Formation of a cooperation network in Mato Grosso on Machine Learning and Image Analysis: Diagnosis of COVID-19 in X-ray images

Jeniffer L. Z. Luz
UNEMAT, Research Group PIXEL
Rondonópolis-MT, Brazil
jeniffer.luz@unemat.br

Scenio M. A. de Araujo
UNEMAT, Research Group PIXEL
Rondonópolis-MT, Brazil
scenio.mathias@unemat.br

Caio C. E. de Abreu
UNEMAT
Alto Araguaia-MT, Brazil
caioenside@unemat.br

Juvenal Silva Neto
UNEMAT
Alto Araguaia-MT, Brazil
juvenalneto@unemat.br

Carlos A. S. J. Gulo
UNEMAT, Research Group PIXEL
Rondonópolis-MT, Brazil
sander@unemat.br

ABSTRACT

Since the beginning of the COVID-19 outbreak, the scientific community has been making efforts in several areas, either by seeking vaccines or improving the early diagnosis of the disease to contribute to the fight against the SARS-CoV-2 virus. The use of X-ray imaging exams becomes an ally in early diagnosis and has been the subject of research by the medical image processing and analysis community. Although the diagnosis of diseases by image is a consolidated research theme, the proposed approach aims to: a) apply state-of-the-art machine learning techniques in X-ray images for the COVID-19 diagnosis; b) identify COVID-19 features in imaging examination; c) to develop an Artificial Intelligence model to reduce the disease diagnosis time; in addition to demonstrating the potential of the Artificial Intelligence area as an incentive for the formation of critical mass and encouraging research in machine learning and processing and analysis of medical images in the State of Mato Grosso, in Brazil. Initial results were obtained from experiments carried out with the SVM (Support Vector Machine) classifier, induced on a publicly available image dataset from Kaggle repository. Six attributes suggested by Haralick, calculated on the gray level co-occurrence matrix, were used to represent the images. The prediction model was able to achieve 82.5% accuracy in recognizing the disease. The next stage of the studies includes the study of deep learning models.

KEYWORDS

machine learning, image analysis, COVID-19, medical imaging

1 INTRODUCTION

Since the end of 2019, the world has been making efforts to contain the spread of the SARS-CoV-2 virus (Severe Acute Respiratory Syndrome Coronavirus 2), which causes COVID-19 disease. Although laboratory tests use a process known as RT-PCR (Real-Time Polymerase Chain Reaction) and serve as a basis to confirm COVID-19 cases, these tests suffer from insufficient sensitivity, as reported in 71% of the researches carried out by [3]. It happens due to several factors, such as sample preparation and quality control [2]. Image analysis has allowed the development of refined techniques able to identify risks, from medical data still not determined, using mainly decision trees, such as those previously used for pneumonia risk prediction [6]. Hence, when achieving extreme precision, each patient is individually monitored through imaging exams, making treatment and diagnosis faster and more accurate and creating what could be identified as a "Corona Pattern" for an algorithm that uses artificial intelligence alongside with computer vision methods [5].

The objective of this work lies in applying and developing frontier research on machine learning and its applications in X-ray images for the COVID-19 diagnosis [4]. Medical imaging equipment is accessible in clinical settings, such as computed tomography and X-rays, and allows more accurate diagnoses also in asymptomatic patients, preventing the spread of the virus [1, 3]. Based on this context, this work presents a preliminary analysis of the machine learning model based on the SVM (Support Vector Machine) classifier, for the automatic classification and diagnosis of COVID-19 using X-ray images. The availability of equipment to perform X-ray examinations is a high and low cost, regarding all aspects. Thus, this system can be applied as a support tool, from screening patients to making decisions about hospitalizations and monitoring patients during its hospitalizations, reducing the need for other imaging tests. Nevertheless, in the medium term, the work developed has contributed to the development of computational solutions using machine learning techniques, artificial intelligence, and computer vision in the state of Mato Grosso, in Brazil.

2 PROPOSED APPROACH

The SVM model was chosen considering its satisfactory results in several aspects with successful application [5]. All the principles previously established by this technique were followed based on the statistical learning theory for obtaining classifiers. This acquisition can also be considered a searching process, where among all the hypotheses generated by the algorithm, we consider the one with the best ability to describe the domain in which computational learning occurs [5]. The machine learning implemented in the project is that supervised, with the algorithm being fed with images divided into two subsets: the training set, used for training and adjusting the classifier; and the test set, used to determine the degree of effectiveness of the built model. The implementation occurred in Python 3.8 with the scikit-learn 0.22.2 library, the tests

being performed on the Google Colaboratory platform (Google Colab).

This work used a public dataset, made available for research purposes on Kaggle platform [1] (online community of data scientists subsidized by Google LLC) [2]. The chosen dataset contains 188 lung X-ray images, among which 94 of patients diagnosed with COVID-19 and 94 pulmonary images without any comorbidities, being these the two groups addressed in this research. For the training stage, 74 images were distributed in each class called "covid" for patients with pneumonia characteristic of COVID-19 and "normal" for lungs without the disease. Sample images of a healthy lung and COVID-19 are shown in Fig. 1. The remaining images were randomly distributed to test the model proposed. Before performing

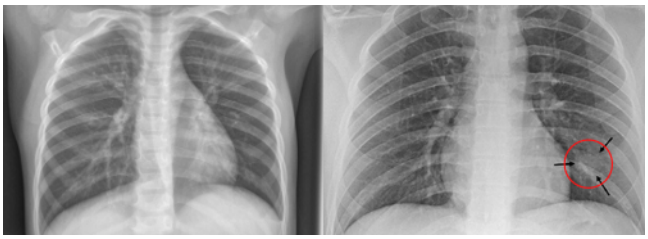

Figure 1: Healthy Lung and Ground-glass opacity indicating COVID-19 pneumonia, respectively

the algorithm's training step, it is valid to highlight that the dataset went through the segmentation process, identifying the features of pulmonary problems [1]. The images went through the stage known as feature extraction, responsible for converting the data into only a relevant group of bits [6]. The feature extraction technique adopted, the Gray-Level Co-occurrence Matrix (GLCM) [6], is based on the texture analysis of the images and is widely used in the field of X-ray images classification [1, 6]. The GLCM stores the number of occurrences that a pixel of a certain gray level intensity occurs with another pixel in the image, considering location patterns, for example, following a pattern of horizontal comparison at 0º, vertical at 90º or diagonally at 45º, or 135º. In that manner, six features of the GLCM matrix were considered for each image: contrast, correlation, dissimilarity, energy, entropy, and homogeneity. Each part of a pulmonary X-ray image has unique attributes and characteristics, which allows identifying categories in the texture patterns of the images analyzed. Texture analysis algorithms capture and encode these attributes into representative numerical values to improve the learning abilities of a machine learning model. The images used in the experiments have up to eight Gray-Levels, resulting in four GLCM matrices, being one matrix for each spatial orientation and the chosen features were normalized to a single one-dimensional vector (1×64) [6].

3 DISCUSSION

The metrics used to calculate the model's performance and most used in the area of artificial intelligence are: 1) Accuracy - the percentage of images classified correctly; 2) Precision - the percentage of images of class C correctly classified among all those predicted as belonging to class C; 3) Recall - the percentage of class C images that were correctly predicted; and 4) F1-score - consists of the harmonic mean of the precision and recall parameters, since precision and recall do not report the number of other images classified incorrectly, or not belonging to class C. F1-score is calculated as follows:

$$F1 = \frac{2 \times (precision \times recall)}{(precision + recall)} \quad (1)$$

The model achieved a high precision rate of 84% for normal class images and 81% for COVID-19 class images, 80% recall for normal class images and 85% for COVID-19 class images, and the F1-score metric of 82% for normal class images and 83% for covid class images, using the RBF kernel with default parameters. In this way, the SVM classifier with RBF kernel diagnosed COVID-19 in 82.5% of the tested images.

Based on the confusion matrix, using the SVM classifier with RBF kernel, the values contained in each type of classification were: 16 images predicted to be true (occurs when the value real is positive and the classifier correctly predicts positive); 4 images for false positive (occurs when the actual value is negative and the classifier incorrectly predicts positive); 17 for false true (occurs when the actual value is negative and the classifier correctly predicts negative), and 3 images for false negative (occurs when the actual value is positive, but the classifier incorrectly predicts negative). It is possible to assume that, with these result, the classifier correctly predicted the diagnosis of COVID-19 in 33 images, in a total of 40.

4 CONCLUSIONS

Analyzing the still growing curve of COVID-19 cases in Brazil and in the world, an effective and fast method for diagnosis is something extremely necessary. The analysis of X-ray images with the aid of machine learning techniques with supervised learning, combined with the knowledge of specialists in the medical field, becomes a model capable of meeting these needs. The results obtained, although preliminary, are already capable of indicating that the research is on the right path for further development in this particular topic and others in the area of machine learning. The next stage of the project includes further studies in the different kernels of the SVM, as well as the initiation of studies of deep learning networks and the use of dataset with a larger number of images.

REFERENCES

[1] Ioannis D. Apostolopoulos and Tzani A. Mpesiana. 2020. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine* 43, 2 (2020), 635–640.
[2] Joseph P. Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. 2020. COVID-19 Image Data Collection: Prospective Predictions Are the Future. (2020).
[3] Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, and Wenbin Ji. 2020. Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. *Radiology* 296, 2 (2020), 115–117.
[4] Daniela L. Freira, Rodrigo F. A. P. de Oliveira, Carmelo J. A. B. Filho, Pedro Buarque, and Ana Clara A. M. V. F. de Medeiros. 2020. Machine Learning Applied in SARS-CoV-2 COVID 19 Screening Using Clinical Analysis Parameters. *IEEE Latin America Transactions* (2020), 978–985.
[5] Ophir Gozes, Maayan Frid-Adar, Hayit Greenspan, Patrick D. Browning, Huangqi Zhang, Wenbin Ji, Adam Bernheim, and Eliot Siegel. 2020. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis. (2020).
[6] Sergio Varela-Santos and Patricia Melin. 2021. A new approach for classifying coronavirus COVID-19 based on its manifestation on chest X-rays using texture features and neural networks. *Information Sciences* 545 (2021), 403–414.

[1] https://www.kaggle.com/khoongweihao/covid19-xray-dataset-train-test-sets