Response to the Reviewers' comment on "Convolutional Neural Network-based automatic analysis of chest radiographs for the detection of COVID-19 pneumonia: a prioritizing tool in the emergency department, phase I study and preliminary "real life" results"

D. Tricarico, M. Calandri, M. Barba, C. Piatti, C. Geninatti, D. Basile, M. Gatti, M. Melis and A. Veltri

February 4th, 2022

We would like to thank the Assistant Editor and the three reviewers for the time and effort in handling this manuscript, and for the insightful comments and suggestions that guided us in improving this paper. The Reviewers 1 and 2 provided suggestions on how to improve the method and discussion section, from both statistical analysis and description of the proposed algorithm, Reviewer 3, recommended this paper for English medical language proof. We summarized the main concerns from the Reviewers as follows:

- **Statistical Analysis**: Referring to another article published in 2021 [1], Reviewer 1 suggested to integrate similar analysis in our manuscript. We would like to better explain the fact that in the work we presented we did not have a scoring label, each image is only labelled by its membership to one class: **COVID-19**, in case of positive RT-PCR test within 24h, **No Finding**, if no radiological finding is present, **Other**, otherwise. In the referenced article, every sample was labelled by using a **score from 0 to 18**, indicating the severity of COVID-19 infection; this difference prevented us from exactly replicating the approach on our data. Moreover, the selection criteria in the reference study include, apart from the RT-PCR test result, **the presence of radiological findings in the chest x-ray** to support the identification of biomarkers discriminatory for the COVID-19. **In our case this selection has not been performed,** with the intention to evaluate the proposed method in a setting similar to

the **triage use case** when this selection criterion cannot be applied a-priori. Finally, the two datasets differ for sample size: in the referenced work, **1000 positive patients** were provided, while **236 positive cases** were available for our study. We modified our manuscript to compare the methodologies in the two papers. To integrate and compare the two results we added the following contents:

- **Distribution of COVID-19 risk indicator** computed by AIPPO, grouped by ground-truth class membership and gender to describe how the score takes higher values in case of COVID-19 positive cases and to illustrate any differences in model behaviour between genders.
- **Distribution of COVID-19 risk indicator computed by the method for COVID-19 positive group, divided by Age Group and Gender**. Because of the different age distribution in our dataset, the Age Groups differ from the referenced article: we grouped together patients with less than 50 years and split the ≥80 years cohort in two cohorts for 80- 89 and ≥90 years. For each age group we studied the score distribution by mean and quantiles, and we tested the differences between genders with Mann-Whitney U test (alpha = 0.05). The difference was significant for 4 out of 6 age groups, highlighting a possible gender bias in model scoring. Unfortunately, with the class labelling we could not prove this is caused by the algorithm itself or by the underlying COVID-19 severity distribution in the data. We thank Reviewer 1 for the insightful methodological suggestion, we will reserve a more detailed analysis for our future endeavours.
- **Performance differences between groups**: Reviewer 1 required a more detailed analysis on how the proposed approach performs when applied to subgroups of the whole data. While in the previously submitted version of this manuscript we took in consideration performance variations by using 6 times iterated 5-fold cross validation strategy, in this newer version we divided the dataset by age group and gender to highlight significant differences. We compared distributions for different age groups by means of Kruskal-Wallis statistical test (alpha = 0.05) and between gender groups by means of Mann-Whitney U statistical test (alpha = 0.05). In both cases the difference resulted significant, with p-value < 1.12e-12 and p-value < 1.60e-4 respectively. This difference suggests a possible gap in performance between groups. Considering age groups, we notice higher than average performance for

younger patients. For gender grouping, the gap in MAP performance can be interpreted by both the effect of male prevalence in the dataset, 305 cases against 231, and a different class distribution, with a higher portion of No Finding in the female group. We thank Reviewer 1 for the useful methodological proposal, which spotlights this aspect, about which we will reserve more detailed study for our future works.

- **Correlation between score and age**: in the article referenced by Reviewer 1, the correlation between COVID-19 severity score elaborated by radiologists and patient age was analysed. In our dataset, COVID-19 severity score is not provided, while only class membership is known, **for this reason it is not possible to compare correlations with predicted score and radiologists' one**. To match our results with ones from [1], we computed the Spearman correlation coefficient [Rs] between patient age and COVID-19 risk indicator computed by the proposed algorithm for patients positive to COVID-19. We made this calculation on both the whole dataset and divided by gender group. For the whole dataset, there was almost no correlation between the two variables, Rs = -0.004, p-value = 0.884; in the referenced article a low positive correlation was obtained, but it was not statistically significant. Grouped by gender, the correlation resulted low and positive in both cases, but not significant, Rs = 0.167 p-value < 5.52e-10 for the female group, and Rs = 0.065 p-value < 5.13e-4 for the male one, confirming the results in the referenced paper.

- **<u>Presentation of available data</u>**: Reviewer 1 suggested reviewing the content of **Table 2 and 3**, to provide information in a more informative way. By reference, the Tables 2 and 3 of referenced paper [1] where indicated. We thank the Reviewer for the suggestion, useful to a more understandable and comprehensive description of data used. We would like to highlight that the **differences in dataset labelling prevented us from fully replicating** the indicated approach. In the section "**3.1 Baseline characteristics**" from past version of this work, we analysed the **age distributions for COVID-19 positive and negative patients** proving they do not differ significantly (Mann-Whitney U test). In the updated manuscript, we analysed the **variability of class membership** (COVID-19, Other, No Finding) **and age** when data is divided into

groups. When grouping by gender, the differences between distributions were **not significant**, chi-square test p-value = 0.986 for class membership, Mann-Whitney U test p-value = 0.178. Finally, we divided samples by age groups, evaluating the differences in COVID-19 positive case prevalence between males and females; this difference **resulted significant** for every age group apart from patients in the 80-89 years old range. We summarised these results in **dedicated tables**.

- **Introduction completeness:** Reviewer 2 highlighted that the Introduction section did not provide enough information to the Reader to fully understand the context of deep learning applications in early detection of COVID-19 by analysing CRXs. To overcome this issue, we updated the manuscript with more references to works on this topic and we summarized the different tasks that have been tackled in this field. We think this broader overview can help the Reader to collect all the necessary information to meaningfully frame our work.

- **Algorithm details:** We would like to thank Reviewer 2 for the indication to improve the description of the algorithm providing more details. We updated the manuscript with comprehensive description of the steps and change of wording, with the aim of providing more clarity:
    - **Image preprocessing**: Reviewer identified some missing details in the description of image preprocessing, asking for clarifications about the usage of U-net model to extract the lung portion of the image. To provide an immediate description of the steps, **we added a new figure** to the manuscript illustrating the steps in detail. The original image **a)** is properly resized limiting the maximum dimension to 1000 pixels but maintaining its original aspect ratio **b)**. The resized picture **c)** is then processed by the U-Net model **d)** to extract a binary mask **e)** indicating where the lungs are located. To define the final cut **g)**, the binary mask **e)** is framed by the smallest rectangle that contains the lungs **f)**. The resized picture **c)** is normalised by mean and variance **h) i)** and, finally cropped **j)** using the computed rectangular cut to obtain the final picture **k).**

- **Score calculation**: Reviewer asked to clarify the **mechanism behind the calculation of COVID-19 risk indicator** based on the similarity metrics. To equip the Reader with the required information, we added a dedicated paragraph in the manuscript. **In the database** of our proposed method, indicated f) in Figure 1, is stored a set of cases from the past about the ground truth is given. This set is referred to as a **"reference set"**; we would like to enforce the fact that **only the feature vectors** extracted from the images **and related ground truth** are contained in the dataset. When a new picture is received by the system and its feature vector is computed by the deep neural network model, the **method calculates the angular similarity** between the computed vector and the ones stored in the database. The system **retrieves the 10 most similar cases and related ground truth**. To compute the final score, the system **counts how many of the 10 nearest neighbours belong to the COVID-19** positive class, summing up their contribution by a **weighting function** that depends on the similarity and position in the neighbours' similarity ranking. This function decays logarithmically with the position and increases linearly with the similarity, so that the most similar cases account for most of the final score. In the function is also included a term to scale the score **in the range 0 to 1**.
- **Deep neural network training**: In the previous version of the manuscript, we mentioned the weights of deep neural network (DNN) model were transferred from the ones obtained by training the same architecture on ImageNet classification task. We would like to underline that fine-tuning had not been performed, so that the DNN weights used to compute performance were the ones from ImageNet. Reviewing the manuscript, we noticed that the usage of words "training set" and "testing set" can be misleading, making the Reader think of fine-tuning phase; for this reason, we substituted the wording "training set" with "reference set", from the content-based image retrieval task nomenclature.

- **Performance metrics:** Reviewer 1 suggested evaluating the performance of the proposed solution also by **comparing the severity score predicted by our solution and the one from the radiologists. This approach cannot be applied to our work**

because available dataset did not contain any information about the severity of the infection, but **only if the patient was infected by COVID-19 or not**, and, if not, a further distinction was made between cases affected by some other diseases and cases where are not present any radiological finding. Reviewer 2 requested more details about the **Mean Average Precision [MAP]** metric we used for performance calculation and why we interpreted higher values of this measure as indication of better case ranking. The proposed method elaborates a Chest X-ray picture **predicting its likelihood to be related to a COVID-19** positive case, that we indicated as COVID-19 risk indicator. Providing to it a set of cases, **we can sort it by the predicted score** and compute the MAP performance. MAP is an **evaluation measure for information retrieval system used to assess how well the system is performing in providing sorted lists of results to a given set of queries**; this metric assumes values in the range between 0 and 1, where 1 indicates that all the relevant results for the queries are placed on top of their lists. This interpretation comes from the way the metric is computed: **Precision** (in the context of information retrieval) of each relevant item in the sorted list of results will decrease depending on how many wrong results are placed before, the less this happen the higher will be the Precision and so the resulting **Average Precision** [AP] of the sorted list. MAP results as the mean of AP of the computed lists. In our work we used this metric to evaluate the goodness of the sorting induced by our prediction, when a set of cases are provided.

- **Usage of English language:** We would like to thank Reviewer 3 for the time spent evaluating our paper and for the indication to improve the use of English. To address this concern, the new version of the manuscript has been extensively revised with a heavy language revision, as can be observed by the document tracked changes. C. Platti, one of the authors previously not involved in the manuscript language revision, collaborated in this task. C. Platti has already revised the English language of several papers from our group, successfully accepted in some of the most influential journals of our discipline.

Finally, for a better readability we listed here the parts of the manuscript added or modified with reference to the concerns/recommendations arisen by the reviewers:

Recommendation 1.1: “This reviewer would like the authors to include the variable severity "score" in the patients classified as COVID-19 positive; the inclusion of this variable would be extremely valuable for the readers. Also, fundamentals are to mention the method they use to classify the severity”

- Modifications to section ***2.2.2 Performance metrics and statistical analysis***, to better explain the prioritization induced by our method and the way we evaluated it

- Added dedicated section ***2.2.3 COVID-19 risk indicator statistical analysis***

Recommendation 1.2: “A second matrix that can measure the performance besides MAP could be the correlation value between the calculated severity and the age of patients and compare these correlations values between the CNN and the senior radiologists.”

- Added dedicated section ***2.2.3 COVID-19 risk indicator statistical analysis***

- Added dedicated section ***3.2.1 COVID-19 risk indicator statistical analysis***

Recommendation 1.3: “There have been differences in correlation between age and severity of Italian and Mexican populations following the same methodology. Please include the reference:

Correlation between Chest X-Ray Severity in COVID-19 and Age in Mexican-Mestizo Patients: An Observational Cross-Sectional Study. Biomed Res Int. 2021 Apr 29;2021:5571144. doi: 10.1155/2021/5571144. PMID: 33997012; PMCID: PMC8090453.

Comment these findings in the discussion of you paper, because, then the CNN should be questioned if its classification methods calculated the same MAP in separed groups: male, females and age groups.”

- Added dedicated section ***2.2.3 COVID-19 risk indicator statistical analysis***

- Added dedicated section ***3.2.1 COVID-19 risk indicator statistical analysis***

Recommendation 1.4: “Authors have sufficient sample size, can they create some groups and compare groups to discard any differences in the classification power of the CNN?”

- Modifications to section ***2.2.2. Performance metrics and statistical analysis***, to introduce the methodology

- Modifications to section ***3.2 Phase I study***, to illustrate and interpret the results

Recommendation 1.5: “Considering to supplement or substitute Tables 2 and 3 with graphs or a more informative table with significance comparison (see examples of figures 2 and 3 from the recommended reference).”

- Modifications to section ***3.1 Baseline characteristics***, to introduce more details about data distribution across age and gender groups

- ***Table 2 and 3***, substituted with more informative tables, showing statistical testing on class membership comparison between age and gender groups

Recommendation 2.1: “The introduction is not enough and vague. Authors should include more related research of using CNN for early detection of COVID-19.”

- Modifications to section ***1. Introduction***, providing more references to the works in this topic and summarizing the different tasks tackled

Recommendation 2.2: “In Figure 1, DNN is responsible for extracting features from new images. Author mentioned the weights are transferred from the ImageNet task. So how authors fine-tune the model on CXR data? Please add more details here.”

- Modifications to section **2.1 Deep learning method**, to specify that fine-tune has not been performed

Recommendation 2.3: “In Figure 1, f represents the database to store previous extracted features. Are they the features extracted from ground truth of CXR images? Please explain.”

- Modifications to section **2.1 Deep learning method**, to better explain the content of database f) in Figure 1, and how it is used in the proposed approach

Recommendation 2.4: “In the preprocessing stage, how was the U-Net lung segmentation model be used here? If a segmentation model is used, the input should be a binary image but from Figure 1, the input of DNN seems to be an original image with only being copped.”

- Modifications to section **2.1 Deep learning method**, to describe in detail the preprocessing stage of our method

- Added **Figure 2**, to provide to the Reader a visual representation of the process steps

Recommendation 2.5: “How did authors finally decide which group the new CXR image is belong to? Is the similarity calculated based on group? Please add more details here.”

- Modifications to section **2.1 Deep learning method**, to describe in detail how the final score is calculated, exploiting the similarity between features

Recommendation 2.6: “In Table 1, MAP is calculated for each class but why the interpretation is: “higher MAP indicate the positive cases are places on top of the diagnostic queue“? Please explain more here.”

- Modifications to section **2.2.2 Performance metrics and statistical analysis**, to explain the MAP metrics and how we applied it to our use case.

[1] Correlation between Chest X-Ray Severity in COVID-19 and Age in Mexican-Mestizo Patients: An Observational Cross-Sectional Study. Biomed Res Int. 2021 Apr 29;2021:5571144. doi: 10.1155/2021/5571144. PMID: 33997012; PMCID: PMC8090453.