capturar o conteúdo textural das imagens. Cada pixel é um código binário baseado em seu valor em relação aos pixels vizinhos. Isso é feito atribuindo um valor de 1 aos pixels que possuem vizinhos com valores maiores ou iguais ao valor do pixel sendo considerado e um valor de zero para aqueles com valores mais baixos. Isto resulta em um código binário de n bits que descreve cada pixel, onde n é o número de vizinhos.

Figura 1 – Exemplo dos Padrões Binários Locais

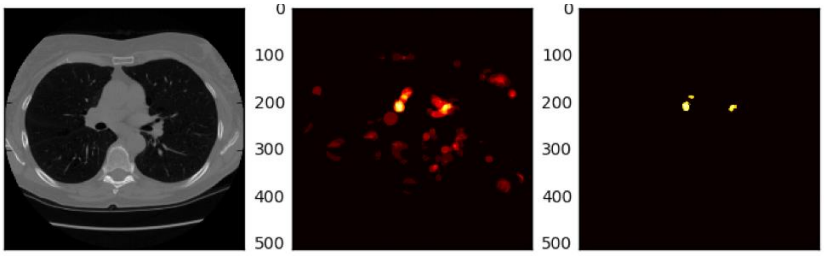


Fonte: Junior (2014).

Após a aplicação do LBP mantem-se apenas os nódulos detectados que foram encontrados em regiões onde ocorreu várias detecções em todas as fatias. Essa abordagem gera um mapa de calor para combinar as detecções da U-Net em todas as fatias de um único paciente, conforme apresentado na Figura 2, todas as regiões dos

nódulos detectados para um paciente com as regiões que são mais vermelhas, indicando mais detecções nesses locais de pixel.

Figura 2 – Mapa de calor



Fonte: LUNA16 - Bram, J; Colin, J (2016).

Nota-se que existe a detecções de falso positivos em vários locais da tomografia computadorizada, mas apenas algumas regiões temos maior intensidade. O mapa de calor é limitado para manter apenas os pixels que possuem repetidas detecções. Assim, as regiões que tiverem mais detecções da segmentação do nódulo têm maiores chances de serem nódulos reais. O Dice Coefficient (DC) é a métrica usada para avaliar quanto um nódulo segmentado se sobrepõe a um dos nódulos do mapa de calor, capturando assim a média de pixels que aparecem com a maior frequência. Após isso o algoritmo estará apto para identificar os possíveis nódulos na tomografia do paciente.

Assim pode-se preparar os dados para o treinamento, aplicando uma normalização, para que os algoritmos os recebam com uma distribuição normal. Existem várias formas de fazer essa etapa, mas a maneira mais comum é aplicando algoritmo de score Z, que consiste basicamente em padronizar minhas variáveis antes de usá-las no meu modelo, para que o efeito de qualquer variável aconteça na mesma escala que nas outras.

Para iniciar o treinamento do modelo, os dados pré-processados na etapa anterior são imputados, juntamente com as informações dos pacientes que acompanham o dataset. Desse modo, para treinarmos o modelo, separamos o conjunto de treinamento em dois novos conjuntos de imagens, cada um com 10 imagens. O primeiro contendo imagens com nódulos e o segundo conjunto de imagens sem os nódulos. Criamos assim uma sequência de treino que intercala as imagens dos dois conjuntos. A arquitetura utilizada ao receber essas imagens as converte em matrizes numéricas aplicando uma sequência de convoluções que isolam e segmentam os possíveis nódulos.

Ao finalizar o treinamento, o modelo pode ser salvo e usado para receber novas imagens diferentes das utilizadas no treino.

3 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Como resultado tem-se um modelo capaz de detectar um nódulo em uma tomografia computacional no padrão de 512 x 512 com uma precisão de quase 70%, o que para utilização médica ainda é baixa. Muitas outras pesquisas em relação a quais tipos de algoritmos se aplicam melhor na detecção de cada tipo de câncer, com resultados impressionantes, podem talvez serem aplicados de modo que em combinação acabem se completando (KOUROUA, 2014).

O modelo, a partir do momento que é alimentado com uma quantidade de dados significante e se tem um poder de processamento capaz de suportar esses dados, pode ser aprimorado de tal modo que traga resultados altamente precisos e de fato auxiliem os médicos em seus diagnósticos.

Mesmo com os avanços tecnológicos tanto em relação a própria tecnologia quanto os métodos que vêm sendo estudados e utilizados, o machine learning pode ser usado para melhorar substancialmente (15-25%) a precisão de prever a suscetibilidade ao câncer, a recorrência e a mortalidade (CRUZ, 2006).

4 CONSIDERAÇÕES FINAIS

O desenvolvimento do presente trabalho possibilitou a análise de como a utilização de algoritmos e técnicas de aprendizado de máquina podem contribuir das mais diversas formas, dentre elas auxiliando no diagnóstico do câncer. Através da criação de modelos baseados em arquiteturas de redes neurais convolucionais como a U-Net, a qual foi estudada no artigo, tem-se a possibilidade de desenvolver modelos altamente precisos que venham contribuir significativamente nas mais diversas áreas da saúde, podendo também serem utilizadas para outras áreas de estudo, o que torna a aplicação das técnicas ilimitadas.

Para estudos futuros propõe-se a análise de outros tipos de algoritmos e técnicas que possam ser aplicadas em conjunto de maneira que se obtenha resultados melhores. Como também o estudo de meios integrar esse tipo de informação entre as comunidades médicas para um melhor aproveitamento e contribuição das partes envolvidas, facilitando a obtenção de dados e proporcionando um modelo de machine learning mais preciso e capaz.

REFERÊNCIAS

ABDULKADIR, A.; BROX, T.; ÇIÇE, O.; LIENKAMP, S.; RONNEBERGER, O. **3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation.** Germany, 2016. Disponível em: <https://lmb.informatik.uni-freiburg.de/Publications/2016/CABR16/cicek16miccai.pdf>. Acesso em diversas datas entre agosto e novembro de 2019.

BRAM, J; COLIN, J. **LUNA16 - LUng Nodule Analysis 2016.** Disponível em: <https://luna16.grand-challenge.org//>. Acesso em diversas datas entre agosto e novembro de 2019.

CRUZ, J.; WISHART, D. **Applications of Machine Learning in Cancer Prediction and Prognosis.** Canadá: University of Alberta Edmonton, 2006. Disponível em: < http://journals.sagepub.com/doi/pdf/10.1177/117693510600200030>. Acesso em diversas datas entre agosto e novembro de 2019.

GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. Rio de Janeiro: Alta Books, 2019.

GLOBAL CANCER OBSERVATORY. **Cancer Today.** IARC, 150 Cours Albert Thomas, France, 2019. Disponível em: < http://gco.iarc.fr/>. Acesso em diversas datas entre agosto e novembro de 2019.

JUNIOR, G. **Introdução a Local Binary Patterns (LBP).** Universidade Federal do Maranhão, 2014. Disponível em: < http://nca.ufma.br/~geraldo/vc/14.b_lbp.pdf>. Acesso em diversas datas entre agosto e novembro de 2019.

KOUROUA K.; EXARCHOSA, T.; EXARCHOSA K.; KARAMOUZIS M.; FOTIADISA, D. **Machine learning applications in cancer prognosis and prediction.** Athens: University of Athen, 2014. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2001037014000464>. Acesso em diversas datas entre agosto e novembro de 2019.

PAINS, C. **Câncer é a principal causa de morte em quase 10% das cidades brasileiras.** São Paulo: O Globo, 2018. Disponível em: < https://oglobo.globo.com/sociedade/saude/cancer-a-principal-causa-de-morte-em-quase-10-das-cidades-brasileiras-22595871>. Acesso em diversas datas entre agosto e novembro de 2019.

SANKESARA, H. **UNet Introducing Symmetry in Segmentation.** Towards Data Science, 2018. Disponível em: < https://towardsdatascience.com/u-net-b229b32b4a71/>. Acesso em diversas datas entre agosto e novembro de 2019.

SILVEIRA, Guilherme; BULLOCK, Bennett. **Machine Learning: Introdução à classificação**. São Paulo: Casa do Livro, 2017.

WENTZEL, M. **Quanto o câncer custa à economia do Brasil?** BCC News Brasil, 2018. Disponível em: <https://www.bbc.com/portuguese/geral-43047430>. Acesso em diversas datas entre outubro e novembro de 2019.