



Development of a Machine Learning based model for early screening for oral cancer

Desenvolvimento de um modelo baseado em Machine Learning para rastreamento precoce do cancer de boca

SILVA, Ivisson Alexandre Pereira da⁽¹⁾; OLIVEIRA, Catarina Rodrigues Rosa de⁽²⁾;
OLIVEIRA, José Marcos dos Santos⁽³⁾; FILHO, Carlos Alberto Correia Lessa⁽⁴⁾;
FERREIRA, Sonia Maria Soares⁽⁵⁾

⁽¹⁾ 0000-0002-1682-3648; Student, researcher and master's student of the Professional Master's Degree in Health Research, in CESMAC University Center (*Centro Universitário CESMAC*), Brazil. Email: iapereira29@gmail.com.

⁽²⁾ 0000-0001-9178-8902; Teacher, Master in Stomatology and Radiology by São Leopoldo Mandic, teacher in the discipline of Clinical Propedeutics in CESMAC University Center (*Centro Universitário CESMAC*), Brazil. Email: catarinarosaodonto@hotmail.com.

⁽³⁾ 0000-0002-4618-1500; Teacher, PhD in Biochemistry and Biotechnology in CESMAC University Center (*Centro Universitário CESMAC*) and Federal University of Alagoas (*Universidade Federal de Alagoas - UFAL*), teacher at the Institute of Chemistry and Biotechnology at UFAL and Pharmacy at CESMAC, Brazil. Email: jose_marcos_cbjr@hotmail.com.

⁽⁴⁾ 0000-0002-4114-1235; Teacher, researcher, specialist in software development and teacher of Artificial Intelligence discipline in CESMAC University Center (*Centro Universitário CESMAC*), Brazil. Email: carlos.filho@cesmac.edu.br.

⁽⁵⁾ 0000-0002-4825-171x; Teacher, researcher, coordinator of the Professional Master's Degree in Health Research, in CESMAC University Center (*Centro Universitário CESMAC*), Brazil. Email: sonia.ferreira@cesmac.edu.br.

The content expressed in this article is the sole responsibility of its authors.

ABSTRACT

Oral Squamous Cell Carcinoma (OSCC) is the most frequent type of oral cancer, accounting for about 40% of malignant head and neck lesions. It's known that the favorable prognosis is associated with early diagnosis, since the survival rate increases as a function of the diagnosis in the early stages of the disease. Thus, the objective of this work was to implement and train a Machine Learning model that can help in the diagnosis of oral cancer. Through technologies such as artificial intelligence (AI) that can use images in their analyses, it's sought to improve the prognosis of oral cancer through its early detection. Using the branch of AI, Machine Learning and its subgroup Deep Learning, it becomes possible through Convolutional Neural Network (CNN) to perform an image screening of malignant and premalignant lesions, in order to identify the presence or not of oral cancer. The RNC structure is based on the MobileNet structure, which separates the images into fragments and after training, showed the identification of cancer in 91% of the images examined and of Leukoplakia in 84% of the analyzed images.

RESUME

O carcinoma espinocelular da cavidade bucal (CECCB) é o tipo de câncer de boca mais frequente, representando cerca de 40% das lesões malignas de cabeça e pescoço. Sabe-se que o prognóstico favorável está associado ao diagnóstico precoce, visto que a taxa de sobrevivência aumenta em função do diagnóstico nas fases iniciais da doença. Desta forma, o objetivo deste trabalho foi implementar e treinar um modelo de Machine Learning que possa auxiliar no diagnóstico do câncer de boca. Através das tecnologias como inteligência artificial (IA), que podem utilizar imagens em suas análises, busca-se melhorar o prognóstico do câncer de boca por meio da detecção precoce dele. Utilizando o ramo da IA, a Machine Learning e seu subgrupo Deep Learning, torna-se possível por intermédio de Rede Neural Convolutacional (RNC) realizar uma triagem de imagens de lesões malignas e pré-malignas, visando identificar a presença ou não do câncer de boca. A estrutura de RNC está baseada na estrutura de MobileNet, que separa as imagens em fragmentos e após treinamento, mostraram a identificação de câncer em 91% das imagens examinadas e de Leucoplasia em 84% das imagens analisadas.

ARTICLE INFORMATION

Article process:
Submitted: 03/12/2022
Approved: 14/04/2023
Published: 03/07/2023



Keywords:

Oral cancer, Artificial intelligence, Machine learning.

Keywords:

Câncer de boca, Inteligência artificial, Machine learning.

Introduction

Head and neck cancer (HNC) refers to a group of biologically similar cancers, which primarily affect the lips, oral cavity (mouth), nasal cavity, pharynx, larynx, and paranasal sinuses. Mouth cancer is one of the most frequent, where it represents approximately 40% of head and neck lesions (SCUTT et al., 2016). According to estimates from the National Cancer Institute (INCA), the number of new cases of oral cancer expected for Brazil, for each year of the 2020-2022 triennium, will be 11,180 cases in men and 4,010 in women. In Alagoas, 190 new cases are expected, with a predominance of males and residents outside the capital. Alagoas has an estimated incidence rate of 8.6/100 thousand in men and 4.06/100 thousand in women (INCA, 2020). These data reflect that cases of oral cancer cannot be overlooked and that the pathology should not be considered a rare disease.

An important ally for the early detection of oral cancer may be the use of technologies that can be accessible to the entire healthcare team and that can be used to suspect and aid in the diagnosis of both mouth cancer and potentially malignant diseases (Al-Rawi et al., 2022; Mahmood et al., 2020). In this way, it's possible to say that through past data and through statistical algorithms the machine can learn how this association should be made. This process, according to Hurwitz and Kirsch (2019), occurs through training. In the training, according to the authors, previous data of entries is reported, which can be purchase histories, locations, customer characteristics and how this data was classified. Without defining any association rules, the algorithm itself through statistical tests will learn and define the rules that define how the input data relates to its classifications (Lin et al., 2021; Uthoff et al., 2018; Welikala et al., 2020).

Due to this process of learning on its own the relationship of the input data with the data of its classifications, Machine Learning (ML) began to be used in the health area. An example that can be given occurs when collecting symptoms that patients had as input and a classification of whether or not the patient had a disease.

Through various tests, in training, through statistical algorithms using a programming language, the algorithm creates the rules that define the ML model. Through this model, whenever a new patient emerges, the algorithm can be used to predict whether or not a patient has a disease, if the model has already been trained to recognize these symptoms with the disease (Al-Rawi et al., 2022; Aubreville et al., 2017; Mahmood et al., 2020; Song et al., 2021; Uthoff et al., 2018; Welikala et al., 2020).

According to Erickson et al. (2017), when using ML for image-based diagnosis, X-ray images or the region where the symptom stands out from previous patients are used as training input data, as well as information on which of these patients had or did not have the disease to be analyzed. Finally, the algorithm would extract information on color tones, textures and

formats from the contents of the images that would be converted to numbers to be used in the training, generating the rules that will be used to predict images of new patients.

One of the most used formats currently to perform this training, still according to Erickson et al (2017), occurs through Deep Learning (DL), or deep learning. DL is a subgroup of ML, which has gained strength thanks to the improvement in current computers and still according to Ravi et al (2016), and yet that is based on the concepts of neural networks. To perform a training based on Neural Networks, according to Marcus (2018), it's necessary to interconnect several nodes, also called neurons, which will be separated by three layers: Input Layer, which is the layer that receives the input data from the algorithm; Hidden layers that perform more mathematical operations to improve the rules of training; and Output Layer that provides the expected result.

Since each node must perform a mathematical operation and pass it on to all nodes of the next layer, this process requires a high level of computational processing, but generates more accurate results than other algorithms not based on neural networks.

Given these concepts, the current research will consist of using LD algorithm, based on photos of patients who have oral cancer or not, to generate a model that is able to predict whether new patients may have this disease through photos, facilitating the screening process. Thus, the general objective was the development, implementation and training of a ML model that can assist in the early diagnosis of oral cancer through images of malignant and non-malignant oral lesions.

Methodology

Clinical images of malignant and non-malignant lesions were collected through photographs of patients living in the State of Alagoas who use the Unified Health System (Sistema Único de Saúde - SUS) and images from the Digital Atlas of Clinical Pathological Correlations of Oral Cancer and from oral pathology books. The model to aid the early diagnosis of oral cancer was developed from the recognition of the patterns of the collected images, as previously described, with the aid of artificial intelligence.

The patterns of malignant lesions and non-malignant lesions that are part of the differential diagnosis of CECCB were informed to the model by means of statistical algorithms that confer attributes to the images. The model was trained to recognize the patterns of lesions through the DL approach, using convolutional neural networks, generating a model capable of predicting whether new patients may have this pathology only through photos. The programs used for the development of the application based on artificial intelligence (AI) were: Keras (Python) and TensorFlow. Keras is a Python library with Deep Learning support that uses TensorFlow in your training. TensorFlow is an open source library intended for developers to create models of will be used as an ML tool based on the DL approach, where the creation of

the convolutional neural network was carried out using multiple levels of abstraction, using as a base the layers of MobileNet.

Development

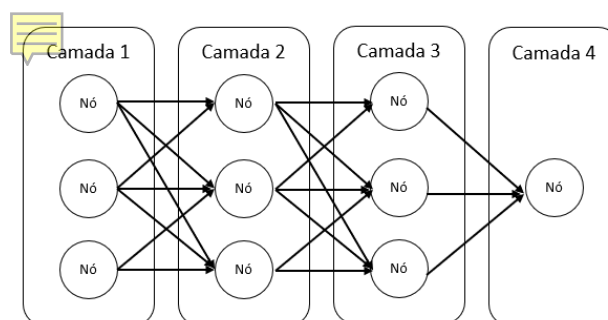
The images collected for the development of the AI model came from the photographs of patients living in the State of Alagoas from 2001 to June 2022, where 4,785 images were used and through a screening it was possible to select 1,000 to be used in the development of the AI model. In the atlas and the images extracted from the oral pathology books we were able to take advantage of 100% of the images since before they were inserted in it, it was passed through a screening process. After the collection and screening of the images that served for the training, the construction of the ML model was initiated.

The first step was to use the algorithm called ImageDataGenerator, which as its name suggests, is a resource capable of increasing the number of images used in training, creating images based on the old ones, using resources such as zooming, rotation or mirroring of the images.

After this step, the creation of the AI model used to screen the cancer photos was initiated. The model was built using the structure of convolutional neural networks based on the MobileNet structure, which consists of separating the images into small fragments, which will be converted to a numerical matrix used by the neural network.

A neural network consists of several nodes, which perform mathematical calculations in search of the desired answers. These nodes process your information and send your responses to other nodes that are interconnected in a next layer, as can be seen in Figure 1.

Figure 1. Representation of a Neural Network.



Source: Research data.

The structure of the neural network created for this model had 4 layers, in which the first and second contain 1024 nodes, the third 512, and the last only 2 nodes, to identify whether it's a positive or negative value for cancer (Figure 2).

Figura 2. Model construction.

```
1 #Modelo base, usando MobileNet
2 base_model=MobileNet(weights='imagenet',include_top=False)
3
4 #Estrutura da Rede Neural
5 x=base_model.output
6 x=GlobalAveragePooling2D()(x)
7 x=Dense(1024,activation='relu')(x)
8 x=Dense(1024,activation='relu')(x)
9 x=Dense(512,activation='relu')(x)
10 preds=Dense(2,activation='softmax')(x)
11
12 #Definição do modelo
13 model= Model(inputs=base_model.input,outputs=preds)
```

Source: Research data.

After all the training, the model was saved in HDF5 format, saving all its rules found in the training, so that it can be easily imported into other algorithms to perform the prediction of values, without having to perform all the training again.

As partial results it was observed that when the ML model performed the training for diagnosis of oral cancer in contrast to their respective differential diagnoses, there was an accuracy of 91% of correct answers for the cases of mouth cancer diagnosed. When the model was trained for the diagnosis of leukoplakia and the other potentially malignant diseases, there was an accuracy of 84% of correct answers.

Compared to the study by Ilhan et al (2021), which showed a positive identification accuracy of mouth cancer of 87%, and of leukoplakia of 70%. And from Song et al (2018), who using the RNC presented an accuracy in the imaging diagnosis of 86.9% of cases of oral cancer, the results of the present study may indicate that the model built is a product that can be used to aid in the diagnosis of oral cancer. In future work it's intended to test the model with different types of images such as photos taken on a cell phone camera in order to assist in the early diagnosis of oral cancer.

Final considerations

The present study showed that the model developed to assist in the early diagnosis of oral cancer can be used for this purpose, since the ML model presented during its training an accuracy of 91% of correct answers for the cases of diagnosed oral cancer and for the training of the diagnosis of leukoplakia and other potentially malignant diseases. There was an accuracy of 84% of correct answers. Thus opening other opportunities for models that have the purpose of assisting in the early diagnosis of oral cancer.

Funding agency

PPSUS - No. 60030.0000000214/2021.

Acknowledgments

Thanks to everyone who was part of this work, especially to my advisors and to the entire team of the Professional Master's Degree in Health Research – MPPS / Cesmac.

REFERENCES

- Al-Rawi, N., Sultan, A., Rajai, B., Shuaeeb, H., Alnajjar, M., Alketbi, M., Mohammad, Y., Shetty, S. R., & Mashrah, M. A. (2022). The Effectiveness of Artificial Intelligence in Detection of Oral Cancer: AI AND ORAL CANCER DIAGNOSIS. In *International Dental Journal* (Vol. 72, pp. 436–447). Elsevier Inc. <https://doi.org/10.1016/j.identj.2022.03.001>
- Aubreville, M., Knipfer, C., Oetter, N., Jaremenko, C., Rodner, E., Denzler, J., Bohr, C., Neumann, H., Stelzle, F., & Maier, A. (2017). Automatic Classification of Cancerous Tissue in Laserendomicroscopy Images of the Oral Cavity using Deep Learning. *Scientific Reports*, 7(1). <https://doi.org/10.1038/s41598-017-12320-8>
- Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. *Radiographics*, 37(2), 505-515.
- Hurwitz, J., Morris, H., Sidner, C., & Kirsch, D. (2019). *Augmented intelligence: the business power of human-machine collaboration*. CRC Press.
- Ilhan, B., Guneri, P., & Wilder-Smith, P. (2021). The contribution of artificial intelligence to reducing the diagnostic delay in oral cancer. In *Oral Oncology* (Vol. 116). Elsevier Ltd. <https://doi.org/10.1016/j.oraloncology.2021.105254>
- Ilhan, B., Lin, K., Guneri, P., & Wilder-Smith, P. (2020). Improving Oral Cancer Outcomes with Imaging and Artificial Intelligence. *Journal of Dental Research*, 99(3), 241–248. <https://doi.org/10.1177/0022034520902128>
- Lin, H., Chen, H., Weng, L., Shao, J., & Lin, J. (2021). Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis. *Journal of Biomedical Optics*, 26(08). <https://doi.org/10.1117/1.jbo.26.8.086007>
- Mahmood, H., Shaban, M., Indave, B. I., Santos-Silva, A. R., Rajpoot, N., & Khurram, S. A. (2020). Use of artificial intelligence in diagnosis of head and neck precancerous and cancerous lesions: A systematic review. In *Oral Oncology* (Vol. 110). Elsevier Ltd. <https://doi.org/10.1016/j.oraloncology.2020.104885>
- Marcus, R., & Papaemmanouil, O. (2018, June). Deep reinforcement learning for join order enumeration. In *Proceedings of the First International Workshop on Exploiting Artificial Intelligence Techniques for Data Management* (pp. 1-4).
- Ministério da Saúde. (2020). Instituto Nacional de Câncer José Alencar Gomes da Silva. *Estimativa 2020-2022: Incidência de Câncer no Brasil*. Rio de Janeiro.
- Ravi, D., Wong, C., Lo, B., & Yang, G.-Z. (2016). *Deep Learning for Human Activity Recognition: A Resource Efficient Implementation on Low-Power Devices*. IEEE Engineering in Medicine and Biology Society. <https://doi.org/10.1109/BSN.2016.7516235>
- Scutti, J. A. B., Pineda, M., Jr, E. E., & Ameida, E. R. de. (2016). *Carcinoma de células escamosas de cabeça e pescoço (HNSCC): desvendando os mistérios do microambiente tumoral*.
- Song, B., Sunny, S., Li, S., Gurushanth, K., Mendonca, P., Mukhia, N., Patrick, S., Gurudath, S., Raghavan, S., Tsusennaro, I., Leivon, S. T., Kolar, T., Shetty, V., Bushan, V. R., Ramesh, R., Peterson, T., Pillai, V., Wilder-Smith, P., Sigamani, A., ... Liang, R. (2021). Bayesian deep learning for reliable oral cancer image classification. *Biomedical Optics Express*, 12(10), 6422. <https://doi.org/10.1364/boe.432365>
- Uthoff, R. D., Song, B., Sunny, S., Patrick, S., Suresh, A., Kolar, T., Keerthi, G., Spires, O., Anbarani, A., Wilder-Smith, P., Kuriakose, M. A., Birur, P., & Liang, R. (2018). Point-of-care, smartphone-based, dual-modality, dual-view, oral cancer screening device with neural network classification for low-resource communities. *PLoS ONE*, 13(12). <https://doi.org/10.1371/journal.pone.0207493>
- Welikala, R. A., Remagnino, P., Lim, J. H., Chan, C. S., Rajendran, S., Kallarakkal, T. G., Zain, R. B., Jayasinghe, R. D., Rimal, J., Kerr, A. R., Amtha, R., Patil, K., Tilakaratne, W. M., Gibson, J., Cheong, S. C., & Barman, S. A. (2020). Automated Detection and Classification of Oral Lesions Using Deep Learning for Early Detection of Oral Cancer. *IEEE Access*, 8, 132677–132693. <https://doi.org/10.1109/ACCESS.2020.3010180>