COVID-19 Lung CT Images Recognition: A Feature-Based Approach

Chiara Losquadro[1][0000–0002–9634–7660], Luca Pallotta[1][0000–0002–6918–0383], and Gaetano Giunta[1][0000–0002–1514–2576]

[1] Roma Tre University, Department of Engineering, Rome, Italy
chiara.losquadro@uniroma3.it, luca.pallotta@uniroma3.it, gaetano.giunta@uniroma3.it

Abstract. The SARS-CoV-2 is quickly spreading worldwide resulting in millions of infection and death cases. As a consequence, it is increasingly important to diagnose the presence of COVID-19 infection regardless of the technique applied. To this end, this work deals with the problem of COVID-19 classification in order to differentiate COVID-19 versus healthy Computed Tomography (CT) images. In particular, first-order statistical measures as well as numerical quantities extracted from the autocorrelation function are investigated with the aim to provide an efficient classification process ensuring satisfactory performance results.

Keywords: COVID-19 · Classification · Feature extraction · Lung Computed Tomography (CT) images.

1 Introduction

On December 31st 2019, the Health Commission of Wuhan in China, informs the World Health Organization (WHO) about a cluster of unknown pneumonia cases in the province, later identified as the SARS-CoV-2. The initially unknown SARS-CoV-2 is a member of the beta-coronavirus family, currently named as COVID-19, and it can lead to mild to severe respiratory tract infections with pneumonia, cold and fever, headache, painful throat and weakness as traditional clinical symptoms. In the worst scenario, the infected patients could also present dyspnoea or hypoxaemia signs but some asymptomatic cases were also observed where any acute COVID-19 symptoms are exhibited [1]. However, for both asymptomatic or acute situation, an early and accurate diagnosis represents a crucial step in order to reduce the COVID-19 mortality and to improve the treatment of this insidious disease. Nowadays, several methods are employed to provide a definitive diagnosis of COVID-19, including reverse transcriptase-polymerase chain reaction (RT-PCR) [2], serology tests, and medical imaging techniques. In particular Computed Tomography (CT) scans and chest X-Ray have been proved to be a great supplement to assess the infection severity [3]. Artificial intelligence technologies are achieving remarkable progress with respect to medical image analysis, with the aim to assist the clinicians classifying different diseases, due to the possibility to extract quantitative features from the acquired images [4]. Recently, several studies have focused on these methodologies in order to provide a secure and automatic way to diagnose also the COVID-19 infection applying several machine learning methods [5–11]. Following the line of reason of these works, this study proposes and investigates a new feature-based methodology in order to classify COVID-19 versus normal CT images. The analyses have been conducted on the challenging CT axial scans data deeply described in [12]. Interestingly, the obtained results have shown the effectiveness of the proposed features in distinguish among COVID-19 and normal patients.

2 Problem Description and Proposed Classification Algorithm

The proposed solution focuses on the exploitation of intrinsic discriminative properties of textural features extracted from lung CT images. In particular, a joint statistical and numerical approach to texture analysis is used starting from the gray-level distribution within the images in order to differentiate them. To this end, the texture information is derived from the histogram computed from the lung CT images, through which both first-order statistic measures and numerical quantities associated with the autocorrelation function are evaluated. The proposed classification algorithm is synthetically represented in Figure 1, and it can be seen as composed by few steps that are described in the following lines.

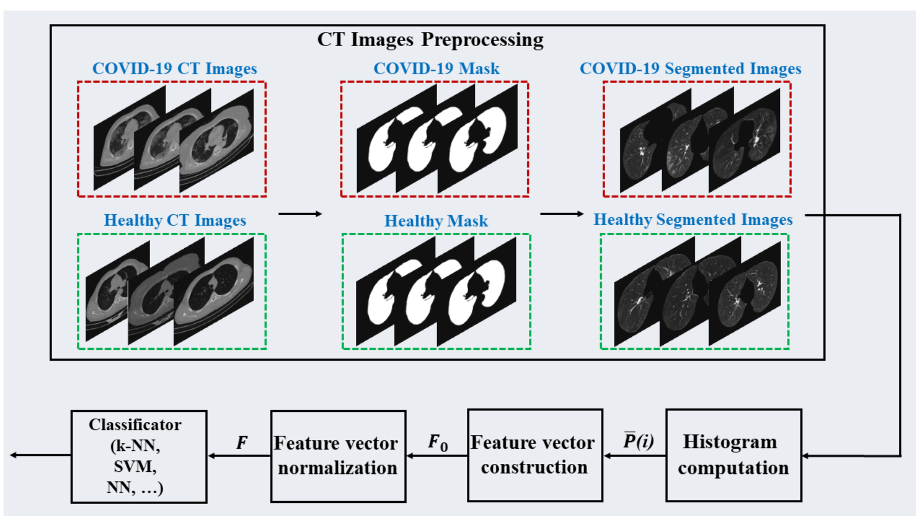

Fig. 1. Block scheme of the proposed CT images feature extraction algorithm.

2.1 Lung Segmentation

In order to achieve accurate classification results, the segmentation process represents a preliminary step of paramount importance [13] when dealing within the framework of medical images. Precisely, lung segmentation aims to basically separate the pixels corresponding to the lung cavity from the surrounding chest anatomy in order to avoid erroneous classification due to the overall anatomy of the patients under observation. For such a purpose, in the proposed procedure, a segmentation technique has been implemented.

2.2 Histogram Computation

In the attempt to differentiate COVID-19 from healthy patients CT scans, a first-order statistical approach is investigated by exploiting the gray-level distribution within the CT images [14]; in particular, as a result of first-order statistical techniques, textural information are derived from the histogram, that is computed for each individual image to classify. A histogram example of a COVID-19 CT scan versus a healthy patient CT scan is shown in Figure 2. It is worth to observe that, through a preliminary visual inspection of Figure 2, a histogram study may reveal useful to discriminate a COVID-19 image from a healthy one.

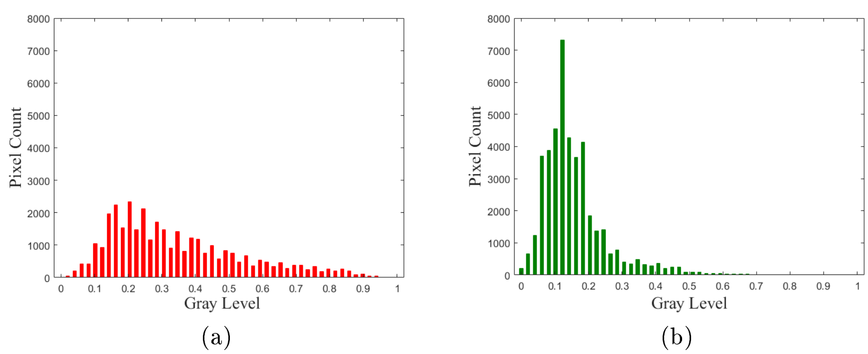

Fig. 2. Histogram computation. Subplots refer to (a) COVID-19 and (b) healthy patients.

2.3 Feature Extraction

Starting from the considerations made in subsection 2.2, a histogram-based feature extraction method is presented to classify COVID-19 and healthy patients CT axial scans. Therefore, the initial point is the histogram normalized to have unit area, say $\overline{P}(i)$, derived at the previous step of the proposed algorithm, from which several quantities are extracted to construct the so-called feature vector indicated with $\boldsymbol{F}_0$ in Figure 1. More precisely, $\boldsymbol{F}_0$ comprises four first-order statistical measures, viz. mean, standard deviation, skewness and kurtosis, lined-up to an additional measure, the median, and three other indices, namely the peak sidelobe level (PSL) ratio and two different definitions of the integrated sidelobe level (ISL) ratio of the autocorrelation function [15] directly computed from $\overline{P}(i)$. As to the PSL and the two different definitions of the ISL, they are computed from the normalized autocorrelation of $\overline{P}(i)$, indicated as $C_P(k)$, $k = 0, ..., K-1$, as follows:

$$\mathrm{PSL} = \max_k \frac{|C_P(k)|}{|C_P(0)|}, \tag{1}$$

$$\mathrm{ISL}_1 = \frac{\sum_{k=1}^{K-1} |C_P(k)|}{|C_P(0)|}, \tag{2}$$

and

$$\mathrm{ISL}_2 = \frac{\sum_{k=1}^{K-1} |C_P(k)|^2}{|C_P(0)|}, \tag{3}$$

having indicated with $|\cdot|$ the modulus of its argument. The last step of the devised algorithm consists in normalizing the resulting feature vector, $\boldsymbol{F}_0$, thanks to the following linear rescaling:

$$\boldsymbol{F} = \frac{\boldsymbol{F}_0 - \mu_{\boldsymbol{F}_0}}{\sigma_{\boldsymbol{F}_0}} \tag{4}$$

meaning $\mu_{\boldsymbol{F}_0}$ as the mean and $\sigma_{\boldsymbol{F}_0}$ as the standard deviation of the feature vector $\boldsymbol{F}_0$. This is done to avoid that a very strong feature value could polarize the classifier's decision.

3 Performance Assessment

This section is aimed at showing the classification capabilities of the algorithm described in Section 2 in terms of CT scans discrimination exploiting their histograms. Therefore, a deep description of the patients dataset is first provided; then, the classification procedure is detailed together with the discussion of the obtained results.

3.1 Patients Dataset

The analyses are conducted on patients with either COVID-19 condition and healthy condition, which are included in the dataset herein exploited, namely the Cov19-Healthy Dataset. The latter is selected from the original COVID-CTset, that has been gathered from Negin Medical Radiology Center located at Sari (Iran) between March and April, 2020 [12]. In particular, COVID-CTset contains the CT axial scans of both COVID-19 and healthy people, acquired using a CT scanner (SOMATOM Scope Power, Siemens Healthcare) and stored in a 512×512 DICOM format.

For the purpose of this study, 180 CT scans (i.e., Cov19-Healthy Dataset) are extracted from the original dataset including 80 COVID-19 infected images and 100 uninfected images, respectively. Table 1 summarizes the patients population details in terms of their sex and age (expressed as mean with its confidence interval in terms of standard deviation (SD), namely, ±SD).

Table 1. Sex and age details of patients enrolled in the study.

		COVID-19	*Healthy*
Sex	*Male*	48	48
	Female	32	52
Age y (±SD)		49.8±14.8	39.8±10.6

3.2 Classification Procedure and Results

This subsection is aimed at showing the effectiveness of the devised methodology in discriminating the COVID-19 and healthy patients described in the previous subsection.

The classification is performed dividing the dataset into two non-overlapped groups: the training set, composed by 70% of the available data and the test set, composed by the remainder 30% (i.e., excluding all the data of training phase). Moreover, since the aim of this work is to show the effectiveness of the proposed feature-extraction based method, we use as classifier the k-nearest neighbour (k-NN) because of its low computational burden, with the parameter k set equal to 5 [16]. Then, to prove the effectiveness of the proposed algorithm, the classification accuracy, say A_{cc}, is used as figure of merit, whose analytic expression is given by

$$A_{cc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}, \tag{5}$$

where TP is the total number of true positives, FP is the total number of false positives, TN is the total number of true negatives, and finally FN represents the total number of false negatives.

Moreover, in order to provide a statistical characterization of the entire classification method, the average classification accuracy is estimated by means of a standard Monte Carlo approach. More precisely, for each independent Monte Carlo trial a different selection of the training and test sets is randomly chosen for each class (i.e., COVID-19 CT scan and healthy CT scan). This process reiterates N number of times (note that in this study N is set equal to 100) the training and test procedures by independent random extractions. The results in terms of average A_{cc} are shown also in comparison with the feature extraction algorithm based on the exploitation of the four statistical measures (i.e., mean, standard deviation, skewness and kurtosis) exclusively, indicated as first order based approach (FOA).

Table 2 shows the classification results in terms of classification accuracy for the above defined two classes. From the table it is evident that the proposed feature extraction method of Section 2.3 is capable of ensuring a satisfactory performance reaching the 84.63% of average correct classification. Conversely, the FOA provides a lower correct classification percentage. In particular, these results are also confirmed by the fact that the maximum achieved classification accuracy of the proposed algorithm is equal to 92.60% still overcoming that reached by the FOA.

Table 2. Classification accuracy (expressed in percentage) for COVID-19 versus healthy patients.

	average A_{cc}	maximum A_{cc}
FOA	76.85%	83.33%
proposed method	84.63%	92.60%

To give further insights about the classification capabilities of the proposed algorithm, the average confusion matrices and the one associated with the maximum A_{cc} for the quoted algorithms are reported in Figure 3.

As can be clearly observed, the confusion matrices corroborate the results obtained in terms of overall classification capabilities. As a matter of fact, the use of all 8 features extracted through the proposed procedure allows to identify COVID-19 patients in spite of the very low computational cost of this method. It is finally worth to highlight that the promising results of the considered feature extraction method could be surely improved considering for instance more sophisticated classifiers such as a support vector machine (SVM) [17], neural networks [18], and so on.

4 Conclusions

In this paper a new feature extraction procedure has been proposed aiming to discriminate COVID-19 patients from healthy ones. The presented method has dealt with the exploitation of existing intrinsic discriminative textural properties from lung CT images, starting from their histograms computation. In particular, the histogram-based approach has allowed to investigate several first-order statistical measures and numerical quantities derived from the histogram autocorrelation which have been used as input to a k-NN classifier. The performance of the proposed extracted features has been assessed on CT axial scans of both

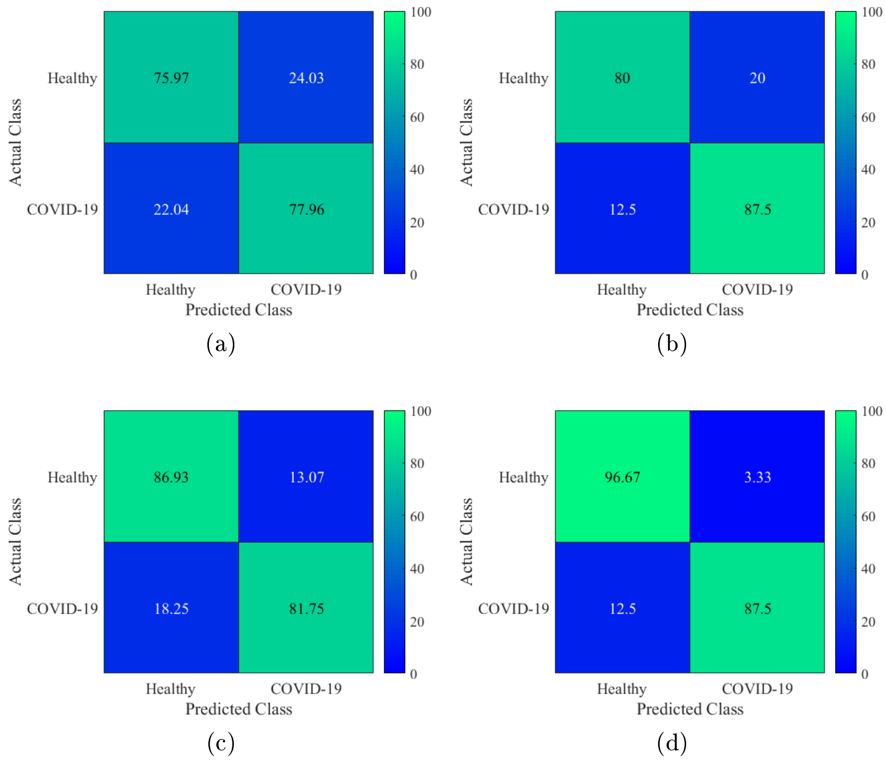

Fig. 3. Confusion matrices, actual class versus predicted class (%). Subplots refer to (a) average confusion matrix of FOA, (b) confusion matrix of the best case of FOA, (c) average confusion matrix of the proposed method, and (d) confusion matrix of the best case of the proposed method.

COVID-19 affected and not affected people, ensuring a satisfactory average correct classification result of 84.63% and reaching, at the end, the high performance of 92.60% of classification accuracy.

Possible future research works might consider the use of other features to be line-up to those proposed in this paper to improve its recognition capabilities as well as the test of the proposed features by means of neural networks.

Aknowledgments

The authors would like to thank the authors [12] for providing the CT axial data. All data supporting this paper are available at https://github.com/mr7495/COVID-CTset.

References

1. S. Das, S. Das, and M. M. Ghangrekar, "The COVID-19 pandemic: biological evolution, treatment options and consequences," *Innovative Infrastructure Solutions*, vol. 5, no. 3, 2020.
2. Y. Li, L. Yao, J. Li, L. Chen, Y. Song, Z. Cai, and C. Yang, "Stability issues of RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed with COVID-19," *Journal of Medical Virology*, vol. 92, pp. 903–908, 2020.
3. T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, "Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases," *Radiology*, vol. 296, pp. 32–40, 2020.
4. J. Zhang, Y. Xia, Y. Xie, M. Fulham, and D. D. Feng, "Classification of Medical Images in the Biomedical Literature by Jointly Using Deep and Handcrafted Visual Features," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1521–1530, 2018.
5. X. Qian, H. Fu, W. Shi, T. Chen, Y. Fu, F. Shan, and X. Xue, "M^3 Lung-Sys: A Deep Learning System for Multi-Class Lung Pneumonia Screening From CT Imaging," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3539–3550, 2020.
6. I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, vol. 43, pp. 635–640, 2020.
7. J. C. Gomes, V. A. de F. Barbosa, M. A. Santana, J. Bandeira, M. J. S. Valença, R. E. de Souza, A. M. Ismael, and W. P. dos Santos, "IKONOS: An intelligent tool to support diagnosis of Covid-19 by texture analysis of x-ray images," *medRxiv*, 2020.
8. M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods," 2020.
9. S. Latif, M. Usman, S. Manzoor, W. Iqbal, J. Qadir, G. Tyson, I. Castro, A. Razi, M. N. K. Boulos, A. Weller, *et al.*, "Leveraging data science to combat covid-19: A comprehensive review," *IEEE Transactions on Artificial Intelligence*, 2020.
10. I. Ozsahin, B. Sekeroglu, M. S. Musa, M. T. Mustapha, and D. Uzun Ozsahin, "Review on diagnosis of covid-19 from chest ct images using artificial intelligence," *Computational and Mathematical Methods in Medicine*, vol. 2020, 2020.
11. R. Ito, S. Iwano, and S. Naganawa, "A review on the use of artificial intelligence for medical imaging of the lungs of patients with coronavirus disease 2019," *Diagnostic and Interventional Radiology*, vol. 26, no. 5, p. 443, 2020.
12. R. Mohammad, A. Abolfazl, and S. Seyed, "A Fully Automated Deep Learning-based Network For Detecting COVID-19 from a New And Large Lung CT Scan Dataset," 2020.
13. P. Aggarwal, V. Renu, S. Bhadoria, and C. Dethe, "Role of Segmentation in Medical Imaging: A Comparative Study," *International Journal of Computer Applications*, vol. 29, no. 1, pp. 54–61, 2011.
14. C. H. Chen, C. K. Chang, C. Y. Tu, W. C. Liao, B. R. Wu, K. T. Chou, Y. R. Chiou, S. N. Yang, G. Zhang, and T. C. Huang, "Radiomic features analysis in computed tomography images of lung nodule classification," *PloS one*, vol. 13, no. 2, p. e0192002, 2018.
15. A. R. Persico, C. Clemente, D. Gaglione, C. V. Ilioudis, J. Cao, L. Pallotta, A. De Maio, I. Proudler, and J. J. Soraghan, "On model, algorithms, and experiment for micro-doppler-based recognition of ballistic targets," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 3, pp. 1088–1108, 2017.

16. R. O. Duda and P. E. H. et al., *Pattern classification and scene analysis*, vol. 3. Wiley New York, 1973.
17. S. Sharma and P. Khanna, "Computer-aided diagnosis of malignant mammograms using Zernike moments and SVM," *Journal of Digital Imaging*, vol. 28, no. 1, pp. 77–90, 2015.
18. L. Meng, D. Dong, L. Li, M. Niu, Y. Bai, M. Wang, X. Qiu, Y. Zha, and J. Tian, "A deep learning prognosis model help alert for covid-19 patients at high-risk of death: A multi-center study," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3576–3584, 2020.