Journal of Medical Systems (2020) 44:135
https://doi.org/10.1007/s10916-020-01597-4

IMAGE & SIGNAL PROCESSING

Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study

Davide Brinati[1] · Andrea Campagner[1] · Davide Ferrari[2] · Massimo Locatelli[3] · Giuseppe Banfi[4] · Federico Cabitza[1]

Received: 23 April 2020 / Accepted: 2 June 2020

Abstract

The COVID-19 pandemia due to the SARS-CoV-2 coronavirus, in its first 4 months since its outbreak, has to date reached more than 200 countries worldwide with more than 2 million confirmed cases (probably a much higher number of infected), and almost 200,000 deaths. Amplification of viral RNA by (real time) reverse transcription polymerase chain reaction (rRT-PCR) is the current gold standard test for confirmation of infection, although it presents known shortcomings: long turnaround times (3-4 hours to generate results), potential shortage of reagents, false-negative rates as large as 15-20%, the need for certified laboratories, expensive equipment and trained personnel. Thus there is a need for alternative, faster, less expensive and more accessible tests. We developed two machine learning classification models using hematochemical values from routine blood exams (namely: white blood cells counts, and the platelets, CRP, AST, ALT, GGT, ALP, LDH plasma levels) drawn from 279 patients who, after being admitted to the San Raffaele Hospital (Milan, Italy) emergency-room with COVID-19 symptoms, were screened with the rRT-PCR test performed on respiratory tract specimens. Of these patients, 177 resulted positive, whereas 102 received a negative response. We have developed two machine learning models, to discriminate between patients who are either positive or negative to the SARS-CoV-2: their accuracy ranges between 82% and 86%, and sensitivity between 92% e 95%, so comparably well with respect to the gold standard. We also developed an interpretable Decision Tree model as a simple decision aid for clinician interpreting blood tests (even off-line) for COVID-19 suspect cases. This study demonstrated the feasibility and clinical soundness of using blood tests analysis and machine learning as an alternative to rRT-PCR for identifying COVID-19 positive patients. This is especially useful in those countries, like developing ones, suffering from shortages of rRT-PCR reagents and specialized laboratories. We made available a Web-based tool for clinical reference and evaluation (This tool is available at https://covid19-blood-ml.herokuapp.com/).

Keywords COVID-19 · RT-PCR test · Blood tests · Machine learning · Random forest · Three-way

Introduction

The pandemic disease caused by the SARS-CoV-2 virus named COVID-19 is requiring unprecedented responses of exceptional intensity and scope to more than 200 states around the world, after having infected, in the first 4 months since its outbreak, a number of people between 2 and 20 million with at least 200,000 deaths. To cope with the spread of the COVID-19 infection, governments all over the world has taken drastic measures like the quarantine of hundreds of millions of residents worldwide.

However, because of the COVID-19 symptomatology, which showed a large number of asymptomatics [12], these efforts are limited by the problem of differentiating between

This article is part of the Topical Collection on *Image & Signal Processing*

✉ Federico Cabitza
federico.cabitza@unimib.it

1 DISCo, Università degli Studi di Milano-Bicocca, Viale Sarca 336, Milano, 20126, Italy

2 SCVSA Department, University of Parma, Parco Area delle Science 11/a, 43124 Parman, Italy

3 Laboratory Medicine Service, San Raffaele Hospital, Via Olgettina, 60, 20132 Milano, Italy

4 IRCCS Istituto Ortopedico Galeazzi, Via Riccardo Galeazzi, 4, 20161 Milano, Italy

Published online: 01 July 2020

COVID-19 positive and negative individuals. Thus, tests to identify the SARS–CoV-2 virus are believed to be crucial to identify positive cases to this infection and thus curb the pandemic.

To this aim, the current test of choice is the reverse transcriptase Polymerase Chain Reaction (rt-PCR)-based assays performed in the laboratory on respiratory specimens. Taking this as a gold standard, machine learning techniques have been employed to detect COVID-19 from lung CT-scans with 90% sensitivity, and high AUROC (0.95) [19, 27]. Although chest CTs have been found associated with high sensitivity for the diagnosis of COVID-19 [1], this kind of exam can hardly be employed for screening tasks, for the radiation doses, the relative low number of devices available, and the related operation costs. A similar attempt was recently performed on chest x-rays [4], which is a low-dose and less expensive test, with promising statistical performance (e.g., sensitivity 97%). However, since almost 60% of chest x-rays taken in patients with confirmed and symptomatic COVID-19 have been found to be normal [45], systems based on this exam need to be thoroughly validated in real-world settings [6]. Further, despite these promising results, some concerns have been raised on these and other works, most of which have not yet undergone peer review: a recent critical survey [46] reported that all of the surveyed studies were possibly subject to high bias and risk of over-fitting, and showed little compliance to reporting and replication standards

The public health emergency requires an unprecedented global effort to increase testing capacity [33]. The large demand for rRT-PCR tests (also commonly known as nasopharyngeal swab tests) due to the worldwide extension of the virus is highlighting the limitations of this type of diagnosis on a large-scale such as: the long turnaround times (on average over 2 to 3 hours to generate results); the need of certified laboratories; trained personnel; expensive equipment and reagents for which demand can easily overcome supply [28]. For instance in Italy, the scarcity of reagents and specialized laboratories forced the government to limit the swab testing to those people who clearly showed symptoms of severe respiratory syndrome, thus leading to a number of infected people and a contagion rate that were largely underestimated [39].

For this reason, and also in light of the predictable wide adoption of mobile apps for contact tracing [15], which will likely increase the demand for population screening, there is an urgent need for alternative (or complementary) testing methods by which to quickly identify infected COVID-19 patients to mitigate virus transmission and guarantee a prompt patients treatment.

On a previous work published in the laboratory medicine literature [14], we showed how simple blood tests might help identifying false positive/negative rRT-PCR tests. This work and the considerations made above strongly motivated us to apply machine learning methods to routine, low-cost[1] blood exams, and to evaluate the feasibility of predictive models in this important task for the mass-screening of potential COVID-19 infected individuals. A comprehensive literature review has been recently published on the use of machine learning [46] for COVID-19 screening and diagnosis; after searching on PubMed, Scopus and Web of Science search engines, we confirm the findings of the above literature review: that no machine learning solution to date is applied to blood counts and other comprehensive routine blood tests for COVID-19 screening and diagnosis. The only study, available so far in the peer-reviewed literature, that applied this approach, although in combination with CT-based diagnosis, was proposed in [32], but it was limited to white blood cell count. In what follows, we report the study that proves the feasibility of our approach.

Table 1 Features of the dataset considered in the present study

Feature	Data Type
Gender	Categorical
Age	Numerical (discrete)
Leukocytes (WBC)	Numerical (continuous)
Platelets	Numerical (continuous)
C-reactive Protein (CRP)	Numerical (continuous)
Transaminases (AST)	Numerical (continuous)
Transaminases (ALT)	Numerical (continuous)
Gamma Glutamil Transferasi (GGT)	Numerical (continuous)
Lactate dehydrogenase (LDH)	Numerical (continuous)
Neutrophils	Numerical (continuous)
Lymphocytes	Numerical (continuous)
Monocytes	Numerical (continuous)
Eosinophils	Numerical (continuous)
Basophils	Numerical (continuous)
Swab	Categorical

Methods

The aim of this work is to develop a predictive model, based on Machine Learning techniques, to predict the positivity or negativity for COVID-19. In the rest of this Section we report on the dataset used for model training and on the data analysis pipeline adopted.

Data description

The dataset used for this study was made available by the *IRCCS Ospedale San Raffaele*[2] and it consisted of

[1] A qualitative estimation of the cost of the exams used for this study is 15 euros per test, approximately five times cheaper than rt-PCR testing.

[2] IRCCS is the Italian acronym for Scientific Institute for Research, Hospitalization and Healthcare

Table 2 Descriptive statistics for the features considered in the present study

Feature	Mean	Std	Median	Kurtosis	Skewness
Age	61.3	18.5	64	-0.1	-0.5
Leukocytes (WBC)	8.5	4.8	7.2	2.3	1.5
Platelets	226.5	100.8	205	1.8	1.1
C-reactive Protein (CRP)	91.1	93.5	57.2	1.9	1.4
Transaminases (AST)	54.2	57.4	37	28.8	4.6
Transaminases (ALT)	46.6	47.1	33	11.8	3.1
Gamma Glutamil Transferasi (GGT)	82	128.8	48	14.8	2.6
Lactate dehydrogenase (LDH)	378	212.9	328	12.6	0.7
Neutrophils	6.6	4.36	5.3	3	1.6
Lymphocytes	1.2	0.7	1.1	16.1	2.7
Monocytes	0.6	0.4	0.5	8.2	2
Eosinophils	0.05	0.1	0	48.7	5.5
Basophils	0.01	0.03	0	14.2	3

279 cases, randomly extracted from patients admitted to that hospital from the end of February 2020 to mid of March 2020. Each case included the patient's age, gender, values from routine blood tests extracted as in [13], and the result of the RT-PCR test for COVID-19, performed by nasopharyngeal swab. The parameters collected by the blood test are reported in Table 1.

The dependent variable "*Swab*" is binary and it is equal to 0 in the absence of COVID-19 infection (negative swab test), and it is equal to 1 in the case of COVID-19 infection (positive to the swab test). The number of occurrences for the negative and positive class was respectively 102 (37%) and 177 (63%), thus the dataset was slightly imbalanced towards positive cases.

Table 2 summarizes the descriptive statistics of the continuous features considered in this work. In Fig. 1 we report the violin plots that show the feature distribution of the most predictive features employed to build the machine learning models of this case study.

Figure 2 shows the pairwise correlation of the features used for this study, while Fig. 3 focuses on variables "*Age*", "*WBC*", "*CRP*", "*AST*" and "*Lymphocytes*".

Data manipulation

First of all, the categorical feature *Gender* has been transformed into two binary features by *one-hot encoding*.

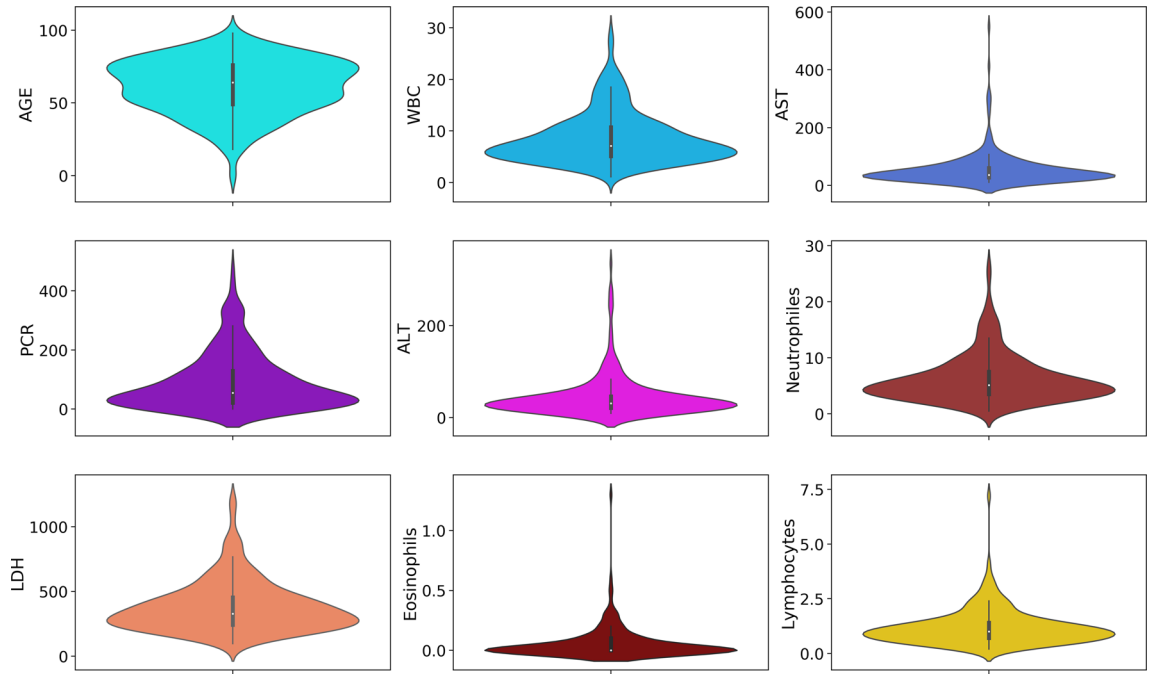

Fig. 1 Violin plots for selected features in the training dataset (chosen for their predictive importance)

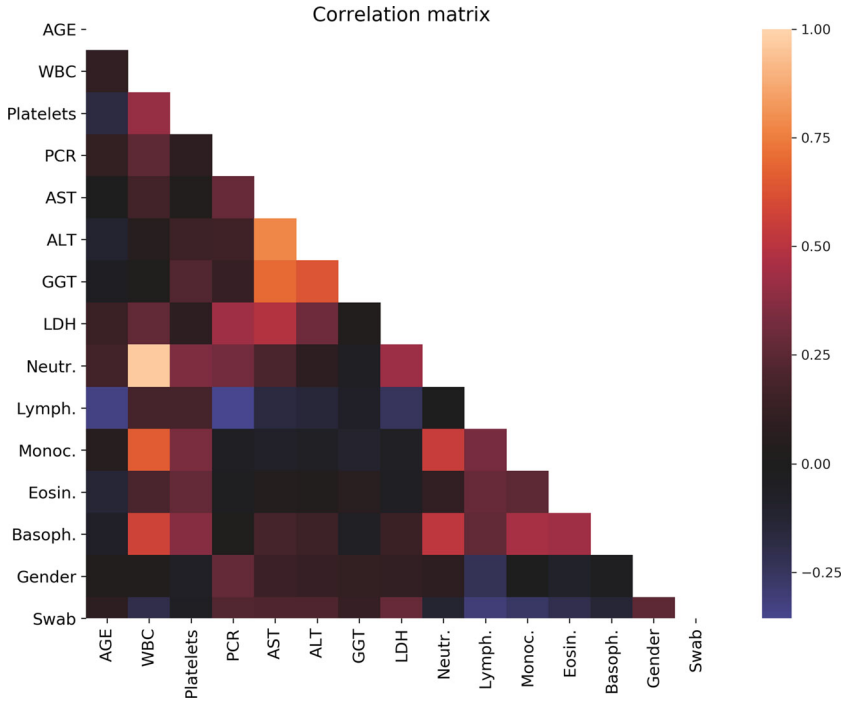

Fig. 2 Pairwise Pearson correlation of the features taken into account for this case study

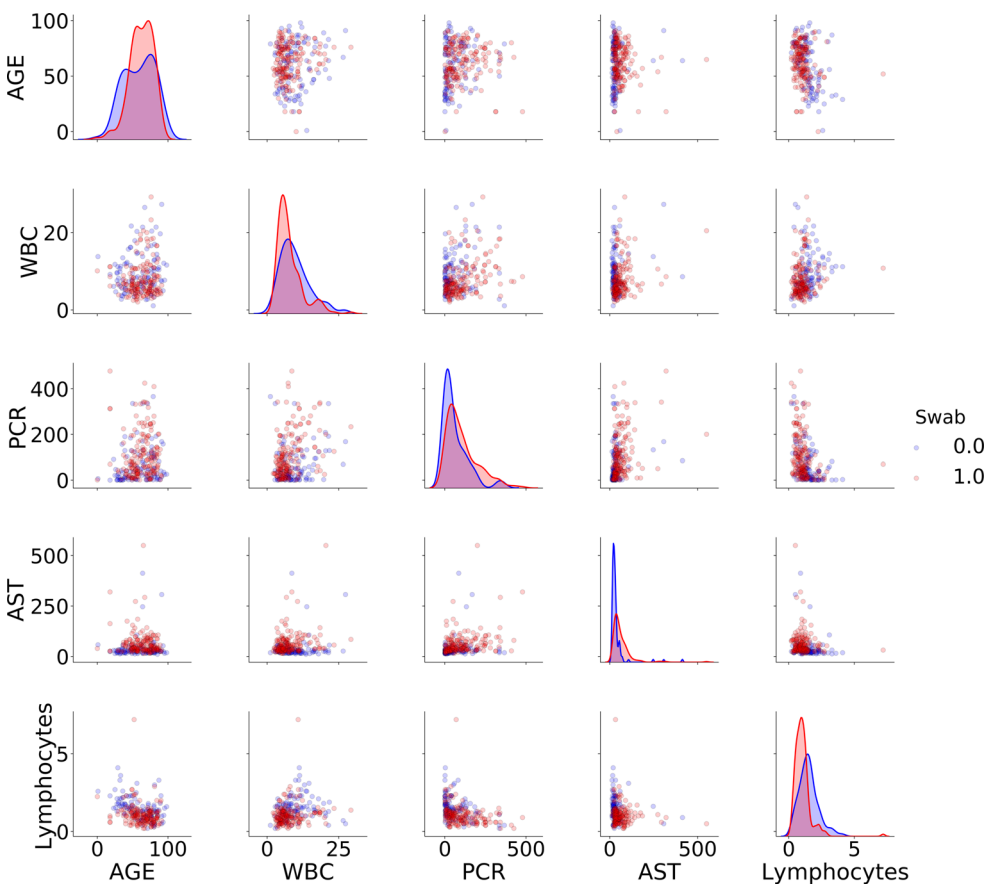

Fig. 3 Distribution plots and pairwise scatter plots of selected features. Red points and red distributions represent positive patients to Covid19, while blue points represent negative patients

Table 3 Features and missing values in the dataset

Feature	$N°$ of missing	% of missing on the total
C-reactive protein (CRP)	6	2.1
Aspartate Aminotransferase (AST)	2	0.7
Alanine Amino Transferase (ALT)	13	4.6
Gamma Glutamil Transferasi (GGT)	143	51.2
Lactate Dehydrogenase (LDH)	85	30.4
Leukocyte Count (WBC)	2	0.7
Platelets	2	0.7
Neutrophils	70	25
Lymphocytes	70	25
Monocytes	70	25
Eosinophils	70	25
Basophils	71	25.4

Further, we notice that the dataset was affected by missing values in most of its features (see Table 3).

To address data incompleteness, we performed missing data imputation by means of the *Multivariate Imputation by Chained Equation* (MICE) [5] method. MICE is a multiple imputation method that works in an iterative fashion: in each imputation round, one feature with missing values is selected and is modeled as a function of all the other features; the estimated values are then used to impute the missing values and re-used in the subsequent imputation rounds.

We chose this method because multiple imputation techniques are known to be more robust and better capable to account for uncertainty, especially when the proportion of missing values on some features may be large, compared with single imputation ones [38] (as they employ the joint distribution of the available features). Further, in order to avoid data leakage and control the bias due to imputation, we performed the missing data imputation during the nested cross-validation (described in the following section), by using for the imputation only the data in each training folds: this allows to quantify the influence of the data imputation on the results by observing the variance of the results across the folds.

Model training, selection and evaluation

We compared different classes of Machine Learning classifiers. In particular, we considered the following classifier models:

- *Decision Tree* [40] (DT);
- *Extremely Randomized Trees* [17] (ET);
- *K-nearest neighbors* [2] (KNN);
- *Logistic Regression* [21] (LR);
- *Naïve Bayes* [25] (NB);
- *Random Forest* [23] (RF);
- *Support Vector Machines* [41] (SVM).

We also considered a modification of the Random Forest algorithm, called three-way Random Forest classifier [7] (TWRF), which allows the model to abstain on instances for which it can express low confidence; in so doing, a TWFR achieves higher accuracy on the effectively classified instances at expense of coverage (i.e., the number of instances on which it makes a prediction). We decided to consider also this class of models as they could provide more reliable predictions in a large part of cases, while exposing the uncertainty regarding other cases so as to suggest further (and more expensive) tests on them.

From a technical point of view, Random Forest is an ensemble algorithm that relies on a collection of Decision Trees (i.e. a forest, hence the name of the algorithm) that are trained on mutually independent subsets of the original data in order to obtain a classifier with lower variance and/or lower bias. The independent datasets, on which the Decision Trees in the forest are trained, are obtained from an original dataset by both sampling with replacement the instance and selecting a random subset of the features (see [20] for more details about the Random Forest algorithm). As Random Forest are a class of probability scoring classifiers (that is, for each instance the model assigns a probability score for every possible class), the abstention is performed on the basis of two thresholds $\alpha, \beta \in [0, 1]$: if we denote with 1 the positive class and 0 the negative class, then each instance is classified as positive if $score(1) > \alpha$ and $score(1) > score(0)$, negative if $score(0) > \beta$ and $score(0) > score(1)$ and, otherwise, the model abstains. In these models the performance is usually evaluated only on the non-abstained instances [16], and the coverage is a further performance element to be considered.

The models mentioned above have been trained, and evaluated, through a *nested cross validation* [9, 20] procedure. This procedure allows for an unbiased generalization error estimation while the hyperparameter search (including feature selection) is performed: an inner cross-validation loop is executed to find the optimal hyperparameters via grid search and an outer loop evaluates the model performance on *five folds*.

Models were evaluated in terms of *accuracy*, *balanced accuracy* [3], *Positive Predictive Value* (PPV)[4], *sensitivity*, *specificity* and, except for the three-way Random Forest, the *area under the ROC curve* (AUC). After discussing this with

[3] We recall that balanced accuracy is defined as the average of sensitivity and specificity. If accuracy and balanced accuracy significantly differ, the data could be interpreted as unbalanced with respect to class prevalence.

[4] We recall here that PPV represents the probability that subjects with a positive screening test truly have the disease.

Table 4 The models' performance: 95% C.I. of model accuracy on *5-folds nested CV*

	DT	ET	KNN	LR	NB	RF	SVM	TWRF
A (all features)	[0.70, 0.78]	[0.68, 0.79]	[0.66, 0.76]	[0.70, 0.81]	[0.64, 0.81]	[0.74, 0.80]	[0.69, 0.80]	[0.83, 0.89]
B (without Gender)	[0.62, 0.75]	[0.74, 0.82]	[0.66, 0.76]	[0.670, 0.79]	[0.65, 0.76]	[0.71, 0.86]	[0.66, 0.79]	[0.83, 0.89]

the clinicians involved in this study, we considered accuracy and sensitivity to be the main quality metrics, since false negatives (that is, patients positive to COVID-10 which are, however, classified as negative, and possibly let go home) are more harmful than false positives in this screening task.

Results

For all the preprocessing steps and tested classifiers, we employed the standard Python data analysis ecosystem, comprising *pandas* [31] (for data loading and pre-processing), *scikit-learn* [35] (for both pre-processing and the classifiers implementations) and *matplotlib* [22] (for visualization purposes). The experiments were executed on a PC with an Intel i7 processor (6 cores, 3.2 GHz clock frequency) and 12 GB RAM: the model selection required around 2 minutes of computation time, while both the model fitting on the training set and the predictions on the test/validation sets required around 1 second of computation time.

Tables 4 and 5 show the 95% confidence intervals of, respectively, average accuracy and average balanced accuracy (that is, the average of sensitivity and specificity) of the models (on the nested cross-validation) trained on the two best-performing sets of features: the first one, dataset A, includes all the variables, while the second one, dataset B, excludes the "*Gender*" variable, as this was found of negligible predictive value.

Figure 4 shows the performance of the traditional models (i.e., the TWRF model was excluded) on the *nested cross-validation*.

To further validate the above findings, the entire dataset has been splitted into training and test/validation sets, respectively the 80% and the 20% of the total instances. The performance of the models, with optimal hyper-parameters as selected through nested cross-validation, is shown in Fig. 5, which depicts the ROC curves for all the models. Two models, Logistic Regression and Random Forest, exhibited comparable performance (difference less than 1%) in terms of AUC (LR = 85%, RF = 84 %) and sensitivity (LR = 93%, RF = 92%), but Random Forest reported higher performance in terms of accuracy (LR = 78%, RF = 82%) and much higher specificity (LR = 50%, RF = 65%): thus, Random Forest was selected as reference best performing model. The best performing model, i.e. the Random Forest classifier, trained on dataset B, achieved the following results on the test/validation set: accuracy = 82% , sensitivity = 92%, PPV = 83%, specificity = 65%, AUC = 84%. Figure 6 shows the performance of this model in the precision/recall space.

The optimal hyperparameters found are shown in Table 6.

Similarly, for the best three-way Random Forest classifier on the validation set we observed: accuracy = 86%, sensitivity = 95%, PPV = 86%, specificity = 75%, coverage = 70% (that is, for 30% of the validation instances the model abstained).

The feature importance assessed for the the best performing model (Random Forest on dataset B), are shown in Fig. 7. The feature importances were computed by estimating, for each feature, the total normalized reduction, across the Decision Trees in the trained Random Forest, to the variance of the target feature (hence, greater importance values denote a greater contribution to explaining the target variance): this computation was performed via the reference Random Forest implementation provided in the scikit-learn library.

Finally, it is worth noting that while the best performing model obtained good predictive performance, Random Forest is known to be a black-box model, that is a model that is not directly able to provide interpretable insight into how its predictions are made, as these predictions are obtained from the averaging of the Decision Trees in the forest. In order to provide an interpretable overview (in the sense of eXplainable AI[18]) of this predictive model, we also developed a Decision Tree model, which is shown in Fig. 8, to approximate the decision-making steps implemented by

Table 5 The models' performance: 95% C.I. of model balanced accuracy on *5-folds nested CV*

	DT	ET	KNN	LR	NB	RF	SVM	TWRF
A (all features)	[0.64, 0.71]	[0.67, 0.81]	[0.60, 0.74]	[0.65, 0.79]	[0.63, 0.77]	[0.70, 0.82]	[0.69, 0.76]	[0.83, 0.87]
B (without Gender)	[0.63, 0.73]	[0.67, 0.84]	[0.61, 0.74]	[0.64, 0.74]	[0.63, 0.76]	[0.70, 0.80]	[0.65, 0.77]	[0.83, 0.87]

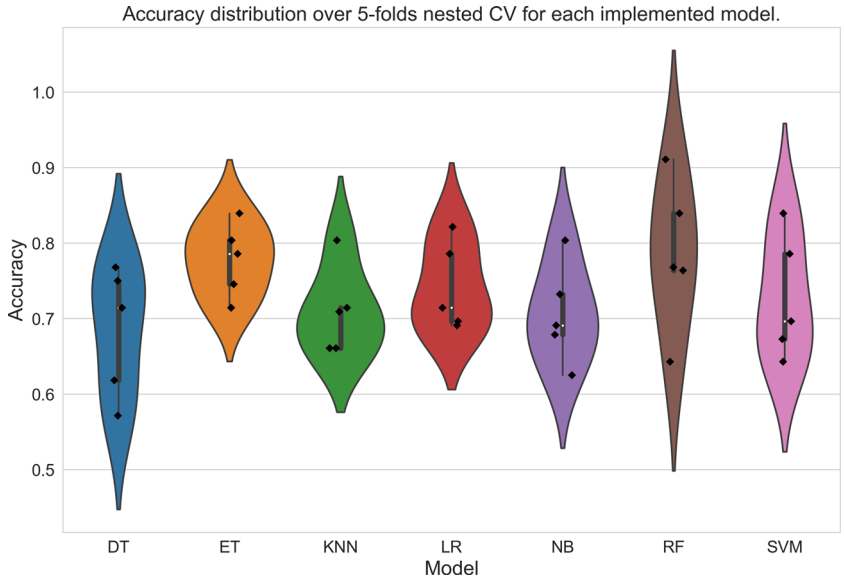

Fig. 4 Violin plots of the accuracy distributions reached by each models on five folds (on dataset B)

the Random Forest model. Although the depicted decision tree is associated with a lower discriminative performance than the two former (inscrutable) models, such a tree can be used as a simple decision aid by clinicians interested in the use of blood values to assess COVID-19 suspect cases.

Discussion

We have developed two machine learning models to discriminate between patients who are either positive or negative to the SARS-CoV-2, which is the coronavirus causing the COVID-19 pandemia. In this task, patients are represented in terms of few basic demographic characteristics (gender, age) and a small array of routine blood tests, chosen for their convenience, low cost and because they are usually available within 30 minutes from the blood draw in regular emergency department. The ground truth was established through RT-PCR swab tests.

We presented the best traditional model, as it is common practice, and a three-way model, which guarantees best sensitivity and positive predictive value: the former is the proportion of infected (and contagious) people who will have a positive result and therefore it is useful to clinicians when deciding which test to use. On the other hand, PPV is useful for patients as it tells the odds of one having COVID-19 if they have a positive result.

The performance achieved by these two best models (sensitivity between 92% and 95%, accuracy between 82%

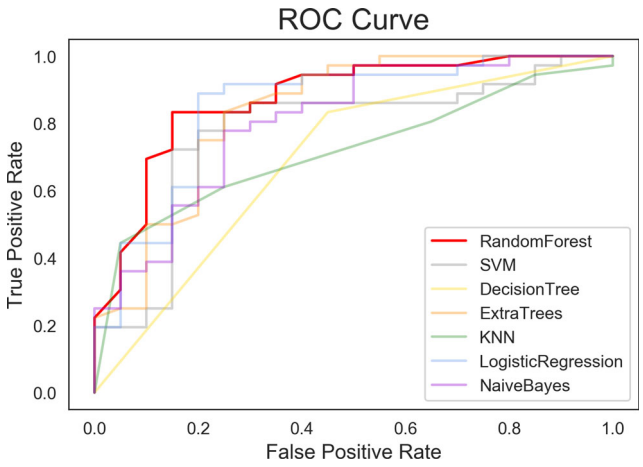

Fig. 5 The sensitivity and specificity curve (i.e., sensitivity /positive predictive value curve or, equivalently true positive rate / false positive rate as depicted in the Figure) of the evaluated models. The best performing algorithm, Random Forest, is highlighted

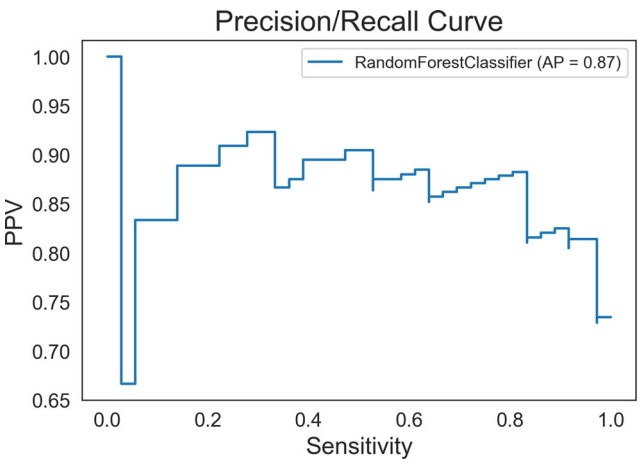

Fig. 6 The precision/recall (i.e., positive predictive value / sensitivity) curve, and the area under this curve

Table 6 Optimal hyperparameters for the Random Forest classifier. For the sake of reproducibility, also the random seed is reported

Hyperparameters	Value
Max Depth	-
Criterion	Gini
$N°$ estimators	100
Random seed for reproducibility	123

and 86%) provides proof that this kind of data, and computational models, *can be used* to discriminate among potential COVID-19 infectious patients with sufficient reliability, and similar sensitivity to the current Gold Standard. This is the most important contribution of our study.

Also from the clinical point of view, the feature selection was considered valid by the clinicians involved. Indeed, the specialist literature has found that COVID-19 positivity is associated with lymphopenia (that is, abnormally low level of white blood cells in the blood), damage to liver and muscle tissue [44, 48], and significantly increased C-reactive protein (CRP) levels [10]. In [29] a comprehensive list of the most frequent abnormalities in COVID-19 patients has been reported: among the 14 conditions considered, they report increased aspartate aminotransferase (AST), decreased lymphocyte count (WBC), increased lactate dehydrogenase (LDH), increased C-reactive protein (CRP), increased white blood cell count (WBC) and increased alanine aminotransferase (ALT).

These parameters are also the most predictive features identified by the best classifier (Random Forest), all together with the Age attribute. Also other studies confirm the relevance of these features and their association with the COVID-19 positivity [8, 34, 37, 50], compared to other kinds of pneumonia [49]. This also gives confirmation that our models ground on clinically relevant features and that most of these values can be extracted from routine blood exams.

The interpretable Decision Tree model provides a further confirmation (see Fig. 8) of the soundness of the approach: the clinicians (ML, GB) and the biochemist (DF) involved in this study found reasonable that the AST would be the first parameter to consider (i.e., mirrored by the fact that AST was the root of the decision tree) and that it was found to be the most important predictive feature. Indeed, values of AST below 25 are good predictors of COVID-19 positivity (accuracy = PPV = 76%), while values below 25 are a good predictor of COVID-19 negativity (accuracy = Negative Predictive Value = 83%). Similar observations can also be made about CRP, Lymphocytes and general WBC counts.

No statistically significant difference was found between the accuracy and the balanced accuracy of the models (as mirrored by the overlap of the 95% confidence intervals), as a sign that the dataset was not significantly unbalanced.

Moreover, we can notice that the best performing ML classifier (Random Forest) exhibited a very high sensitivity ($\sim$ 90%) but, in comparison, a limited specificity of only 65%. That gives the main motivation for the three-way classifier: this model offers a trade-off between increased specificity (a 10% increment compared with the best traditional ML model) and reduced coverage, as the three-way approach abstains on uncertain instances (i.e., the cases that cannot be classified with high confidence neither as positive nor negative). This means that the model yields more robust and reliable prediction for the classified instances (as it is mirrored by the increase in all of the performance measures), while for the other ones it is anyway useful in suggesting further tests, e.g., by either a PCR-RNA swab test or a chest x-ray.

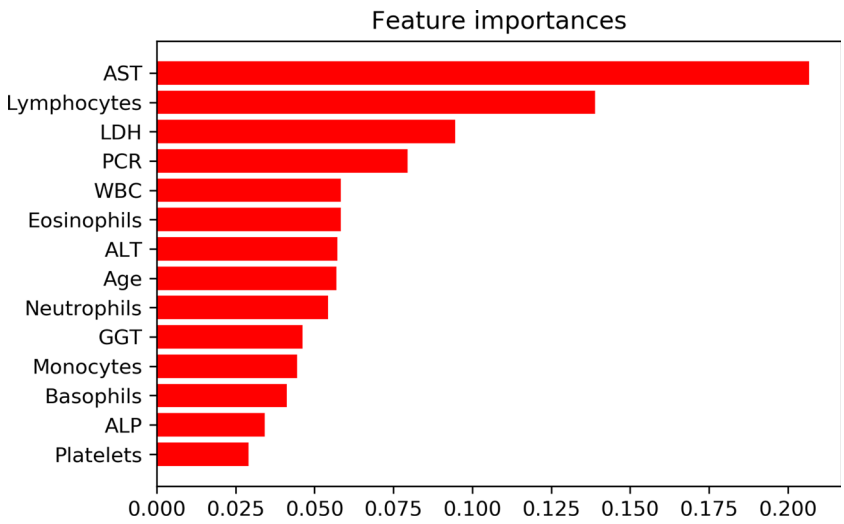

Fig. 7 Feature importance scores for the best performing model

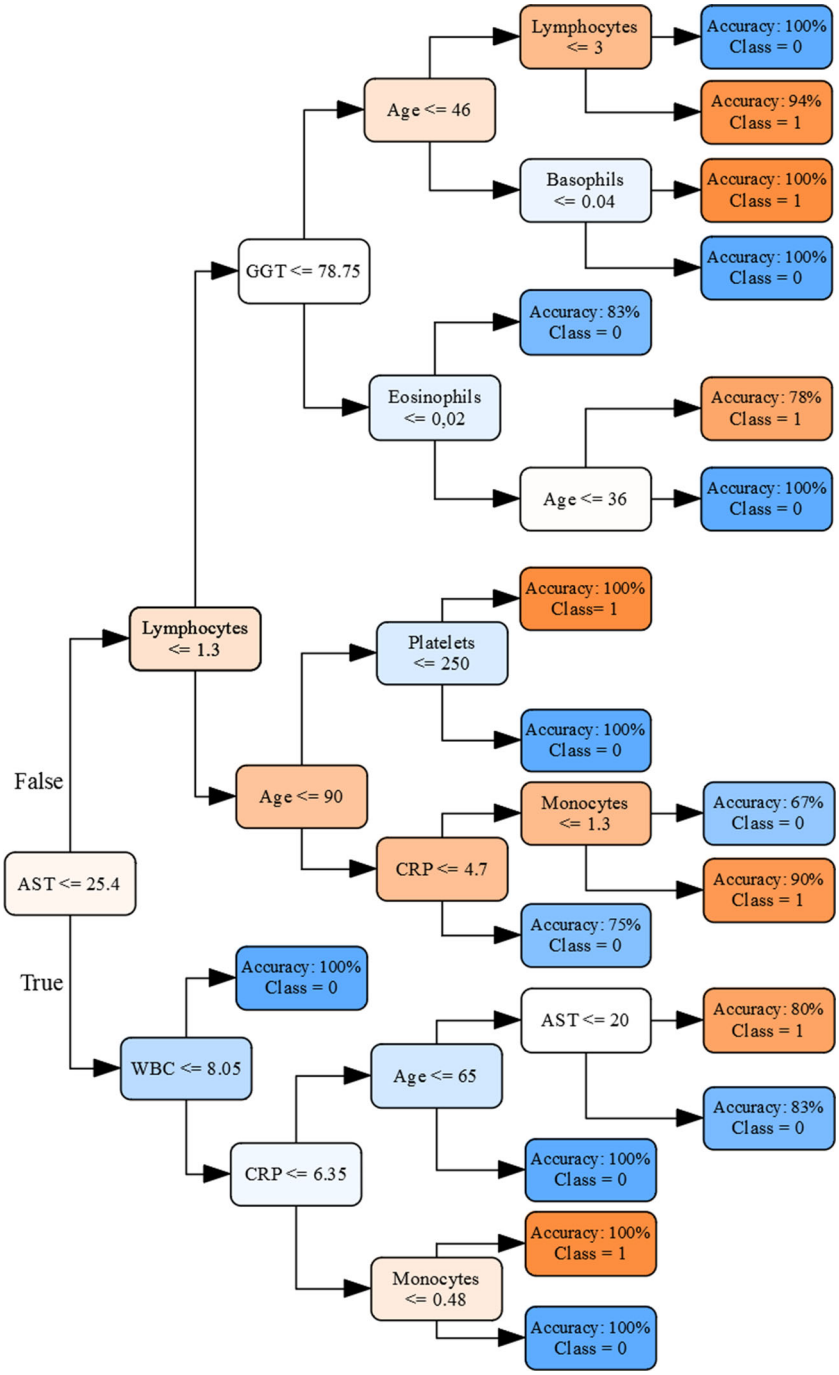

Fig. 8 An interpretable Decision Tree, developed in order to support the interpretation of the predictions from the other models. Color gradients denote predictivity for either classes (shades of blue correspond to COVID-19 negativity, shades of orange to positivity)

In regard to the specificity exhibited by our models, we can further notice that even while these values are relatively low compared with other tests (which are more specific but slower and less accessible), this may not be too much of a limitation as there is a significant disparity between the costs of false positives and false negatives and in fact our models favors sensitivity (thus, they avoid false negatives). Further, the high PPV (> 80%) of our models suggest that the large majority of cases identified as positives by our models would likely be COVID-19 positive cases.

That said, the study presents two main limitations: the first, and more obvious one, regards the relatively low number of cases considered. This was tackled by performing nested cross-validation in order to control for bias [43], and by employing models that are known to be effective also with moderately sized samples [3, 36, 42]. Nonetheless, further research should be aimed at

confirming our findings, by integrating hematochemical data from multiple centers and increasing the number of the cases considered. The second limitation may be less obvious, as it regards the reliability of the ground truth itself. Although this was built by means of the current gold standard for COVID-19 detection, i.e., the rRt-PCR test, a recent study observed that the accuracy of this test may be highly affected by problems like inadequate procedures for collection, handling, transport and storage of the swabs, sample contamination, and presence of interfering substances, among the others [30]. As a result, some recent studies have reported up to 20% false-negative results for the rRt-PCR test [24, 26, 47], and a recent systematic review reported an average sensitivity of 92% and cautioned that "up to 29% of patients could have an initial RT-PCR false-negative result". Thus, contrary to common belief and some preliminary study (e.g., [11]), the accuracy of this test could be less than optimal, and this could have affected the reliability of the ground truth also in this study (as in any other using this test for ground truthing, unless cases are annotated after multiple tests. However, besides being a limitation, this is also a further motivation to pursue alternative ways to perform the diagnosis of SARS-CoV-2 infection, such as our methods are.

Future work will be devoted to the inclusion of more hematochemical parameters, including those from arterial blood gas assays (ABG), to evaluate their predictiveness with respect to COVID-19 positiveness, and the inclusion of cases whose probability to be COVID-positive is almost 100%, as they resulted positive to two or more swabs or to serologic antibody tests. This would allow to associate a higher weight with misidentifying those cases, so as, we conjecture, improve the sensitivity further.

Moreover, we want to investigate the interpretability of our models further, by both having more clinicians validate the current Decision Tree, and possibly construct a more accurate one, so that clinicians can use it as a convenient decision aid to interpret blood tests in regard to COVID-19 suspect cases (even off-line).

Finally, this was conceived as a feasibility study for an alternative COVID-19 test on the basis of hematochimical values. IN virtue of this ambitious goal, the success of this study does not exempt us from pursuing a real-world, *ecological* validation of the models [6]. To this aim, we deployed an online Web-based tool[5] by which clinicians can test the model, by feeding it with clinical values, and considering the sensibleness and usefulness of the indications provided back by the model. After this successful feasibility study, we will conceive proper external validation tasks and undertake an ecological validation to assess the cost-effectiveness and utility of these models for the screening of COVID-19 infection in all the real-world settings (e.g., hospitals, workplaces) where routine blood tests are a viable test of choice.

[5]The tool is available at the following address: https://covid19-blood-ml.herokuapp.com/.

Code and data availability

Availability of data and material

The developed web tool is available at the following address: https://covid19-blood-ml.herokuapp.com/ The complete dataset will be made available on the Zenodo platform as soon as the work gets accepted for publication.

Code availability

The complete code will be made available on the Zenodo platform as soon as the work gets accepted for publication.

Funding Information This research was sponsered by the Research on the Major Scientific Instrument of National Natural Science Foundation of China (61727809).

Compliance with Ethical Standards

Conflict of interests The authors have declared no conflict of interest.

Ethical approval Research involving human subjects complied with all relevant national regulations, institutional policies and is in accordance with the tenets of the Helsinki Declaration (as revised in 2013), and was approved by the authors' Institutional Review Board on the 20th of April.

Informed consent Individuals signed an informed consent authorizing the use of their anonymously collected data for retrospective observational studies (article 9.2.j; EU general data protection regulation 2016/679 [GDPR]), according to the IRCCS San Raffaele Hospital policy (IOG075/2016), and the appropriate institutional forms have been archived.

References

1. Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., and Xia, L., Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. Radiology p 200642, 2020.
2. Altman, N. S., An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46(3):175–185, 1992.
3. Anguita, D., Ghio, A., Greco, N. et al., Model selection for support vector machines: Advantages and disadvantages of the machine learning theory. In: *IJCNN-2010*, pp. 1–8, 2010. https://doi.org/10.1109/IJCNN.2010.5596450.

4. Apostolopoulos, I. D., and Mpesiana, T. A., Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Physical and Engineering Sciences in Medicine 1, 2020.
5. van Buuren, S., and Groothuis-Oudshoorn, K., mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software, Articles* 45(3):1–67, 2011. https://doi.org/10.18637/jss.v045.i03 https://www.jstatsoft.org/v045/i03.
6. Cabitza, F., and Zeitoun, J. D., The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. Annals of translational medicine 7(8), 2019.
7. Campagner, A., Cabitza, F., and Ciucci, D., The three-way-in and three-way-out framework to treat and exploit ambiguity in data. *International Journal of Approximate Reasoning* 119:292–312, 2020.
8. Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., and Di Napoli, R., Features, evaluation and treatment coronavirus (covid-19). In: *StatPearls [Internet], StatPearls Publishing*, 2020.
9. Cawley, G. C., and Talbot, N. L., On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11(Jul):2079–2107, 2010.
10. Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y. et al., Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The Lancet* 395(10223):507–513, 2020.
11. Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K., Bleicker, T., Brünink, S, Schneider, J., Schmidt, M. L., and at al., Detection of 2019 novel coronavirus (2019-ncov) by real-time rt-pcr. *Eurosurveillance* 25(3):2000045, 2020.
12. Day, M., Covid-19: identifying and isolating asymptomatic people helped eliminate virus in italian village. *BMJ* 368:m1165, 2020.
13. Ferrari, D., Lombardi, G., Strollo, M., Pontillo, M., Motta, A., and Locatelli, M., Association between solar ultraviolet doses and vitamin d clinical routine data in european mid-latitude population between 2006 and 2018. *Photochemical & Photobiological Sciences* 18(11):2696–2706, 2019.
14. Ferrari, D., Motta, A., Strollo, M., Banfi, G., and Massimo, L., Routine blood tests as a potential diagnostic tool for covid-19. Clinical Chemistry and Laboratory Medicine (2) https://doi.org/10.1515/cclm-2020-0398, 2020.
15. Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L, Parker, M., Bonsall, D., and Fraser, C., Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. Science, 2020.
16. Ferri, C., and Hernández-Orallo, J, Cautious classifiers. *ROCAI* 4:27–36, 2004.
17. Geurts, P., Ernst, D., and Wehenkel, L., Extremely randomized trees. *Machine learning* 63(1):3–42, 2006.
18. Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., and Holzinger, A., *Explainable ai: the new 42? In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 295–303. New York: Springer, 2018.
19. Gozes, O., Frid-Adar, M., Sagie, N., Zhang, H., Ji, W., and Greenspan, H., Coronavirus detection and analysis on chest ct with deep learning. arXiv:200402640, 2020.
20. Hastie, T., Tibshirani, R., and Friedman, J., The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.
21. Hosmer, DW Jr., Lemeshow, S., and Sturdivant, R. X., *Applied logistic regression*. Vol. 398. New York: John Wiley & Sons, 2013.
22. Hunter, J. D., Matplotlib: a 2d graphics environment. *Computing in science & engineering* 9(3):90–95, 2007.
23. Kam, H. T., Random decision forest. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada, August, vol 1416, p 278282*, 1995.
24. Kim, S., Kim, D. M., and Lee, B., Insufficient sensitivity of rna dependent rna polymerase gene of sars-cov-2 viral genome as confirmatory test using korean covid-19 cases, 2020.
25. Lewis, D. D., Naive (bayes) at forty: The independence assumption in information retrieval. In: *European conference on machine learning*, pp. 4–15. New York: Springer, 1998.
26. Li, D., Wang, D., Dong, J., Wang, N., Huang, H., Xu, H., and Xia, C., False-negative results of real-time reverse-transcriptase polymerase chain reaction for severe acute respiratory syndrome coronavirus 2: Role of deep-learning-based ct diagnosis and insights from two cases. *Korean journal of radiology* 21(4):505–508, 2020a.
27. Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q. et al., Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology p 200905, 2020b.
28. Li, Z., Yi, Y., Luo, X., Xiong, N., Liu, Y., Li, S., Sun, R., Wang, Y., Hu, B., Chen, W. et al., Development and clinical application of a rapid igm-igg combined antibody test for sars-cov-2 infection diagnosis. Journal of medical virology, 2020c.
29. Lippi, G., and Plebani, M., Laboratory abnormalities in patients with covid-2019 infection. Clinical Chemistry and Laboratory Medicine (CCLM)1(ahead-of-print), 2020.
30. Lippi, G., Simundic, A. M., and Plebani, M., Potential preanalytical and analytical vulnerabilities in the laboratory diagnosis of coronavirus disease 2019 (covid-19). Clinical Chemistry and Laboratory Medicine (CCLM) 1(ahead-of-print), 2020.
31. McKinney, W., et al., Pandas: a foundational python library for data analysis and statistics. Python for High Performance and Scientific Computing 14(9), 2011.
32. Mei, X., Lee, H. C., Ky, D., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P. M., Chung, M. et al., Artificial intelligence–enabled rapid diagnosis of patients with covid-19. Nature Medicine pp 1–55, 2020.
33. MP, C., et al., and PJ, Diagnostic testing for severe acute respiratory syndrome–related coronavirus-2: A narrative review. Annals of Internal Medicine. https://doi.org/10.7326/M20-1301, 2020.
34. Pan, F., Ye, T., Sun, P., Gui, S., Liang, B., Li, L., Zheng, D., Wang, J., Hesketh, R. L., Yang, L. et al., Time course of lung changes on chest ct during recovery from 2019 novel coronavirus (covid-19) pneumonia. Radiology p 200370, 2020.
35. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al., Scikit-learn:, Machine learning in python. *The Journal of machine Learning research* 12:2825–2830, 2011.
36. Qi, Y., Random forest for bioinformatics. In: *Ensemble machine learning*, pp. 307–323. New York: Springer, 2012.
37. Qin, C., Zhou, L., Hu, Z., Zhang, S., Yang, S., Tao, Y., Xie, C., Ma, K., Shang, K., Wang, W. et al., Dysregulation of immune response in patients with covid-19 in wuhan, china. China (February 17, 2020), 2020.
38. Rubin, D. B., *Multiple imputation for nonresponse in surveys*. Vol. 81. New York: John Wiley & Sons, 2004.
39. Rubino, S., Kelvin, N., Bermejo-Martin, J. F., and Kelvin, D., As covid-19 cases, deaths and fatality rates surge in italy, underlying causes require investigation. *The Journal of Infection in Developing Countries* 14(03):265–267, 2020.
40. Safavian, S. R., and Landgrebe, D., A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics* 21(3):660–674, 1991.

41. Schölkopf, B, Smola, A. J., Bach, F., et al., and 2002, Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.
42. Skurichina, M., and Duin, R. P. W., Stabilizing classifiers for very small sample sizes. In: *Proceedings of ICPR-1996*, Vol. 2, pp. 891–896, 1996.
43. Varma, S., and Simon, R., Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics* 7(1):91, 2006.
44. Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y., and et al, Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china. *Jama* 323(11):1061–1069, 2020.
45. Mea, W., Chest x-ray findings in 636 ambulatory patients with covid-19 presenting to an urgent care center: a normal chest x-ray is no guarantee. The Journal of Urgent Care Medicin (2):1–9, 2020.
46. Wynants, L., Van Calster, B., Bonten, M. M., Collins, G. S., Debray, T. P., De Vos, M., Haller, M. C., Heinze, G., Moons, K. G., Riley, R. D. et al., Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. bmj 369, 2020.
47. Xie, X., Zhong, Z., Zhao, W., Zheng, C., Wang, F., and Liu, J., Chest ct for typical 2019-ncov pneumonia: relationship to negative rt-pcr testing. Radiology p 200343, 2020.
48. Zhang, C., Shi, L., and Wang, F. S., Liver injury in covid-19: management and challenges. The Lancet Gastroenterology & Hepatology, 2020.
49. Zhao, D., Yao, F., Wang, L., Zheng, L., Gao, Y., Ye, J., Guo, F., Zhao, H., and Gao, R., A comparative study on the clinical features of COVID-19 pneumonia to other pneumonias. Clinical Infectious Diseases. https://doi.org/10.1093/cid/ciaa247 https://academic.oup.com/cid/article-pdf/doi/10.1093/cid/ciaa247/32894214/ciaa247.pdf, 2020.
50. Zheng, M., Gao, Y., Wang, G., Song, G., Liu, S., Sun, D., Xu, Y., and Tian, Z., Functional exhaustion of antiviral lymphocytes in covid-19 patients. Cellular & Molecular Immunology 1–3, 2020.