scientific reports

Check for updates

OPEN

Machine learning is the key to diagnose COVID-19: a proof-of-concept study

Cedric Gangloff[1]✉, Sonia Rafi[1], Guillaume Bouzillé[1], Louis Soulat[2] & Marc Cuggia[1]

The reverse transcription-polymerase chain reaction (RT-PCR) assay is the accepted standard for coronavirus disease 2019 (COVID-19) diagnosis. As any test, RT-PCR provides false negative results that can be rectified by clinicians by confronting clinical, biological and imaging data. The combination of RT-PCR and chest-CT could improve diagnosis performance, but this would requires considerable resources for its rapid use in all patients with suspected COVID-19. The potential contribution of machine learning in this situation has not been fully evaluated. The objective of this study was to develop and evaluate machine learning models using routine clinical and laboratory data to improve the performance of RT-PCR and chest-CT for COVID-19 diagnosis among post-emergency hospitalized patients. All adults admitted to the ED for suspected COVID-19, and then hospitalized at Rennes academic hospital, France, between March 20, 2020 and May 5, 2020 were included in the study. Three model types were created: logistic regression, random forest, and neural network. Each model was trained to diagnose COVID-19 using different sets of variables. Area under the receiving operator characteristics curve (AUC) was the primary outcome to evaluate model's performances. 536 patients were included in the study: 106 in the COVID group, 430 in the NOT-COVID group. The AUC values of chest-CT and RT-PCR increased from 0.778 to 0.892 and from 0.852 to 0.930, respectively, with the contribution of machine learning. After generalization, machine learning models will allow increasing chest-CT and RT-PCR performances for COVID-19 diagnosis.

The severe acute respiratory syndrome coronavirus 2 (SARS-coV-2) outbreak started in December 2019 in the Hubei province, China. The associated disease, coronavirus disease 2019 (COVID-19)[1], has now spread worldwide. The World Health Organization currently reports more than 10 million confirmed cases and 500,000 deaths. Increased mortality rates and the collapse of healthcare systems have been reported in several regions[2–4]. Indeed, due to SARS-coV-2 contagiousness, promiscuity within health systems can promote patient-to-patient transmission[5,6] and the contamination of healthcare workers[7], rapidly leading to the saturation of health systems[8]. To limit this effect, patients with COVID-19 infection are hospitalized in specific units after being emergency department (ED) triage[9]. Therefore, it is essential to have a reliable and easy-to-use tool for COVID-19 diagnosis. SARS-coV-2 real-time RT-PCR reverse transcription–polymerase chain reaction (RT-PCR) is the accepted standard for COVID-19 diagnosis[10]. However, RT-PCR performances are sub-optimal and, like for any other test, there are false negatives results[11,12]. Therefore, additional investigations should be performed in patients with negative RT-PCR results but high clinical probability of COVID-19. In this context, chest-CT is an interesting tool because it allows detecting virus-induced lung tissue damages and alternative diagnoses[13]. Thus, when a patient presents a high clinical probability of COVID-19, a negative RT-PCR and a chest-CT showing typical COVID-19 lesions with no sign of alternative diagnosis, it is possible to consider that the patient has COVID-19 with a false negative RT-PCR result. The use of chest-CT alone cannot be recommended, but its combined use with clinic and RT-PCR allows to resolve diagnostic ambiguities[14]. However, RT-PCR and chest-CT cannot be performed in all patients suspected to have COVID-19 for many reasons, including reagent shortage[15], device unavailability, lack of human resources, and high costs. Moreover, the time required to perform both tests increase the risk of ED overcrowding by patients waiting for their results. Therefore, health professionals must adapt their diagnostic strategies in function of their resources[16]. To our knowledge, the potential contribution of machine learning using imaging, clinical and laboratory data has been poorly evaluated in this context. Machine learning is an inherited artificial intelligence approach that enables computers to extract or classify patterns. It allows predicting whether a patient belongs to a predefined group using explanatory variables. The recent increase in machine learning models in the healthcare field suggests that these methods could improve the COVID-19

[1]LTSI Laboratory, INSERM U1099, Université de Rennes 1, Rennes, France. [2]Department of Emergency Medicine, Pontchaillou University Hospital, 35033 Rennes, France. ✉email: cedric.ganglof@gmail.com

diagnostic strategy[17]. The objective of this study was to develop and evaluate machine learning models using clinico-biological data from health records to improve the RT-PCR and chest-CT performances for COVID-19 diagnosis among post-emergency hospitalized patients.

Materials and methods

This study protocol was approved by the Medical Ethics Committee of Rennes academic hospital (approval number 0020.93 issued on July 7, 2020). All methods were performed in accordance with the relevant guidelines and regulations. Authorization to conduct research from the Clinical Data Warehouse of Hospital of Rennes was given by CNIL—Commission Nationale Informatique et Liberté (Authorization number 2020-028 issued on February 27, 2020). Informed written consent from each participant was not required for this study according to the French Data Protection Act of 6 January 1978, as this study only included information from existing medical records and did not involve interaction with patients or collection of identifiable private information. Each entry of sample data was deidentified to ensure confidentiality.

Software. Data extractions, manipulations, statistical analyses, and model buildings were performed with "R-studio", version 1.3.1093, RStudio PBC, 2009–2020. Specialized packages and functions were used for specific analysis: "Dplyr", version 1.0.0 was used for data manipulation, "Purrr", version 0.3.4 for data simplification, and "missForest", version 1.4 for missing data imputation. Variable importance calculations and K-fold cross-validation were performed with the "Caret" package, version 6.0-86. Correlations matrix were calculated with the "corrplot" package, version 0.84. Random forests were built with "randomForest" version 4.6-14 and artificial neural networks with "neuralnet" version 1.44.2. "pROC" version 1.16.2 was used to generate the receiver operating characteristic (ROC) curves and calculate the area under the curve (AUC) for each model.

Setting. Data were collected retrospectively from patients admitted to the adult E.D. of Rennes Academic Hospital, France.

Patient selection. All post-emergency hospitalized patients ≥ 18 years old admitted between March 20, 2020 and May 5, 2020 and suspected to have COVID-19 were included in the study.

Data collection. Data were automatically collected from "eHOP", a local clinical data warehouse in which health data are integrated and de-identified in real time[18]. Structured data, such as laboratory results, were directly collected from the data warehouse. Text fields were structured by using regular expressions[19].

Data pre-processing. In the raw data-frame, all values were associated with a unique identifier (ID) corresponding to each patient's admission. This data-frame contained multiple lines per ID (Fig. 1, step 1). Variables collected more than once during the patient journey appeared as lists (Fig. 1, step 2). Lists were simplified according to the type of variable (Fig. 1, step 3).

Predicted variable. The predicted variable for each patient was the presence of COVID-19. COVID-19 was diagnosed as follows. Patients were triaged at ED admission and considered "suspected COVID-19" when they had at least one symptom compatible with COVID-19. Symptoms considered as compatible with COVID-19 were the followings: cough, dyspnea, hyperthermia, myalgias, asthenia, diarrhea, confusion or anosmia. After triage, patients were examined by an ED physician who estimated the clinical probability of COVID-19 (low, intermediate or high) and the need for hospitalization. RT-PCR and chest-CT were performed in all hospitalized patient suspected to have COVID-19. When suspected COVID-19 patients had positive RT-PCR, they were considered "COVID-19 positive", regardless of the level of clinical probability. When clinical probability was "high" and chest-CT showed typical COVID-19 images with no sign of alternate diagnosis, the patient was considered "COVID-19 positive", even if RT-PCR was negative. In this case, the RT-PCR result was considered a false negative[20,21]. Patients were allocated to "COVID" and "NOT-COVID" groups accordingly.

Predicting variables selection. All clinical and laboratory variables present in the database were collected. The Student's *t*- and chi-square tests were used to compare means between groups for numerical and binary variables, respectively. A p value < 0.05 was considered statistically significant. Variables with a p value < 0.2 were considered variables of interest. To avoid multicollinearity, correlation coefficients were calculated for each pair of variables of interest. When correlation coefficient was higher than 0.8, one of the two variables was excluded.

Data split. Data were randomly divided in two parts: the train data-frame, and the test data- frame. The train data-frame corresponded to 80% of the whole data-frame and was used to build the models. Models performances were evaluated using the test data-frame that corresponded to the remaining 20%.

Missing data imputation. Before the training process, missing values were imputed independently for each data-frame with a non-parametric procedure developed by Stekhoven and Buhlmann. This method called "missForest" is well-suited for mixed datasets requiring categorical and continuous variables imputations and is based on a random forest model trained iteratively. In this method, an evaluation is made after each iteration by calculation of the normalized root mean squared error and implementation is stopped when the evaluation indicates a decrease in performance. Three iteration were performed with 100 tree per random forest in this study.

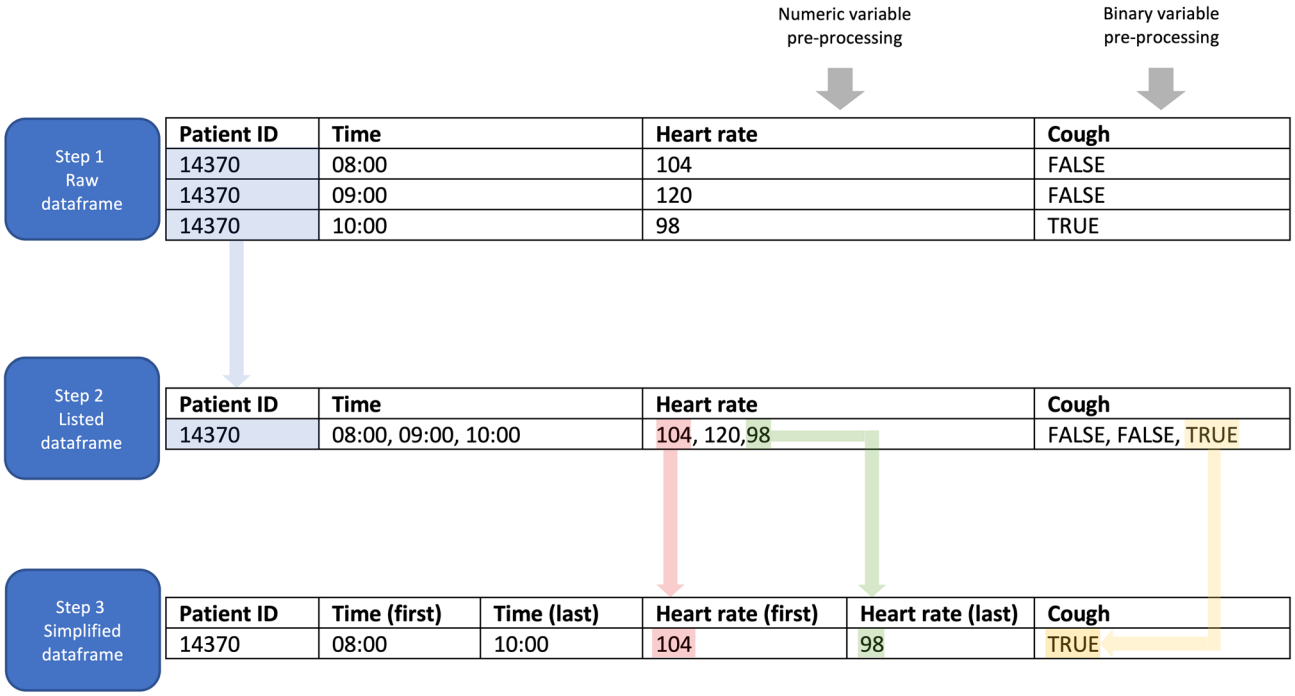

Figure 1. Data pre-processing. *The first step* corresponded to raw data, as they were initially stored in the database. Each ID was characterized by multiple rows. *On the second step*, data were listed in chronological order, with a single row per ID (blue arrow). *In the third step*, data were simplified. For numeric variables, only the first value was selected (red arrow). For binary variables, the value "true" was retained when it was present at least once in the list (yellow arrow).

Model training. Three model types were constructed: binary logistic regressions, random forests and artificial neural networks. Random forest models were trained with 500 trees; neural networks were composed of three layers. Each model type was trained with three sets of variables: clinico-biological variables, clinico-biological variables with chest-CT, and clinico-biological variables with RT-PCR. A k-fold cross validation was performed in order to prevent over-fitting. Overfitting occurs when à machine learning algorithm captures the noise of the data. In this case, high performances are observed on the training data, but poor results are observed on new data. In other words, overfitted models cannot give suitable predictions on new patients. K-folds cross validation is a high-performance method to prevent overfitting. In this approach, the data-frame is divided into k parts called "folds". A model is trained by using k − 1 folds, and the remaining fold is used to validate the model. The same procedure is applied k times (once per fold). This approach is well-suited for small datasets, but requires more calculations. In this study, k = 10 folds were used to build the models.

Variable importance. To compare the importance of the different variables, the value of the most important variable in each model was arbitrarily set at 100 and the relative importance of each variable was determined with an adequate method depending on the model. In binary logistic regressions, the absolute value of the t-statistic for each parameter was used to calculate the importance of each variable. In random forests, the prediction error on the out-of-bag portion of the data was recorded for each tree and he same was done after permuting each predictor variable. The difference between the two were averaged over all trees and normalized by the standard deviation of the differences to determine each variables importance. In the neural network models, the method was based on combinations of the absolute values of the weights[22].

Performance measurement. Models were built with the train data-frame and their performances were assessed on the test data-frame, whose data were not used for model-building. This procedure guarantees non-biased performances measurements by confronting the models to unseen data, as if they were challenged to predict the presence of COVID-19 among new patients. The area under ROC curves is commonly used to evaluate and compare classifiers in machine learning, biomedical and bioinformatics applications[23]. In this study, models' predictions were compared to the "COVID" variable in the test data-frame and ROC curves were constructed accordingly. The AUC was the primary outcome used to evaluate each model performance.

Ethics approval. This study was approved by the ethic committee of Rennes academic hospital (number of approval: 20.93).

Consent for publication. All methods were performed in accordance with the relevant guidelines and regulations. Authorization to conduct research from the Clinical Data Warehouse of Hospital of Rennes was

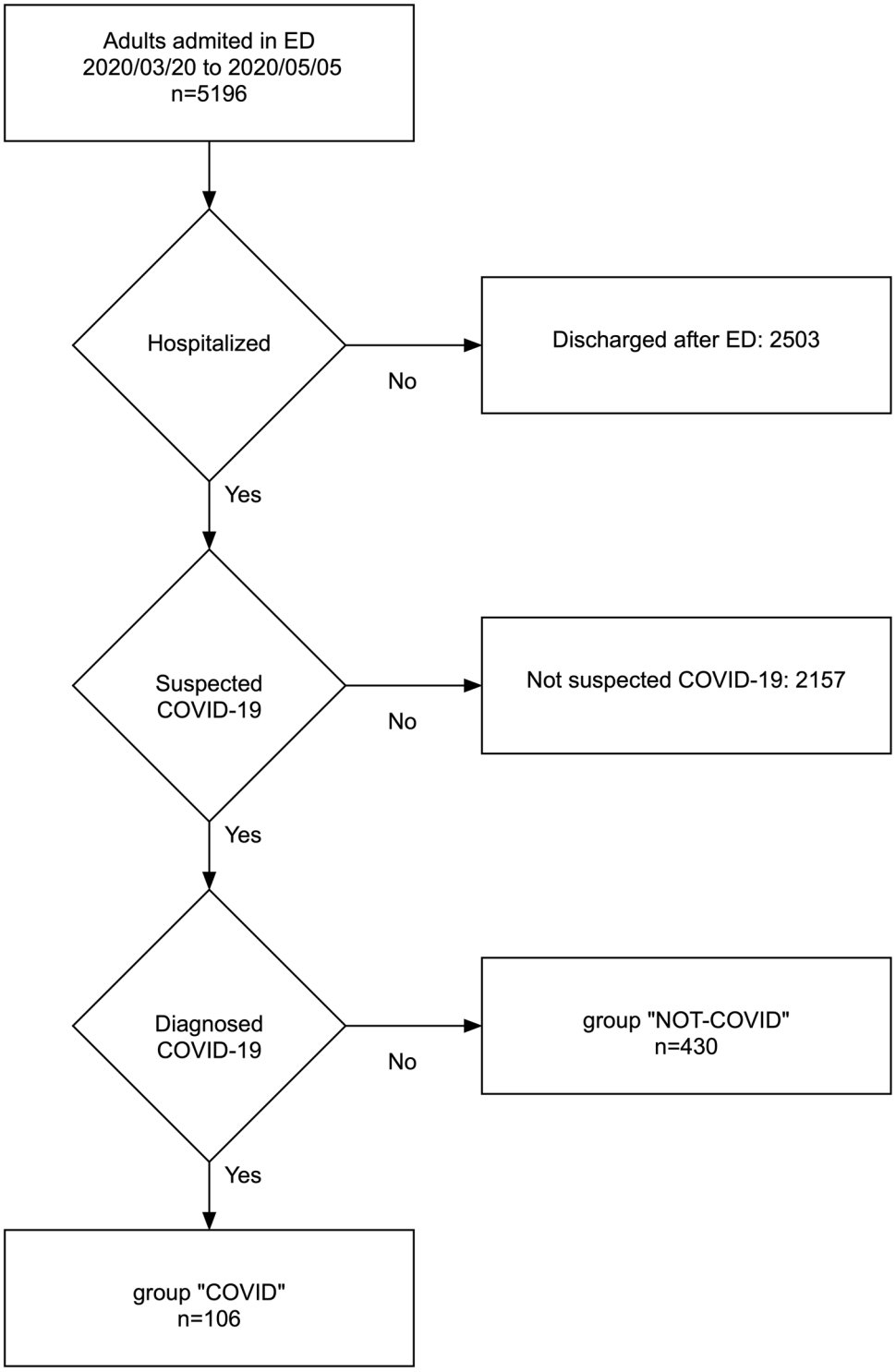

Figure 2. Flow chart of patient selection. Patients suspected to have COVID-19 had at least one of the following symptoms: cough, dyspnea, hyperthermia, myalgias, asthenia, diarrhea, confusion or anosmia. Both chest-CT and RT-PCR were performed in all patients with suspected COVID-19 who were hospitalized.

given by CNIL—Commission Nationale Informatique et Liberté (Authorization number 2020-028 issued on February 27, 2020). Informed written consent from each participant was not required for this study according to the French Data Protection Act of 6 January 1978, as this study only included information from existing medical records and did not involve interaction with patients or collection of identifiable private information. Each entry of sample data was deidentified to ensure confidentiality.

Diagnostic	n (%)
COVID-19	106 (19.8)
Cardiac insufficiency	98 (18.3)
Pneumonia	74 (13.8)
Chronic obstructive pulmonary disease (COPD)	52 (9.7)
Influenza-like illness	38 (7.1)
Intra-abdominal infection	34 (6.3)
Asthma	20 (3.7)
Non organic dyspnea	19 (3.5)
Urinary tract infection	19 (3.5)
Confusion in the elderly (delirium)	14 (2.6)
Transcient fever	14 (2.6)
Cancer	12 (2.2)
Pulmonary embolism	12 (2.2)
Skin infection	6 (1.1)
Others	5 (0.9)
Central nervous system infection	4 (0.7)
Heart infection (pericarditis, myocarditis, endocarditis)	4 (0.7)
Prothesis-related infection	3 (0.6)
Traumatic dyspnea	2 (0.4)
Total	536 (100)

Table 1. Diagnostic categories for the 536 suspected COVID-19 hospitalized patients. All patients presented at last one clinical sign compatible with COVID-19 and underwent chest-CT and RT-PCR. 106 were classified in the COVID group, 430 in the NOT-COVID group.

Results

Patient selection. The patient selection flow chart is presented Fig. 2.

Diagnostics. Diagnostics for the 536 patients selected in this study are represented in Table 1.

Selected variables. Twenty-three clinico-biological variables were considered as variables of interest (Table 2). Variables not selected as variables of interests are presented in supplementary Table 1.

Variables correlations. Calculation of the correlation coefficients for each pair of the 23 variables of interest (Fig. 3) showed that two variables were highly correlated with a correlation coefficient > 0.8: neutrophil count and leukocyte count. Leukocyte count was removed from model building. Therefore, the final set of clinical and laboratory variables selected for model building included 22 variables.

Chest-CT and RT-PCR performances. AUCs of chest-CT and RT-PCR used alone for COVID-19 diagnosis were 0.778 (CI 95% 0.682–0.873) and 0.852 (CI 95% 0.764–0.940), respectively.

Models performance. The AUC values for the three model types trained with each set of variables are presented in Table 3.

The ROC curves for the binary logistic regression models are presented Fig. 4.

Importance of clinico-biological variables. The importance of the different variables in each model type is presented Table 4.

Discussion

Models presented in this study were trained on typical suspected COVID-19 patients. All models were trained and evaluated using data from patients with diseases (e.g. heart failure, pneumonia, asthma, COPD; Table 1) that are frequently observed in ED and that share clinical symptoms with COVID-19. The finding that our machine learning models could differentiate between these diseases and COVID-19 suggests that they could be implemented in other EDs with similar patient populations.

The variables selected for model-building were consistent with the clinico-biological signs of COVID-19. These variables belong to five categories: clinical signs, arterial blood gas, blood cell count, ionogram, hemostasis and liver enzymes. Clinical signs: the proportion of cough, hyperthermia, myalgia, asthenia, diarrhea, and confusion was significantly higher in the COVID-19 than in the NO-COVID-19 group. Such

	NOT-COVID (n = 430)	COVID (n = 106)	P value
Clinicals and treatments			
Cough, %	83.0 (79.1–87.0	92.4 (86.5–98.2)	0.0563
Hyperthermia, %	66.7 (61.8–71.7)	77.2 (67.9–86.4)	0.0940
Myalgias, %	17.1 (13.2–21.1)	34.1 (23.7–44.6)	0.0012*
Asthenia, %	30.9 (26.1–35.8)	45.5 (34.5–56.5)	0.0187*
Diarrhea, %	22.9 (18.5–27.3)	32.9 (22.5–43.2)	0.0867
Confusion, %	21.7 (17.4–26.1)	7.5 (1.7–13.4)	0.0063*
Furosemid (usual treatment), %	16.0 (12.2–19.9)	6.3 (0.9–11.7)	0.0401*
Arterial blood gas			
Base excess, mmol/L	3.0 (2.6–3.4)	2.7 (1.8–3.6)	0.0151*
Lactates, mmol/L	1.7 (1.5–1.9)	1.3 (1.1–1.5)	<0.001*
Complete blood count			
Red blood cell count, Tera/L	4.2 (4.1–4.3)	4.5 (4.3–4.7)	<0.001*
Mean platelet volume, fL	8.6 (8.4–8.8)	8.8 (8.5–9.1)	0.0269*
Leukocytes, G/L	10.2 (9.6–10.8)	7.7 (6.7–8.7)	0.0568
Neutrophils, G/L	7.9 (7.4–8.4)	6 (5.1–6.9)	0.1488
Platelet count	236.1 (225.8–246.4)	198.9 (182.1–215.7)	0.0482*
Eosinophils percentage	1.4 (1.1–1.7)	0.8 (0.4–1.2)	0.0873
Basophils percentage	0.6 (0.5–0.7)	0.4 (0.3–0.5)	<0.001*
Lymphocytes, G/L	1.3 (1.2–1.4)	1 (0.8–1.2)	<0.001*
Monocytes, G/L	0.8 (0.7–0.9)	0.6 (0.5–0.7)	<0.001*
Ionogram			
Potassium, mmol/L	4.1 (4–4.2)	4 (3.8–4.2)	0.0039*
Phosphor, mmol/L	1 (0.9–1.1)	1.1 (0.9–1.3)	<0.001*
Hemostasis and liver enzymes			
Alanine aminotransferase, mmol/L	64.5 (47.1–81.9)	46.2 (33.9–58.5)	0.1845
International normalized ratio	1.3 (1.2–1.4)	1.2 (1.1–1.3)	<0.001*
D-Dimer, ng/ml	2200 (1600–2800)	2800 (1400–4200)	<0.001*

Table 2. Variables of interest. Means and percentage between groups were compared with Student's t- and chi-square tests, respectively. Only variables with $p < 0.02$ were considered as variables of interest and were listed in this table. Values in brackets represent the 95% confidence intervals. *$p < 0.005$.

symptoms have previously been reported in numerous studies[24–28]. Interestingly, anosmia was not selected as a variable of interest, suggesting a lack of relevance of this symptom in our setting[29,30]. Arterial blood gas: in the NOT-COVID group, serum lactate concentration was higher, and base-excess was lower than in the COVID group, revealing the presence of patients with circulatory failure, a frequently reported complication of bacteremia[31]. Therefore, serum lactate concentration and base-excess are relevant for differentiating between patients with COVID-19 and with bacterial infections. Blood cell count: the mean leukocyte, lymphocyte, and platelet counts were lower in the COVID than in the NOT-COVID group. Previous authors have reported similar results. Indeed, a meta-analysis from Zhu et al. showed that patients with COVID-19 do not have hyperleukocytosis, except when associated with bacteremia[32]. COVID-19-associated lymphopenia correlates with the disease severity and is related to an immune response deficiency[33]. Similarly, thrombocytopenia was previously identified as a poor prognosis factor in this context[34]. Indeed, a meta- analysis by Lippi et al. revealed that platelet count was significantly lower in patients with severe COVID-19[35], suggesting an inappropriate activation of the coagulation process. Ionogram: the mean potassium concentration was lower in the COVID group. This could be due to hyperventilation, but further investigation must be conducted to confirm this hypothesis. Hemostasis and liver enzymes: the mean D-dimer concentration was higher in the COVID group than in the NOT-COVID group. Elevated D-dimers are associated with higher rates of thromboembolic events[36]. These results are in line with the theory of an increased thromboembolic risk in patients with COVID-19[37–39]. This finding could be associated with the presence of antiphospholipid antibodies, but the pathophysiology of this phenomenon is still debated[40]. Variables selected for model building were therefore consistent with previous studies that have reported clinico-biological signs of COVID-19.

Machine learning models will help to triage COVID-19 patients. RT-PCR and chest-CT are expensive, require qualified professionals to perform them and it is a real challenge to be able to get efficiently these two examinations in the context of a pandemic. An increasing number of patients awaiting results of these tests can lead to ED overcrowding and increased mortality rates in an epidemic context[41,42]. The logistic regression model presented in this study and trained only with clinico-biological variables had an AUC value of 0.754. This model only requires clinical examination and routine biology assays: complete blood cell count, ionogram,

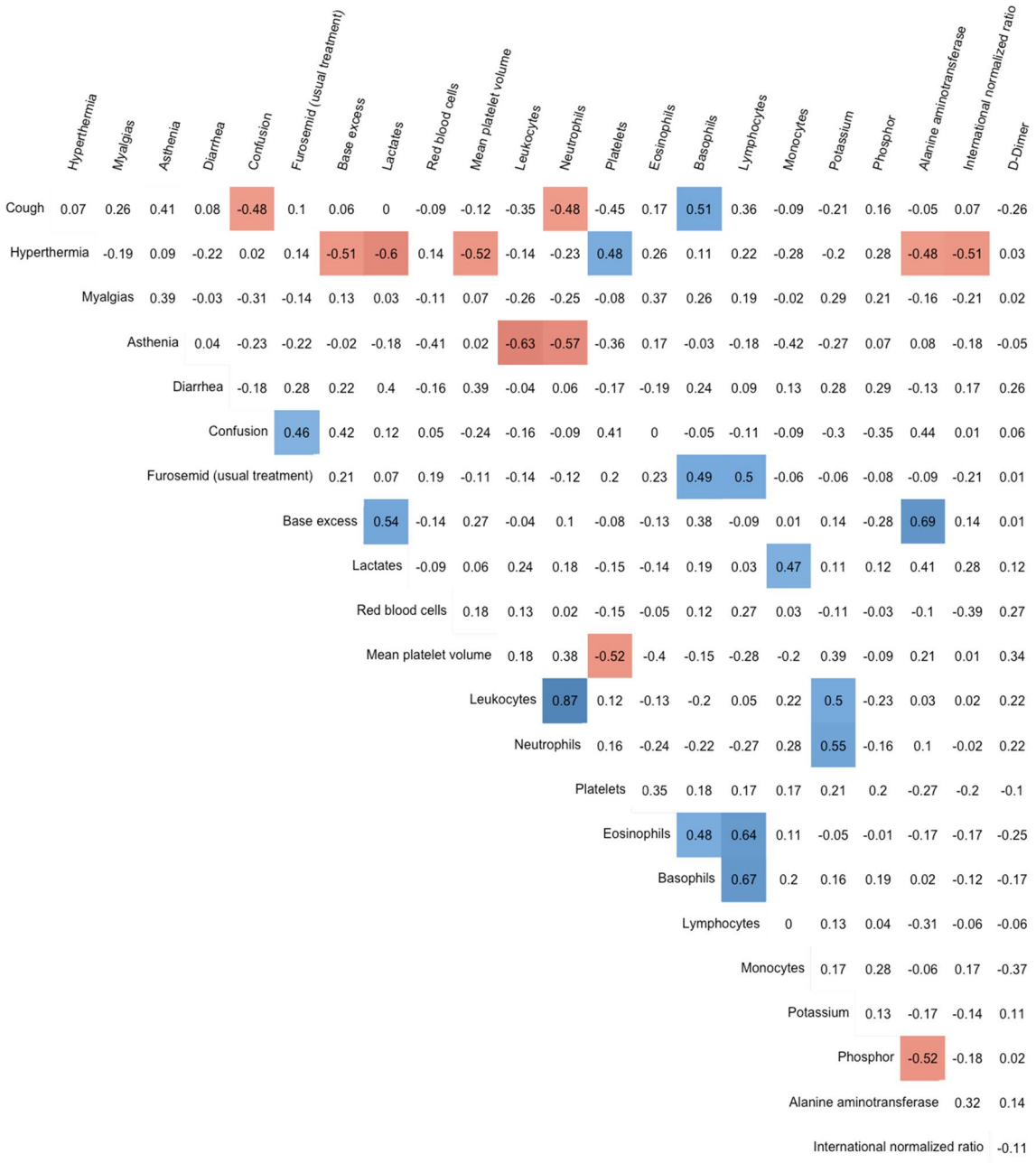

Figure 3. Correlation coefficients for all pairs of variables of interest. Pearson and Spearman coefficients were calculated for continuous and binary variables, respectively. Correlations not significantly different from 0 are in white cells. Positive correlations are in blue cells, and negative correlations in red cells. Leukocyte count and neutrophil count were identified as highly correlated, and leukocyte count was removed from model building.

	Clinico-biological	Clinico-biological + chest-CT	Clinico-biological + RT-PCR
Binary logistic regression	0.772 (0.668–0.875)	0.886 (0.804–0.968)	0.930 (0.867–0.992)
Random forest	0.754 (0.638–0.871)	0.829 (0.724–0.935)	0.903 (0.816–0.989)
Artificial neural network	0.728 (0.617–0.840)	0.892 (0.811–0.973)	0.844 (0.731–0.957)

Table 3. AUC for each machine learning model. Three model types were constructed: binary logistic regression, random forest, and artificial neural network. Each model was trained with three sets of variables: clinico-biological, clinico-biological with chest-CT, and clinico-biological with RT-PCR. Models were built and assessed on two separate data-frames. Values in brackets represent the 95% confidence intervals.

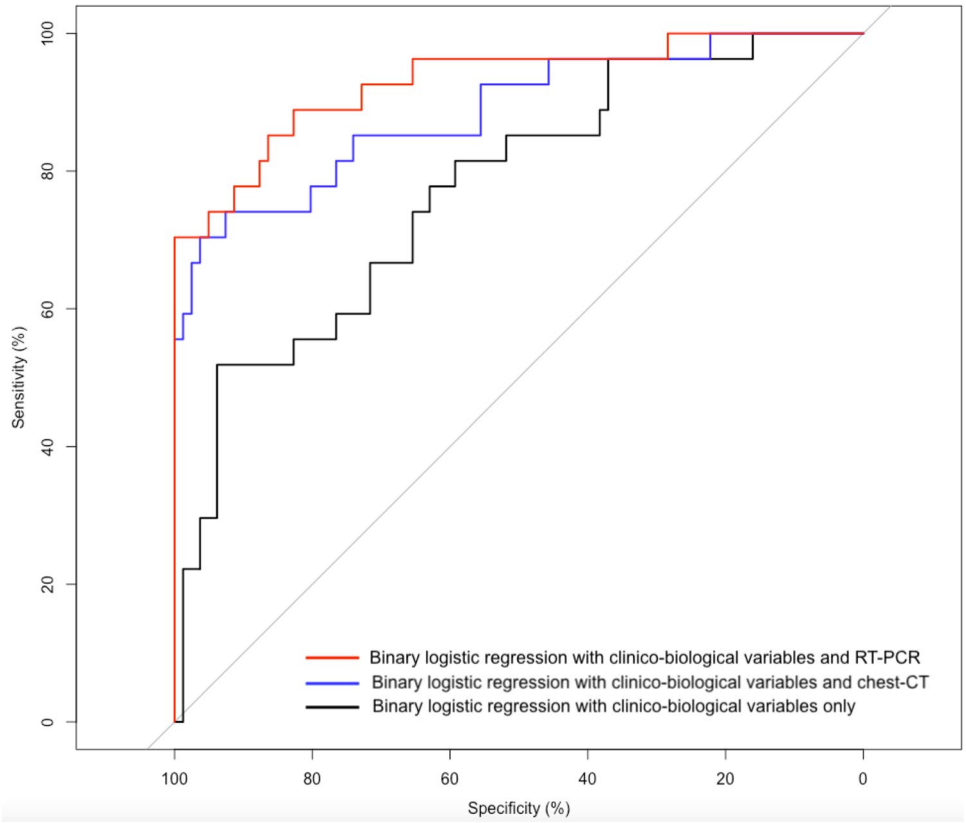

Figure 4. ROC curves for the 3 logistic regression models based on common clinico-biological variables alone, clinico-biological variables with chest-CT and common clinico-biological variables with RT-PCR. The "Binary logistic regression with clinico-biological variables and RT-PCR" was the best performing model in this study.

standard hemostasis tests, liver enzymes, and arterial blood gas. Such tests are low- cost and can be realized worldwide using automated devices. Therefore, in ED, a first triage identifying patients requiring isolation might be done by using machine learning while waiting for the RT-PCR result.

Machine learning will improve RT-PCR and chest-CT performance for COVID-19 diagnosis. Several studies found sub-optimal performances of these tests when only one is used for COVID-19 diagnosis[11,13,43,44]. Indeed, the sensitivity of the RT-PCR test depends on the number of cycles used to determine the cut-off value for positivity[45] and one of the issues by using chest-CT alone for COVID-19 diagnosis is the risk of false positive[20,46]. Artificial intelligence methods could be used to overcome these drawbacks. Some studies have already investigated the use of artificial intelligence for COVID-19 diagnosis and their number is progressively increasing with the pandemic duration[47,48]. Many of these studies are based on deep neural networks to improve COVID-19 diagnosis by chest-CT or X-ray imaging, particularly to help to differentiate between COVID-19 lesions and bacterial lung diseases[49–52]. For examples, the COVID-net tool based on 16,756 chest radiography images across 13,645 patients has an accuracy of 92.4%, the COVID-19 detection neural network (COVNet) based on 4356 chest-CT from 3322 patients has an accuracy of 95%[53,54]. However, few studies used laboratory, clinical, and imaging data together for COVID-19 diagnosis. Our results are in line with studies that used machine learning models based on clinico-biological variables for COVID-19 diagnosis[55–57]. The performances of these models were low, excepted for the model described by Plante and al. that used data from a large sample, but did not include imaging data[58]. Another study integrated RT-PCR, chest-CT and clinico-biological data, like in the present work, but the study population was smaller, and the performance was slightly lower[59]. In our study, the AUC values for chest-CT and RT-PCR increased from 0.778 to 0.892 and from 0.852 to 0.930 with the contribution of machine learning. The generalization of such models will allow increasing the diagnostic performances of both chest-CT and RT-PCR for COVID-19 diagnosis.

Limitations. Our study has some limitations. First, the machine learning models developed in this experimentation are not directly transferrable to other hospitals due to it's monocentric design. Such models must be further developed and tested on a larger scale to be generalized. However, the predictive variables selected and identified as highly important in this study are similar to the clinical and biological signs reported by previous authors, suggesting the absence of major obstacles for model generalization. Second, the study population included only hospitalized patients suspected to have COVID-19. It would be interesting to perform a similar study in non-hospitalized patient to test the model performances for COVID-19 diagnosis in paucisympto-

	Binary logistic regression	Random forest	Artificial neural network
Clinicals and treatments			
Cough	62	1.5	21.3
Hyperthermia	24.7	3.1	24.1
Myalgias	40.5	9.36	39.7
Asthenia	24.1	6.7	34.1
Diarrhea	57.1	6	22.9
Confusion	91.8	5.8	33
Furosemid (usual treatment)	72.8	0	22.3
Arterial blood gas			
Base excess	31.6	37.3	16.2
Lactates	100	100	77.9
Complete blood count			
Red blood cell count	85.1	86.4	41
Mean platelet volume	0	34.5	0
Neutrophils	64.6	50.6	55.1
Platelet count	36.8	46.8	47.3
Eosinophils	35.9	36.5	57.1
Basophils	74.8	79.9	100
Lymphocytes	67.4	51.8	46
Monocytes	27.6	57.3	52.6
Ionogram			
Potassium	17.5	39.4	19.4
Phosphor	6.2	45.5	4.2
Hemostasis and liver enzymes			
Alanine aminotransferase	64.6	39.8	3.1
International normalized ratio	9	39.2	0.2
D-Dimer	57.6	58.4	10

Table 4. Importance of clinico-biological variables by decreasing order in each model type. The relative importance of each variable was calculated in comparison with the most important variable in the model, whose importance was arbitrarily set at 100.

matic patients. Finally, the classification of chest-CT as negative on the basis of the absence of typical images of COVID-19 might need to be reviewed in line with recent publications on COVID-19 diagnosis using deep learning methods.

Conclusion. Our study demonstrates that machine learning models can be developed for improving COVID-19 diagnosis in patients hospitalized through the ED. Models based on chest-CT or RT-PCR will increase the performance of these tests by using clinico-biological variables. After generalization, machine learning should play a key role in the management of the outbreak by improving the performances of chest-CT and RT-PCR for COVID-19 diagnosis.

Data availability

After publication, the data will be made available to others on reasonable requests to the corresponding author.

Received: 30 October 2020; Accepted: 16 March 2021
Published online: 30 March 2021

References

1. Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J. & Prescott, H. C. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): A review. *JAMA* **324**, 782 (2020).
2. Li, Q. *et al.* Early transmission dynamics in Wuhan, China, of Novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).
3. Korean Society of Infectious Diseases and Korea Centers for Disease Control and Prevention. Analysis on 54 mortality cases of coronavirus disease 2019 in the Republic of Korea from January 19 to March 10, 2020. *J. Korean Med. Sci.* **35**, e132 (2020).
4. Peng, L. *et al.* Improved early recognition of coronavirus disease-2019 (COVID-19): Single-center data from a Shanghai Screening Hospital. *Arch. Iran. Med.* **23**, 272–276 (2020).
5. Wong, S. C. Y. *et al.* Risk of nosocomial transmission of coronavirus disease 2019: An experience in a general ward setting in Hong Kong. *J. Hosp. Infect.* **105**, 119–127 (2020).
6. For the Singapore 2019 Novel Coronavirus Outbreak Research Team *et al.* Detection of air and surface contamination by SARS-CoV-2 in hospital rooms of infected patients. *Nat. Commun.* **11**, 2800 (2020).

7. Xiao, J., Fang, M., Chen, Q. & He, B. SARS, MERS and COVID-19 among healthcare workers: A narrative review. *J. Infect. Public Health* **13**, 843–848 (2020).
8. Coccolini, F. *et al.* COVID-19 the showdown for mass casualty preparedness and management: The Cassandra Syndrome. *World J. Emerg. Surg.* **15**, 26 (2020).
9. Maves, R. C. *et al.* Triage of scarce critical care resources in COVID-19 an implementation guide for regional allocation. *Chest* **158**, 212–225 (2020).
10. Hanson, K. E. *et al.* Infectious diseases society of America Guidelines on the Diagnosis of COVID-19. *Clin. Infect. Dis.* https://doi.org/10.1093/cid/ciaa760 (2020).
11. Xiao, A. T. False negative of RT-PCR and prolonged nucleic acid conversion in COVID-19: Rather than recurrence. 2.
12. Li, Y. *et al.* Stability issues of RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed with COVID-19. *J. Med. Virol.* **92**, 903–908 (2020).
13. Ai, T. *et al.* Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases. 23.
14. Rubin, G. D. *et al.* The role of chest imaging in patient management during the COVID-19 pandemic: A multinational consensus statement from the Fleischner Society. *Radiology* **296**, 172–180 (2020).
15. Beetz, C. *et al.* Rapid large-scale COVID-19 testing during shortages. *Diagnostics* **10**, 464 (2020).
16. Lone, S. A. & Ahmad, A. COVID-19 pandemic—an African perspective. *Emerg. Microbes Infect.* **9**, 1300–1308 (2020).
17. Furlow, B. Deep learning poised to revolutionise diagnostic imaging. *Lancet Respir. Med.* **5**, 779 (2017).
18. Delamarre, D., Bouzille, G., Dalleau, K., Courtel, D. & Cuggia, M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. 5.
19. Tang, X., Zeng, Q., Cui, T. & Wu, Z. Regular expression-based reference metadata extraction from the web. in *2010 IEEE 2nd Symposium on Web Society* 5607427 (IEEE, 2010). https://doi.org/10.1109/SWS.2010.5607427.
20. Caruso, D. *et al.* Chest CT features of COVID-19 in Rome, Italy. *Radiology* **296**, E79–E85 (2020).
21. Chung, M. *et al.* CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology* **295**, 202–207 (2020).
22. Gevrey, M., Dimopoulos, I. & Lek, S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol. Model.* **160**, 249–264 (2003).
23. Robin, X. *et al.* pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **12**, 77 (2011).
24. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
25. Chen, N. *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *Lancet* **395**, 507–513 (2020).
26. Jiang, F. *et al.* Review of the clinical characteristics of coronavirus disease 2019 (COVID-19). *J. Gen. Intern. Med.* **35**, 1545–1549 (2020).
27. Goyal, P. *et al.* Clinical characteristics of covid-19 in New York City. *N. Engl. J. Med.* **382**, 2372–2374 (2020).
28. Wang, D. *et al.* Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* **323**, 1061 (2020).
29. Lechien, J. R. *et al.* Olfactory and gustatory dysfunctions as a clinical presentation of mild-to-moderate forms of the coronavirus disease (COVID-19): A multicenter European study. *Eur. Arch. Otorhinolaryngol.* **277**, 2251–2261 (2020).
30. Beltrán-Corbellini, Á. *et al.* Acute-onset smell and taste disorders in the context of COVID-19: A pilot multicentre polymerase chain reaction based case–control study. *Eur. J. Neurol.* **27**, 1738–1741 (2020).
31. Shankar-Hari, M. *et al.* Developing a new definition and assessing new clinical criteria for septic shock: For the third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* **315**, 775 (2016).
32. Zhu, J. *et al.* Clinicopathological characteristics of 8697 patients with COVID-19 in China: A meta-analysis. *Fam. Med. Community Health* **8**, e000406 (2020).
33. Azkur, A. K. *et al.* Immune response to SARS-CoV-2 and mechanisms of immunopathological changes in COVID-19. *Allergy* **75**, 1564–1581 (2020).
34. Hu, L. *et al.* Risk factors associated with clinical outcomes in 323 COVID-19 hospitalized patients in Wuhan, China. 33.
35. Lippi, G., Plebani, M. & Henry, B. M. Thrombocytopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: A meta-analysis. *Clin. Chim. Acta* **506**, 145–148 (2020).
36. Crawford, F. *et al.* D-dimer test for excluding the diagnosis of pulmonary embolism. *Cochrane Database Syst. Rev.* https://doi.org/10.1002/14651858.CD010864.pub2 (2016).
37. Spiezia, L. *et al.* COVID-19-related severe hypercoagulability in patients admitted to intensive care unit for acute respiratory failure. *Thromb. Haemost.* **120**, 998–1000 (2020).
38. Oxley, T. J. *et al.* Large-vessel stroke as a presenting feature of Covid-19 in the young. *N. Engl. J. Med.* **382**, e60 (2020).
39. Zhang, L. *et al.* D-dimer levels on admission to predict in-hospital mortality in patients with Covid-19. *J. Thromb. Haemost.* **18**, 1324–1329 (2020).
40. Zhang, Y. *et al.* Coagulopathy and antiphospholipid antibodies in patients with Covid-19. *N. Engl. J. Med.* **382**, e38 (2020).
41. Geelhoed, G. C. & Klerk, N. H. Emergency department overcrowding, mortality and the 4-hour rule in Western Australia. *Med. J. Aust.* **196**, 122–126 (2012).
42. Kim, J. *et al.* Maximum emergency department overcrowding is correlated with occurrence of unexpected cardiac arrest. *Crit. Care* **24**, 305 (2020).
43. Long, C. *et al.* Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT?. *Eur. J. Radiol.* **126**, 108961 (2020).
44. Liu, R. *et al.* Positive rate of RT-PCR detection of SARS-CoV-2 infection in 4880 cases from one hospital in Wuhan, China, from Jan to Feb 2020. *Clin. Chim. Acta* **505**, 172–175 (2020).
45. Liu, Z. High sensitivity detection of SARS-CoV-2 using multiplex PCR and a multiplex-PCR-based metagenomic method. 24.
46. Yang, W. *et al.* The role of imaging in 2019 novel coronavirus pneumonia (COVID-19). *Eur. Radiol.* **30**, 4874–4882 (2020).
47. Albahri, A. S. *et al.* Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): A systematic review. *J. Med. Syst.* **44**, 122 (2020).
48. Dananjayan, S. & Raj, G. M. Artificial Intelligence during a pandemic: The COVID-19 example. *Int. J. Health Plann. Manag.* **35**, 1260–1262 (2020).
49. Wu, X. *et al.* Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: A multicentre study. *Eur. J. Radiol.* **128**, 109041 (2020).
50. Kang, H. *et al.* Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning. *IEEE Trans. Med. Imaging* **39**, 2606–2614 (2020).
51. Fan, Z., Jamil, M., Sadiq, M. T., Huang, X. & Yu, X. Exploiting multiple optimizers with transfer learning techniques for the identification of COVID-19 patients. *J. Healthc. Eng.* **2020**, 1–13 (2020).
52. Jang, S. B. *et al.* Deep-learning algorithms for the interpretation of chest radiographs to aid in the triage of COVID-19 patients: A multicenter retrospective study. *PLoS One* **15**, e0242759 (2020).
53. 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).
54. Kumar, A., Gupta, P. K. & Srivastava, A. A review of modern technologies for tackling COVID-19 pandemic. *Diabetes Metab. Syndr. Clin. Res. Rev.* **14**, 569–573 (2020).

55. Goodman-Meza, D. *et al.* A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity. *PLoS One* **15**, e0239474 (2020).
56. D'Ambrosia, C., Christensen, H. & Aronoff-Spencer, E. Computing SARS-CoV-2 infection risk from symptoms, imaging, and test data: Diagnostic model development. *J. Med. Internet Res.* **22**, e24478 (2020).
57. Langer, T. *et al.* Development of machine learning models to predict RT-PCR results for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in patients with influenza-like symptoms using only basic clinical data. *Scand. J. Trauma Resusc. Emerg. Med.* **28**, 113 (2020).
58. Plante, T. B. *et al.* Development and external validation of a machine learning tool to rule out COVID-19 among adults in the emergency department using routine blood tests: A large, multicentre, real-world study. *J. Med. Internet Res.* **22**, e24048 (2020).
59. Hermans, J. J. R. *et al.* Chest CT for triage during COVID-19 on the emergency department: Myth or truth?. *Emerg. Radiol.* **27**, 641–651 (2020).

Author contributions

C.G. designed the experiment, collected data, performed statistical analysis and build machine-learning models. C.G. and S.R. wrote the manuscript. G.B., L.S. and M.C. read and approved the final manuscript.

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/s41598-021-86735-9.

Correspondence and requests for materials should be addressed to C.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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021