scientific reports

OPEN

MultiCOVID: a multi modal deep learning approach for COVID-19 diagnosis

Max Hardy-Werbin[1,3], José Maria Maiques[2], Marcos Busto[2], Isabel Cirera[3], Alfons Aguirre[3], Nieves Garcia-Gisbert[1], Flavio Zuccarino[2], Santiago Carbullanca[2], Luis Alexander Del Carpio[2], Didac Ramal[2], Ángel Gayete[2], Jordi Martínez-Roldan[4], Albert Marquez-Colome[5], Beatriz Bellosillo[1,6] & Joan Gibert[1,6]✉

The rapid spread of the severe acute respiratory syndrome coronavirus 2 led to a global overextension of healthcare. Both Chest X-rays (CXR) and blood test have been demonstrated to have predictive value on Coronavirus Disease 2019 (COVID-19) diagnosis on different prevalence scenarios. With the objective of improving and accelerating the diagnosis of COVID-19, a multi modal prediction algorithm (MultiCOVID) based on CXR and blood test was developed, to discriminate between COVID-19, Heart Failure and Non-COVID Pneumonia and healthy (Control) patients. This retrospective single-center study includes CXR and blood test obtained between January 2017 and May 2020. Multi modal prediction models were generated using opensource DL algorithms. Performance of the MultiCOVID algorithm was compared with interpretations from five experienced thoracic radiologists on 300 random test images using the McNemar–Bowker test. A total of 8578 samples from 6123 patients (mean age 66 ± 18 years of standard deviation, 3523 men) were evaluated across datasets. For the entire test set, the overall accuracy of MultiCOVID was 84%, with a mean AUC of 0.92 (0.89–0.94). For 300 random test images, overall accuracy of MultiCOVID was significantly higher (69.6%) compared with individual radiologists (range, 43.7–58.7%) and the consensus of all five radiologists (59.3%, *P* < .001). Overall, we have developed a multimodal deep learning algorithm, MultiCOVID, that discriminates among COVID-19, heart failure, non-COVID pneumonia and healthy patients using both CXR and blood test with a significantly better performance than experienced thoracic radiologists.

Abbreviations

DL	Deep learning
CXR	Chest X-rays
AUC	Area under the receiver operating characteristic curve
COVID-19	Coronavirus disease 2019
RT-PCR	Reverse-transcription polymerase chain reaction
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
HF	Heart failure
NCP	Non-COVID pneumonia

The outbreak of Coronavirus Disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), stroke the worldwide population with more than 200 million cases and 4.5 million deaths by August 2021. The rapid spread of the pandemic led to a global overexertion of health care and research facilities in order to counteract the growing rate of infection. However, a collapse of the sanitary system was imminent and inevitable worldwide, and new technologies were needed to speed up the diagnostic process.

The reference for COVID-19 diagnosis is the detection of SARS-CoV-2 viral RNA by real-time polymerase chain reaction (RT-PCR). However, the massive requests for sample processing at the beginning of the pandemic caused serious delays to obtain results.

[1]Cancer Research Program, IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain. [2]Radiology Department, Hospital del Mar, Barcelona, Spain. [3]Emergency Department, Hospital del Mar, Barcelona, Spain. [4]Innovation and Digital Transformation Department, Hospital del Mar, Barcelona, Spain. [5]Information Systems Department, Hospital del Mar, Barcelona, Spain. [6]Pathology Department, Hospital del Mar, Barcelona, Spain. ✉email: jgibert@imim.es

As lung involvement is one of the main causes of morbidity and mortality in SARS-CoV-2 infection, a quick identification of characteristic findings in chest imaging can support the diagnosis and speed up the identification of COVID-19 positive patients at the emergency units.

Several studies have shown that implementation of deep learning (DL) tools to detect chest X-rays (CXR) findings typically associated with SARS-CoV-2 infection, deliver comparable results to those acquired by interpretation of radiologists. However, most of the trained models have a drop in their prediction performance when tested over external datasets[1]. In addition, one of the main hurdles to overcome when training an algorithm to detect Sars-CoV-2 infection in CXR is the similarity of findings with other entities like bacterial pneumonias or heart failure[2]. On the other hand, models based on laboratory results of peripheral blood also give predictive results on diagnosis[3] and prognosis[4].

A key fact to highlight is how the incursion of COVID-19 caused a dramatic drop in the emergency room consultations of other pathologies. Later on, after the initial peak, the decline of the COVID-19 prevalence made the non-COVID diseases emerge once again at the hospitals. This is relevant due to the challenge of performing an efficient differential diagnosis with selected pathologies during a pandemic. It is well known that the predictive value of a diagnostic test is conditioned by the prevalence of the disease and that of COVID varies widely throughout the different waves of the pandemic[5]. A multicategory approach that takes into account differential diagnoses that are more stable in their prevalence could reduce this variability.

With the objective of improving and accelerating the diagnosis of COVID-19, we developed a tool to assist physicians in reaching a diagnosis. This tool is a multi-modal prediction algorithm (MultiCOVID) based on CXR and blood test with the ability to discriminate between COVID-19, Heart Failure (HF), Non-COVID Pneumonia (NCP) and healthy (Control) samples.

Materials and methods
Dataset
We retrospectively collected CXR images and hemogram values from 8578 samples from 6123 patients and healthy subjects (mean age 66 ± 18 years of standard deviation, 3523 men) from Parc Salut Mar (PSMAR) Consortium, Barcelona, Spain. Four cohorts were designed: (i) 1171 samples from patients diagnosed with COVID-19 by RT-PCR from March to May 2020; (ii) 1008 samples of patients who suffered an episode of heart failure between 2012 to 2019; (iii) 490 samples of patients diagnosed with non-COVID pneumonia (NCP) from 2018 to 2019; (iv) 5909 samples of standard preoperatory studies of healthy subjects from 2017 to 2019 (Fig. 1). HR and NCP diagnosis were selected as defined by the International Classification of Diseases, Tenth Revision (ICD-10) code. All the CXR images from groups i-iii were validated by two independent radiologists (MB and JM).
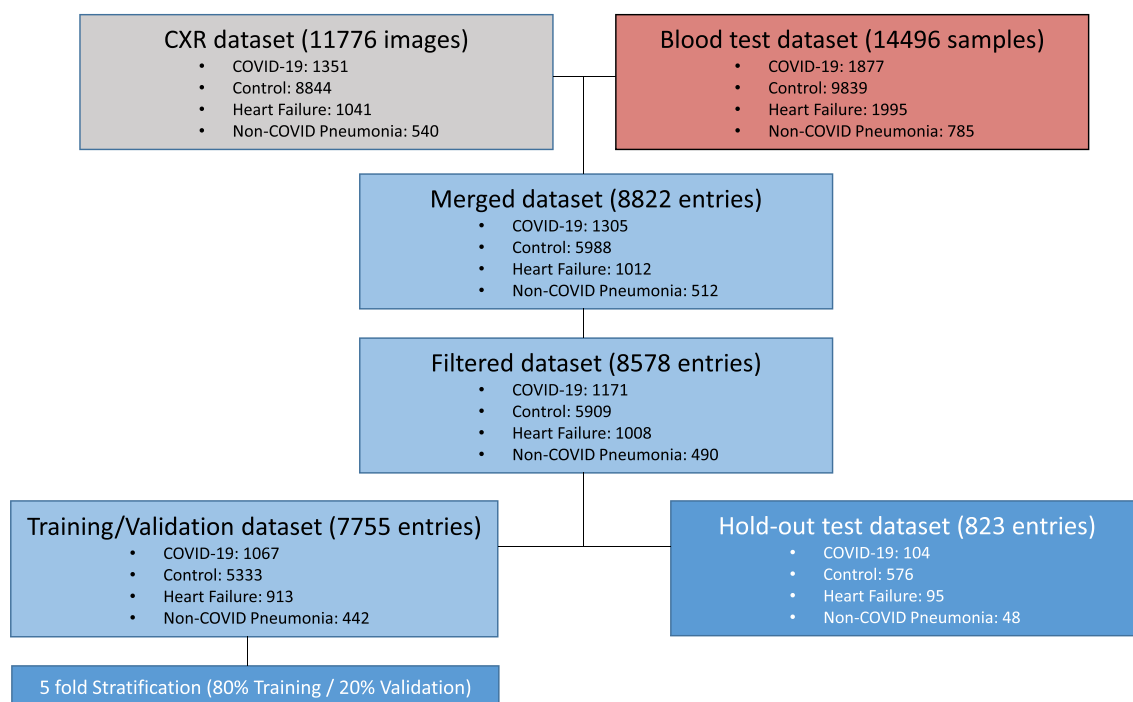

Figure 1. Flowchart for sample selection and patient inclusion in the study and breakdown of training, validation, and hold-out test data sets. Around 25,000 entries were obtained using both CXR images and blood test in a time wise manner. The whole dataset totals 8822 entries of paired CXR and blood test data. Samples with low completeness (less than 80% of blood test data available) were discarded for the model building.

Acquisition of blood sample and image data

We included CXR images performed in a period ranging from 1 day before the patient's diagnosis to 7 days after. The images were filtered to include only frontal projections regardless of the quality and the radiography system used. Blood sample results were collected within a range of 2 days before or 7 days after the CXR acquisition date using PSMAR lab record system, except for control samples whose measurements ranged for 2 weeks. If two or more blood test results were collected, measurements were averaged.

CXR images and blood test results were combined in the same dataset and split into train/validation set (90%), and hold-out test (10%) set. For training/validation split, we divided the dataset in training (80%) and validation (20%) sets with 5 different random seeds. We ensured that there were no cross-over patients between groups.

Deep learning models

Detailed description of the models, training policy and image preprocessing are provided in Supplementary Material. In brief, segmentation model is based on a U-Net architecture[6]. The CXR-only classification model consists of a validated Convolutional neural network (CNN) resnet-34 architecture[7]. Tabular only-model is an Attention-based network (TabNet)[8]. Joint model is a multi-modal deep learning algorithm which merges the CXR-only and the Blood-only models and uses both CXR image and blood tests as input values. It uses Gradient Blending in order to prevent overfitting and improve generalization[9]. MultiCOVID model is an ensemble predictor of 5 different Joint models that would classify independently between the different classes. Then it uses majority vote to assign a final classification. The whole pipeline development and training was performed using fastai deep learning API[10].

Comparison with thoracic radiologist interpretations

Hold-out test dataset consisting of 300 samples (ensuring no patient overlap with training or validation sets) was used for expert interpretation. Each sample consisted of a CXR with matched blood results. Expert interpretations were independently provided by five board-certified thoracic radiologists (FZ, SC, LdC, DR, AG) with 2–30 years post-residency training experience. Radiologists were able to check both non segmented images and blood test results without any other additional information in a platform created ad-hoc for prediction. They provided a classification for each image in one of the four categories (COVID-19, control, HF and NCP). A consensus interpretation for the radiologist was obtained by the majority vote for each paired CHX-blood test analyzed.

Statistical analysis

A two-tailed t-test P value was reported when clinical and population blood test differences were assessed. McNemar–Bowker test was used to compare model performance against radiologist majority vote using FDR correction. Plotting and statistical analyses were performed using the packages ggplot, ggpubr and rcompanion in R, version 3.6 (R Core Team; R Foundation for Statistical Computing).

Ethical approval

The study was designed to use radiology images and associated clinical/demographic/ laboratory patient information already collected for the purpose of performing clinical COVID-19 research by Hospital del Mar. The study was conducted in accordance with the relevant institutional guidelines and regulations. The experimental protocols, data acquisition and analysis were approved by the Parc de Salut Mar Clinical Research Ethics Committee (2020/9199/I). Informed consent was obtained, when possible, from patients or legal representatives or waived by the local Parc de Salut Mar Clinical Research Ethics Committee (2020/9199/I) if informed consent was not available due to the pandemic situation.

Results

Patient characteristics

A total of 8578 samples were evaluated across datasets. Patient characteristics and blood test parameters are shown in Table 1. A highly significant difference in age was found between the cohort of patients with heart failure (82.8 ± 10 years) and the other three cohorts (66.0 ± 16 years for COVID-19 samples, 63.2 ± 18 years for control samples and 67.8 ± 17 years for NCP samples, $P < 0.001$ for each comparison) and was not considered as a valid variable for further classification.

Whole CXR models learn spurious characteristics for classification

Previous studies have demonstrated that deep learning (DL)-based algorithms should be rigorously evaluated due to their ability to learn non relevant features in order to increase its prediction accuracy[1]. For this reason, we first developed a segmentation algorithm able to segment lung parenchyma at a 95%-pixel accuracy. Then, after segmentation, we evaluated the accuracy of the algorithms for three complementary datasets: non-segmented images, segmented regions and excluded regions. After a few training epochs the three different models achieved nonrandom accuracies between 67 and 74% (Fig. 2A). However, attention map exploration on the images showed that the different models based their predictions not only inside but also outside of the lung parenchyma (Fig. 2B).

These observations showed that, although there are important features outside the lung parenchyma that may help the model to classify between the different entities (eg. heart size), there are other elements (eg. oxygen nasal cannulas or intravenous (IV) catheters) that might confound the model. Thus, we decided to first segment all the CXR before training our models for prediction of diagnosis. In order to accomplish this task, we generated a 785-radiology level lung segmentation dataset and trained a U-net model to regenerate the whole CXR dataset keeping only the lung parenchyma.

Characteristic		Overall	COVID-19	Control	HF	NCP
n		8578	1171	5909	1008	490
Age, mean (SD)	[Years]	66.128 (18.353)	66.013 (16.612)	63.165 (18.281)	82.823 (10.731)	67.786 (17.011)
Sex, n (%)	H	5023 (58.557)	677 (57.814)	3598 (60.890)	460 (45.635)	288 (58.776)
	M	3555 (41.443)	494 (42.186)	2311 (39.110)	548 (54.365)	202 (41.224)
% basophils	[%]	0.300 [0.200,0.525]	0.200 [0.100,0.300]	0.400 [0.200,0.600]	0.333 [0.200,0.500]	0.300 [0.175,0.500]
Total basophils	[×10³/μL]	0.030 [0.020,0.050]	0.010 [0.010,0.020]	0.040 [0.020,0.055]	0.030 [0.020,0.050]	0.030 [0.017,0.060]
% eosinophils	[%]	0.700 [0.100,1.900]	0.000 [0.000,0.300]	0.950 [0.175,2.150]	0.900 [0.200,2.100]	0.600 [0.000,2.200]
Total eosinophils	[× 10³/μL]	0.060 [0.010,0.160]	0.000 [0.000,0.020]	0.080 [0.020,0.180]	0.070 [0.020,0.160]	0.060 [0.005,0.200]
MCH	[pg]	29.650 [28.300,30.900]	29.400 [28.300,30.550]	29.800 [28.433,31.050]	29.200 [27.387,30.700]	29.600 [28.500,30.700]
Hematrocrit	[%]	38.000 [33.100,42.250]	40.000 [36.600,43.300]	38.500 [33.600,42.700]	35.200 [31.400,39.200]	32.025 [28.512,36.100]
Red Blood Cells	[×10⁶/μL]	4.290 [3.710,4.809]	4.550 [4.080,4.935]	4.360 [3.743,4.860]	3.930 [3.494,4.370]	3.630 [3.160,4.070]
Hemoglobin	[g/dL]	12.600 [10.800,14.150]	13.300 [12.000,14.433]	12.900 [11.067,14.400]	11.300 [10.000,12.700]	10.500 [9.200,12.100]
Leukocytes	[×10³/μL]	9.020 [6.771,12.100]	6.420 [5.060,8.860]	9.450 [7.260,12.655]	8.660 [6.894,11.072]	10.950 [7.700,14.781]
% lymphocytes	[%]	15.350 [8.333,25.100]	14.900 [9.475,21.900]	16.800 [8.400,27.600]	12.700 [8.200,18.912]	10.100 [5.763,17.850]
Totallymphocytes	[×10³/μL]	1.310 [0.800,2.010]	0.930 [0.688,1.270]	1.525 [0.920,2.240]	1.070 [0.740,1.560]	1.140 [0.605,1.680]
MCHC	[g/dL]	33.233 [32.200,34.167]	33.000 [32.100,33.900]	33.467 [32.500,34.400]	32.300 [31.300,33.200]	32.717 [31.600,33.750]
% monocytes	[%]	7.100 [5.300,8.900]	6.500 [4.400,9.000]	7.100 [5.400,8.800]	8.000 [6.237,9.812]	6.300 [4.308,8.600]
Total monocytes	[×10³/μL]	0.635 [0.442,0.840]	0.415 [0.290,0.600]	0.660 [0.480,0.870]	0.690 [0.510,0.890]	0.690 [0.440,0.890]
% neutrophils	[%]	75.200 [63.600,84.400]	77.000 [69.100,84.600]	73.467 [60.900,84.100]	76.800 [69.183,82.900]	81.050 [70.925,89.100]
Total neutrophils	[×10³/μL]	6.450 [4.400,9.650]	4.870 [3.470,7.095]	6.673 [4.520,10.080]	6.470 [4.950,8.654]	8.418 [5.500,12.360]
P-LCR	[%]	30.700 [25.167,36.900]	30.800 [25.300,36.300]	30.317 [24.950,36.500]	32.858 [26.992,38.400]	32.300 [24.000,40.413]
PDW	[fL]	12.600 [11.100,14.400]	12.600 [11.075,14.200]	12.500 [11.100,14.300]	13.000 [11.400,14.900]	12.850 [10.800,15.400]
Platelets	[×10^3/μL]	222.333 [173.000,281.000]	201.000 [157.500,267.000]	227.000 [178.000,280.333]	215.500 [168.000,269.000]	229.500 [161.250,373.625]
RDW-CV	[%]	13.900 [13.050,15.300]	13.300 [12.600,14.100]	13.750 [13.000,15.100]	15.325 [14.350,17.100]	14.800 [13.813,16.275]
RDW-SD	[fL]	45.100 [41.700,49.650]	43.200 [40.600,46.225]	44.600 [41.337,48.850]	49.833 [46.300,54.975]	48.100 [45.100,52.425]
MCV	[fL]	89.000 [85.221,92.700]	88.800 [85.700,92.300]	88.800 [85.000,92.400]	89.950 [85.300,94.263]	90.250 [86.806,93.681]
MPV	[fL]	10.700 [10.000,11.500]	10.700 [10.100,11.450]	10.667 [10.000,11.400]	11.000 [10.300,11.700]	10.900 [9.950,11.950]

Table 1. Patient characteristics. COVID-19, coronavirus disease 2019; HF, heart failure; NCP, non-COVID pneumonia; MHC, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; P-LCR, platelet-large cell ratio; PDW, platelet distribution width; RDW, red cell distribution width; MCV, Mean corpuscular volume; MPV, mean platelet volume.

Performance of single and multimodal models

In order to evaluate the prediction capacity of both segmented CXR and blood sample data, we built different DL models using both sources alone or in combination. Metrics comparison of all the single vision (CXR-only) and tabular (Blood-only) models are detailed in Supplementary Material. As expected, CXR-only models had a more robust prediction of all 4 categories tested compared to Blood-only models (Fig. 3). This difference is stronger in the classes with less samples (HF, and NCP) where CXR-only models could identify features in the CXR images which are characteristic of these two entities whereas this was not possible with Blood-only models.
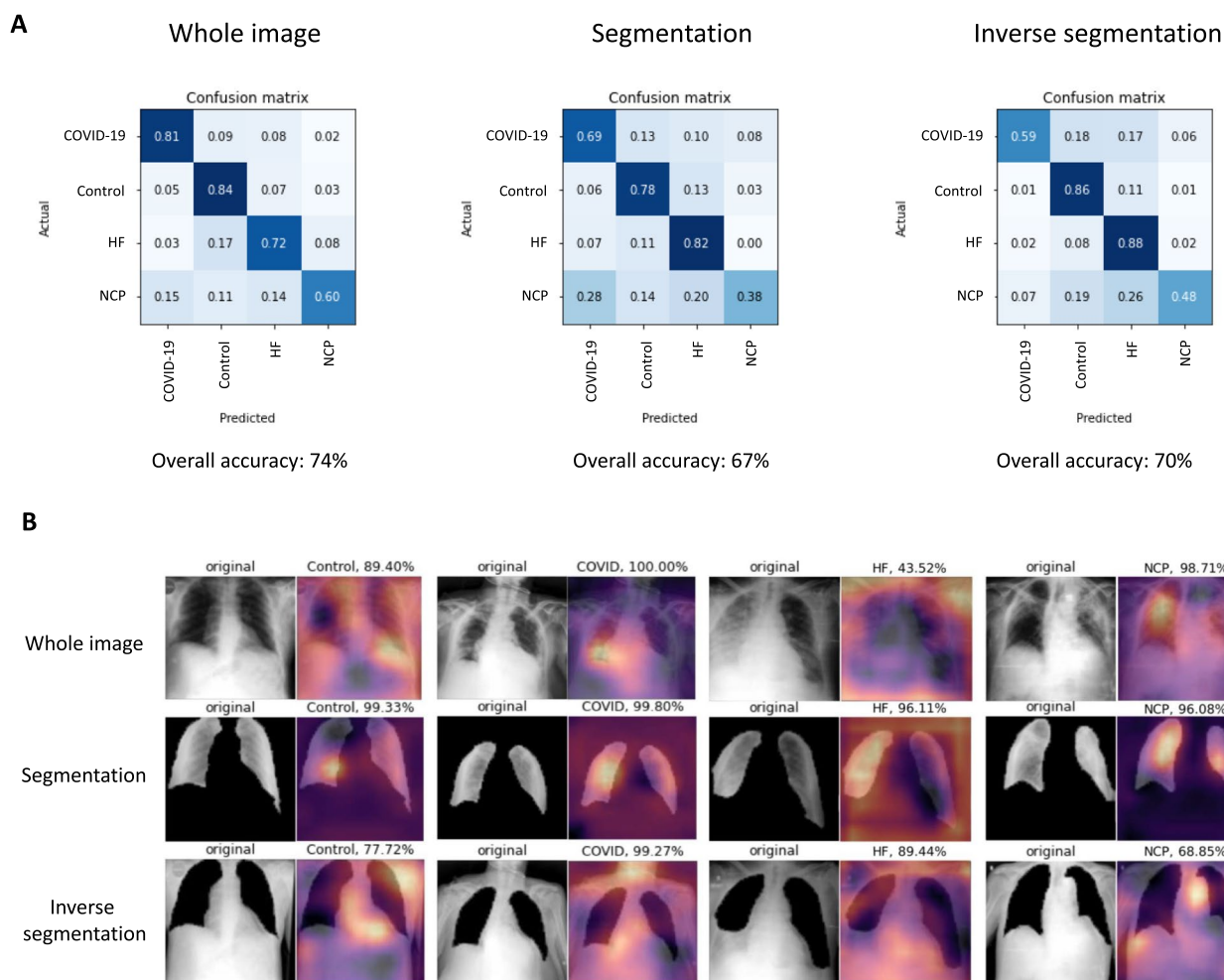
Figure 2. Performance of visual models on whole CXR images. (**A**) Confusion matrix and overall accuracy using whole image, segmented and inverse segmented images, respectively for each category tested. (**B**) Raw image and Grad-CAM heatmap representation of an image for each category and model trained.

Model interpretability of Blood-only models by analyzing feature importance using Shapley Additive explanations[12] showed that patient classification was related to two different axes: the immune compartment and the red blood cell (RBC) compartment, respectively (Fig. 4A). The first axis seems to be strongly associated with COVID-19 classification and shows a specific signature looking at the blood counts (Fig. 4B-top). However, the second axis seems to subdivide patients between COVID-19/Control and HF/NCP, although COVID-19 blood counts seems to be statistically different from Control samples, too (Fig. 4B-bottom).

The combination of CXR and blood tests using multimodal models that combine inputs from tabular and image data to perform a global prediction, slightly increased the prediction capacity of the single models even when DL tabular models are worse than machine learning (ML—XGBoost) models alone (Supplementary Table 1). This underpins the concept that adding new sources of information to the data could increase the ability of the models to generate better predictions [13]. Moreover, the joint approach used for building Multi-COVID algorithm resulted on an improved performance in the majority of the metrics analyzed (Fig. 3 and Supplementary Table 1).

Comparison with expert thoracic radiologists

Finally, we compared the performance of MultiCOVID algorithm with the interpretation of expert chest radiologists. This comparison was performed with 300 CXR randomly selected from the hold-out test set that were independently reviewed by 5 radiologists together with the blood test results. The independent results from radiologists showed an accuracy ranging from 43.7 to 58.7%. This value rose to 59.3% (178/300) when the consensus interpretation of all 5 radiologists based on the majority vote was considered. Of note, the overall accuracy achieved by MultiCOVID was 69.6% (209/300) that was significantly higher than consensus interpretation ($P < 0.001$). In addition, for COVID-19 prediction individually, MultiCOVID showed similar sensitivity to the radiologists' consensus but with a much higher specificity, leading to significantly better performance when discerning between COVID-19 versus Control and COVID-19 vs HF patients ($P < 0.05$ for both comparisons; Fig. 5).
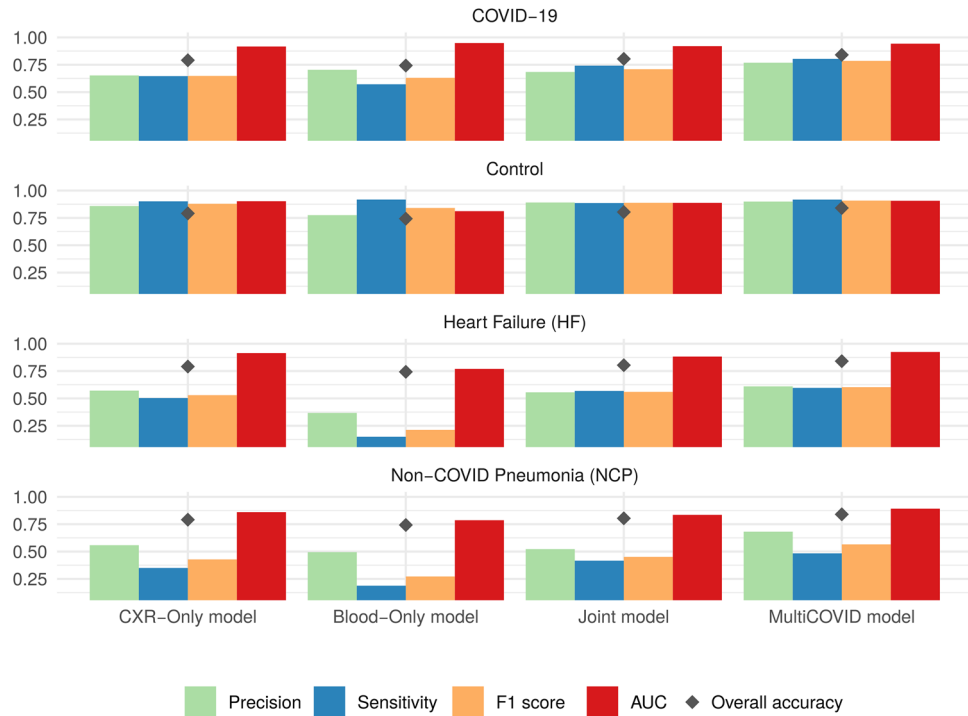
Figure 3. Performance of different models on the entries from hold-out test datasets. Means for precision (green), sensitivity (blue), F1 score (yellow), AUC (red) and accuracy (black diamond) for each model type and category assessed, respectively. CXR-only models use only CXR images for 4 category classification. Blood-only models use blood test a source of information. Joint model uses both CXR and blood test as input for classification and MultiCOVID is the majority vote of 5 different Joint models.
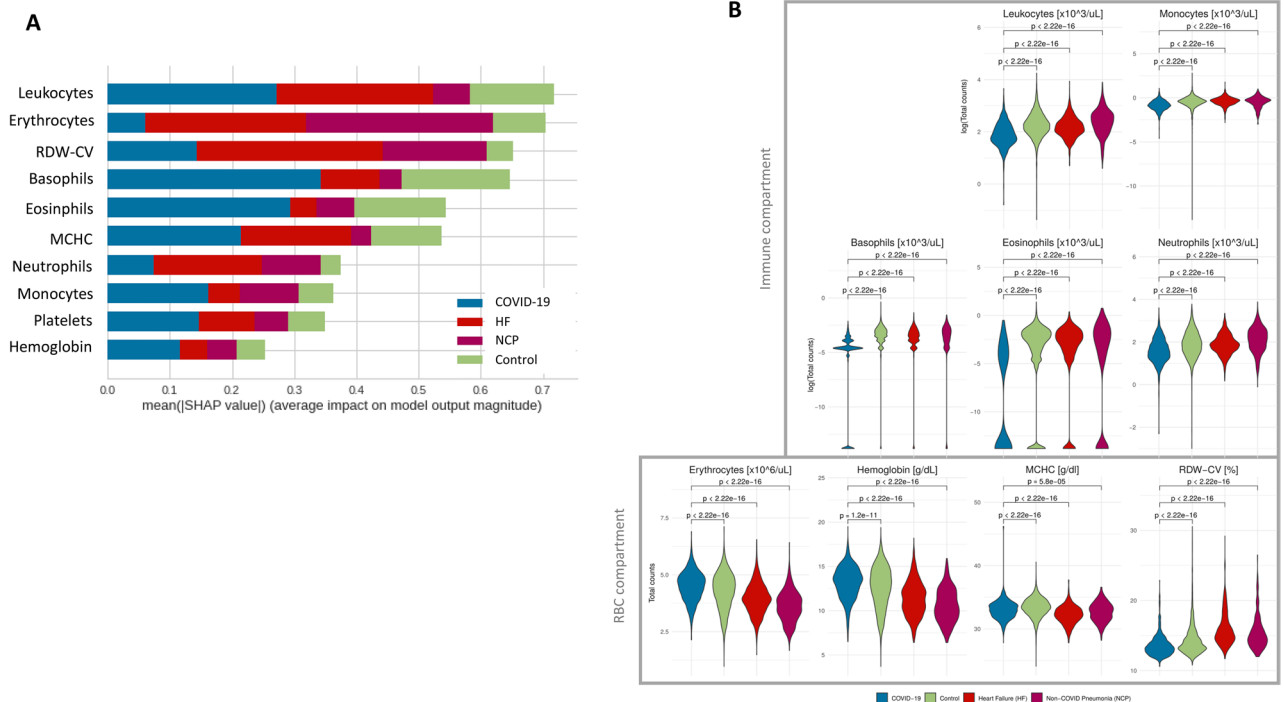

Figure 4. Blood-only model interpretability by SHAP analysis. (**A**) Summary plot showing the mean absolute SHAP value of the ten most important features for the four classes. (**B**) Blood test values of the different features identified by SHAP analysis. RDW-CV: red cell distribution width; MCHC: Mean Corpuscular Hemoglobin Concentration; RBC: red blood cells.
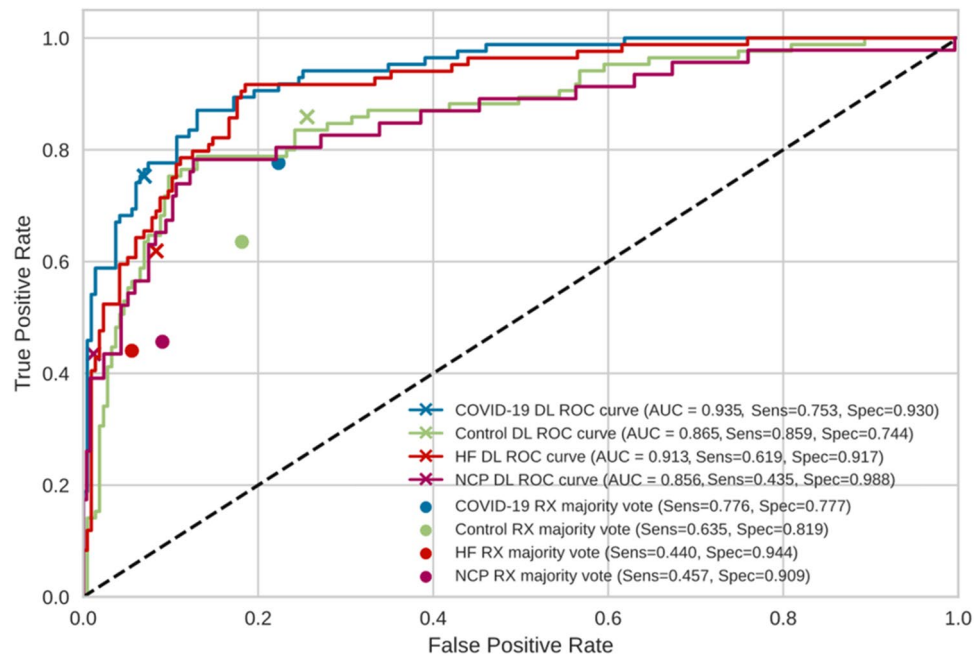
Figure 5. Comparison of the performance of MultiCOVID model with consensus expert radiologist interpretations on random sample of 300 images from the test set. The receiver operating characteristic (ROC) curves for each category (COVID-19 – blue; Control – green; Heart Failure (HF) – red and Non-COVID Pneumonia (NCP) – magenta) are shown for MultiCOVID (DL) and for the consensus interpretation of radiologists (majority vote). Sensitivity (Sens) and specificity (Spec) are also plotted for each category assessed. DL: deep learning.

Discussion

Diagnosis of COVID-19 is an evolving challenge. During the beginning of the pandemic and the successive peaks with high prevalence rates, a prompt and effective diagnosis was critical for proper patient isolation and evaluation. However, since the prevalence of the COVID-19 cases oscillated, showing fewer cases between waves, and more non-COVID cases, it was important to differentiate patients with other diseases than COVID-19 presenting similar visual characteristics in the CXR.

During patient assessment in the emergency room, clinicians take into account different inputs for a proper diagnosis. First, the anamnesis, symptoms, vitals and physical findings guide the physician to an initial assumption. Based on this information, additional tests are requested (CXR, blood test, ECG and SARS-CoV-2 detection). The integration of these results allows the team to diagnose a patient accurately. However, this process is time consuming and sometimes findings are difficult to interpret, leading to misdiagnosis.

To improve this diagnostic process, we have developed and trained a multimodal deep learning algorithm based in a multiple input approach combining CXR images together with blood sample data to identify COVID-19 diagnosis with high sensitivity. This way we were able to manage the increased complexity of the dataset. These data from multiple sources are somehow correlated and complementary to each other and could reflect patterns that are not present in single models alone[13].

Hence, MultiCOVID is fed by two of the most common and fast clinical tests requested in the emergency room (CXR and Blood test) and can predict the presence of three different diseases (COVID-19, heart failure and non-COVID pneumonia) with similar CXR characteristics.

Analysis of single models shows the importance of model interpretation. While CXR-only models could identify patterns outside the lung parenchyma that could diminish its generalization capacity[9], Blood-only models could point to interesting population of cells that are differently represented in COVID-19 patients, leveraging its prediction capacity. In this context, the immune compartment plays an important role in the COVID-19 response, and it has been already published that COVID-19 patients present fewer overall leukocytes counts and, more concretely, eosinophil counts[14, 15]. Furthermore, oxygen transport seems to be somehow affected, modulating the red cell population. In this regard, in our work we found significant differences in the erythrocyte count and the hemoglobin concentration. Although most of the studies correlate the reduction of this values to severe COVID-19 patients[16], this is the first dataset to compare them in these four different categories at the time of diagnosis.

Moreover, although a huge amount of literature about COVID-19 diagnosis and prognosis has been published using only blood tests[17–20] or CXR[21–28] this is the first study that combines both parameters and compares its prediction capacity at diagnosis. Of note, only one previously published study integrates both blood test and CXR severity scores in order to determine in-hospital death of COVID-19 patients[29]. Hence, it is clear that merging both sources of data leads to a better prediction performance when compared with the two single models alone and that this difference is more pronounced where the number of cases is scarce. It is important to stress that

this combination of data sources addresses the variable prevalence of COVID-19 cases during the pandemic, which is an issue that could not be solved in previous studies[23, 24].

Our study has several limitations. First, the algorithm was evaluated on a single center; thus, there was likely some degree of bias. Additionally, the sample collection was performed in different time periods for each group of patients, which could present some kind of differences in the CXR image acquisition although this was partially solved using the lung segmentation model which removes the noise signal present outside the lung parenchyma. And finally, model performance could be influenced by potential shifts in the disease landscape due to COVID-19 variants and vaccination efforts, which could influence the generalizability and interpretation of our findings.

Conclusions

We have developed a multimodal deep learning algorithm, MultiCOVID, that discriminates among COVID-19, heart failure, non-COVID pneumonia and healthy patients using both CXR and blood test with a significantly better performance than experienced thoracic radiologists.

Our approach and results suggest an innovative scenario where COVID-19 prediction could be identified from other similar diseases and facilitate triage within the emergency room in a COVID-19 low prevalence situation.

Data availability

Our code base is provided on GitHub at https://github.com/Tato14/MultiCOVID, including weights for each of the individually trained neural network architectures and respective model weights for the weighted ensemble model. The datasets used and analyzed during the current study will be available from the corresponding author on reasonable request. In order to correct samples bias[11], additional metadata information present in the DICOM image headers from the CXR would be also available upon request.

References

1. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
2. Cleverley, J., Piper, J. & Jones, M. M. The role of chest radiography in confirming covid-19 pneumonia. *BMJ* https://doi.org/10.1136/bmj.m2426 (2020).
3. Avila, E., Kahmann, A., Alho, C. & Dorn, M. Hemogram data as a tool for decision-making in COVID-19 management: Applications to resource scarcity scenarios. *PeerJ* **8**, e9482 (2020).
4. Razavian, N. *et al.* A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. *npj Digit. Med.* **3**, 130 (2020).
5. Trevethan, R. Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Front. Public Heal.* **5**, 307 (2017).
6. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation* 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28.
7. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016). https://doi.org/10.1109/CVPR.2016.90.
8. Arik, S. Ö. & Pfister, T. TabNet: Attentive interpretable tabular learning. *Proc. AAAI Conf. Artif. Intell.* **35**(8), 6679–6687 (2021).
9. Wang, W., Tran, D. & Feiszli, M. What makes training multi-modal classification networks hard?. In *Proceedings / IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 12692–12702 (2019) https://doi.org/10.1109/CVPR42600.2020.01271.
10. Howard, J. & Gugger, S. Fastai: A layered API for deep learning. *Information* **11**, 108 (2020).
11. Garcia Santa Cruz, B., Bossa, M. N., Sölter, J. & Husch, A. D. Public Covid-19 X-ray datasets and their impact on model bias—A systematic review of a significant problem. *Med. Image Anal.* **74**, 102225 (2021).
12. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
13. Ngiam, J. *et al.* Multimodal deep learning. *ICML* (2011).
14. Tan, Y., Zhou, J., Zhou, Q., Hu, L. & Long, Y. Role of eosinophils in the diagnosis and prognostic evaluation of COVID-19. *J. Med. Virol.* **93**, 1105–1110 (2021).
15. Rahman, A. *et al.* Hematological abnormalities in COVID-19: A narrative review. *Am. J. Trop. Med. Hyg.* **104**, 1188–1201 (2021).
16. Lippi, G. & Mattiuzzi, C. Hemoglobin value may be decreased in patients with severe coronavirus disease 2019. *Hematol. Transfus. Cell Ther.* **42**, 116–117 (2020).
17. Kukar, M. *et al.* COVID-19 diagnosis by routine blood tests using machine learning. *Sci. Rep.* **11**, 10738 (2021).
18. Bayat, V. *et al.* A Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) prediction model from standard laboratory tests. *Clin. Infect. Dis.* **73**, e2901–e2907 (2021).
19. Soltan, A. A. S. *et al.* Rapid triage for COVID-19 using routine clinical data for patients attending hospital: Development and prospective validation of an artificial intelligence screening test. *Lancet Digit. Heal.* **3**, e78–e87 (2021).
20. Chen, J. *et al.* Distinguishing between COVID-19 and influenza during the early stages by measurement of peripheral blood parameters. *J. Med. Virol.* **93**, 1029–1037 (2021).
21. Hwang, E. J. *et al.* Deep learning for chest radiograph diagnosis in the emergency department. *Radiology* **293**, 573–580 (2019).
22. Wang, L., Lin, Z. Q. & Wong, A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci. Rep.* **10**, 19549 (2020).
23. Wehbe, R. M. *et al.* DeepCOVID-XR: An artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large U.S. clinical data set. *Radiology* **299**, E167–E176 (2021).
24. Zhang, R. *et al.* Diagnosis of coronavirus disease 2019 pneumonia by using chest radiography: Value of artificial intelligence. *Radiology* **298**, E88–E97 (2021).
25. Baikpour, M. *et al.* Role of a chest x-ray severity score in a multivariable predictive model for mortality in patients with COVID-19: A single-center, retrospective study. *J. Clin. Med.* **11**, 2157 (2022).
26. Nishio, M. *et al.* Deep learning model for the automatic classification of COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy: A multi-center retrospective study. *Sci. Rep.* **12**, 8214 (2022).

27. Sun, Y. *et al.* Use of machine learning to assess the prognostic utility of radiomic features for in-hospital COVID-19 mortality. *Sci. Rep.* **13**, 7318 (2023).
28. Nishio, M., Noguchi, S., Matsuo, H. & Murakami, T. Automatic classification between COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy on chest X-ray image: Combination of data augmentation methods. *Sci. Rep.* **10**, 17532 (2020).
29. Garrafa, E. *et al.* Early prediction of in-hospital death of COVID-19 patients: A machine-learning model based on age, blood analyses, and chest x-ray score. *Elife* **10**, e70640 (2021).

Author contributions

M.H.-W.: Data curation , Validation, Formal analysis, Investigation, Project administration, Supervision, Roles/Writing—original draft, Writing—review & editing; J.M.M.: Data curation , Formal analysis, Investigation, Validation, Roles/Writing—original draft, Writing—review & editing; M.B.: Data curation, Formal analysis, Investigation Validation, Roles/Writing—original draft, Writing—review & editing; I.C.: Data curation, Validation, Roles/Writing—original draft, Writing—review & editing; A.A.: Data curation, Validation, Roles/Writing—original draft, Writing—review & editing; N.G.-G.: Investigation, Visualization, Project administration, Roles/Writing—original draft, Writing—review & editin; F.Z.: Validation, Roles/Writing—original draft, Writing—review & editing; S.C.: Validation, Roles/Writing—original draft, Writing—review & editing, L.A.D.C.: Validation, Roles/Writing—original draft, Writing—review & editing, D.R.: Validation, Roles/Writing—original draft, Writing—review & editing; Á.G.: Validation, Roles/Writing—original draft, Writing—review & editin; J.M.-R.: Project administration, Supervision, Roles/Writing—original draft, Writing—review & editing; A.M.-C.: Data curation, Project administration, Supervision, Roles/Writing—original draft, Writing—review & editing; B.B.: Data curation, Formal analysis, Investigation, Project administration, Supervision, Roles/Writing—original draft, Writing—review & editin; J.G.: Data curation, Formal analysis; Investigation, Visualization, Project administration, Supervision, Roles/Writing—original draft, Writing—review & editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-46126-8.

Correspondence and requests for materials should be addressed to J.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.