Received October 16, 2020, accepted October 19, 2020, date of publication October 26, 2020, date of current version November 10, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3034032*

# Explainable Machine Learning for Early Assessment of COVID-19 Risk Prediction in Emergency Departments

**ELENA CASIRAGHI[1,2], (Member, IEEE), DARIO MALCHIODI[1,2,3], GABRIELLA TRUCCO[1], MARCO FRASCA[1], LUCA CAPPELLETTI[1], TOMMASO FONTANA[4], ALESSANDRO ANDREA ESPOSITO[5], EMANUELE AVOLA[6], ALESSANDRO JACHETTI[7], JUSTIN REESE[8], ALESSANDRO RIZZI[1], (Member, IEEE), PETER N. ROBINSON[9], AND GIORGIO VALENTINI[1,2,3]**

[1]Department of Computer Science "Giovanni degli Antoni," Università degli Studi di Milano, 20133 Milan, Italy
[2]CINI National Laboratory of Artificial Intelligence and Intelligent Systems (AIIS), Università di Roma, 00185 Rome, Italy
[3]Data Science Research Center, Università degli Studi di Milano, 20133 Milan, Italy
[4]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy
[5]Radiology Department, Fondazione IRCCS Ca Granda Ospedale Maggiore Policlinico, 20122 Milan, Italy
[6]Postgraduate School in Radiodiagnostics, Università degli Studi di Milano, 20122 Milan, Italy
[7]Accident and Emergency Department, Fondazione IRCCS Ca Granda Ospedale Maggiore Policlinico, 20122 Milan, Italy
[8]Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[9]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA

Corresponding authors: Elena Casiraghi (elena.casiraghi@unimi.it) and Dario Malchiodi (dario.malchiodi@unimi.it)

This work was supported in part by the Università degli Studi di Milano through the Piano di Sostegno alla ricerca 2019 Grant.

**ABSTRACT** Between January and October of 2020, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus has infected more than 34 million persons in a worldwide pandemic leading to over one million deaths worldwide (data from the Johns Hopkins University). Since the virus begun to spread, emergency departments were busy with COVID-19 patients for whom a quick decision regarding in- or outpatient care was required. The virus can cause characteristic abnormalities in chest radiographs (CXR), but, due to the low sensitivity of CXR, additional variables and criteria are needed to accurately predict risk. Here, we describe a computerized system primarily aimed at extracting the most relevant radiological, clinical, and laboratory variables for improving patient risk prediction, and secondarily at presenting an explainable machine learning system, which may provide simple decision criteria to be used by clinicians as a support for assessing patient risk. To achieve robust and reliable variable selection, Boruta and Random Forest (RF) are combined in a 10-fold cross-validation scheme to produce a variable importance estimate not biased by the presence of surrogates. The most important variables are then selected to train a RF classifier, whose rules may be extracted, simplified, and pruned to finally build an associative tree, particularly appealing for its simplicity. Results show that the radiological score automatically computed through a neural network is highly correlated with the score computed by radiologists, and that laboratory variables, together with the number of comorbidities, aid risk prediction. The prediction performance of our approach was compared to that that of generalized linear models and shown to be effective and robust. The proposed machine learning-based computational system can be easily deployed and used in emergency departments for rapid and accurate risk prediction in COVID-19 patients.

**INDEX TERMS** Associative tree, Boruta feature selection, clinical data analysis, COVID-19, generalized linear models, missing data imputation, random forest classifier, risk prediction.

## I. INTRODUCTION

Coronavirus disease 2019 (COVID-19), caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), emerged in Wuhan, China, in December 2019. COVID-19 quickly became a pandemic [1] and is still threatening the lives of populations worldwide. Given the promising results achieved by studies exploiting Artificial Intelligence (AI) and/or probabilistic models for outcome prediction [2]–[4] in bio-medical problems where

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

human skill and know-how are not able to provide precise and reproducible solutions, in this worldwide health crisis, a great deal of research effort has been devoted to the development of Robotics and Data Analytics techniques [5], [6] exploiting the potentials of AI methods to either predict, monitor, and combat the virus by simulating the virus spread or the time needed to recover [7] ensuring and promoting social distancing [5], [8], identifying early COVID-19 infections, or predicting patient outcome to improve patient care [8]–[11]. Thanks to such AI techniques, wearable devices and Web applications may be used by affected individuals for self-monitoring COVID-19 related symptoms, while clinicians are aided in the diagnosis of COVID-19 infection from either CT [12], [13] or XRay lung images [11], [14], or in the prediction of patient mortality risk, progression to severe disease, intensive care unit admission, ventilation, intubation, or length of hospital stay [8], [15].

In particular, an effective risk prediction model would contribute to precision medicine strategies for tailoring clinical management to the needs of individual patients and thereby increasing the probability of complete recovery. It would also allow emergency departments to optimize patient flow and reduce waiting times.

A substantial amount of research has therefore been been conducted with the goal of predicting patient outcome by analysing different types of data, including clinical, laboratory, and radiological features [15]–[17]. Although promising results have been reported by several authors, a recent survey of COVID-19 prognosis/risk prediction methods [15] reported that most of them are biased due to one of two reasons. Firstly, many published studies lack clinical follow-up data, implying that the categories (labels) used for machine learning may be inaccurate, because patients may develop severe complications subsequently to the initial clinical encounter used for ML. Secondly, many studies use the last available predictor measurements from electronic health records, rather than the predictor values acquired at the time when the model is intended for use. Moreover, several methods do not include any description of the study population, or the intended use of the developed models, are not explained clearly, or are not exhaustively tested. In other cases, parameter values are arbitrarily set, or the experiments at the base of hyper-parameter setting are not robust or are not reported. These considerations lead to the conclusion that most of the presented methods are poorly described and at high risk of bias, raising concern that their predictions could be unreliable when applied in clinical practice [15].

In this article, we therefore aimed to develop a rigorous and explainable risk prediction model that avoids weaknesses mentioned above. Each of the relevant steps of our algorithm was critically designed, tested, and compared to state-of-the-art techniques from the published literature. Moreover, the dataset used to develop and test the algorithms is described in detail (see Section III), and each of the methods and parameter settings used by our approach is described and motivated. The principal aim of this study is to develop an unbiased automatic system primarily devoted to selecting the most important clinical and laboratory variables to be used for COVID-19 risk assessment. Importantly, the variables considered in the present study also include two radiological scores resulting from radiologists' evaluation of CXRs and two ''lung involvement'' scores computed by one of the best performing deep neural networks aimed at COVID-19 risk diagnosis. This allows assessing the integration of a radiological score computed by humans and radiological scores computed by a deep network (see Section II.B), to assess the trustworthiness of computerized AI systems, whose disadvantage is often related to their ''black box'' nature.

To properly manage the missing data issue that arises from the integration of multiple sources of data for COVID-19 risk prediction, we assessed a number of imputation techniques (see Section IV), including methods that do not assume Normality of the data [18]–[20] and several other methods that have been shown to be effective [21]–[23].

Secondarily, in Section V we present a novel feature selection technique exploiting a cross-validation strategy to combine the Boruta [24]–[26] algorithm and a permutation-based feature selector embedded into Random Forests (RFs, [27]). The proposed feature selection method enables robust and stable feature selection (Section V-A1).

The selected features are then used as input to RFs [27] (see Section V-A2) and to the derived Associative Trees (AT, [28], [29], see Section V-A3). While RFs produce a great number of rules, sometimes difficult to be understood, ATs are constructed from RFs, to essentially summarize them producing a simpler rule set that can easily be evaluated and interpreted by clinicians.

RFs and ATs were chosen since their interpretability does not require any (post hoc, proxy) explainer model to analyze their predictions, therefore avoiding unreliable and misleading explanations [30]. Results computed by these algorithms were compared to those obtained by Generalized Linear Models (GLM [31], Section V-B), which assume a binomial distribution for the response variable, therefore removing the normality hypothesis, and estimate a linear regression model ''linked'' to the response variable through a logit distribution. To avoid any bias [15], since the rules are intended to be used at the time of admission of patients to the Emergency Department (ED), our dataset is composed of data acquired upon ED admission of 300 patients, for whom five-months of follow-up data are available, and whose CXR was evaluated by two experienced radiologists blinded to patient status.

In sum, the main contributions and novelties of this article are:

- A machine learning-based computational system that can be easily deployed and used in emergency departments for an early and fast assessment of risk prediction in COVID-19 patients.
- Integration of clinical, laboratory and radiological data for the prediction of COVID-19 disease risk. The integrated prediction system includes radiological scores

estimated by both expert radiologists and by specialized state-of-the-art deep neural networks.
- A novel, robust feature selection algorithm that combines the Boruta algorithm [24], [25] with permutation-based feature selection methods embedded in RFs [27], [32], [33] to select variables that are most relevant for COVID-19 risk prediction.
- An explainable machine learning decision system based on Additive Trees that can support physicians in the early COVID-19 risk assessment through a set of simple and human-interpretable decision rules.
- A thorough comparative evaluation of different imputation techniques to manage the problem of missing data in the context of outcome prediction for COVID-19 patients.

This retrospective study was approved by the ethics committee of the hospital where data were collected, which also waived the requirement for informed patient consent because of its retrospective nature.

## II. RELATED WORK

In this section we overview related works concerning: missing data imputation methods (Section II-A) underlying the algorithms we have studied and compared in Section IV, deep learning models for diagnosis of COVID-19 from lung CT or chest CXR images (Section II-B), which are related to the deep model we use to compute automatic COVID-19 severity scores from CXR images (Section III-B), and risk-prediction methodologies (Section II-C) linked to the proposed risk-prediction models (Section V).

### A. MISSING DATA IMPUTATION

Medical/clinical research is often performed on datasets with a limited number of samples, some of which are described by vectors containing missing values, and where the missing data can be described by one of the following mechanisms [18], [20], [34], [35]:
- Missing-Completely-At-Random (MCAR), meaning that the event of a value being missing is independent from both observed and missing values, and occurs totally at random;
- Missing-At-Random (MAR), occurring when the probability of missing values only depends on observed data, i.e., the latter define groups within which the probability of being missing is constant;
- Missing-Not-At-Random (MNAR), taking place when data are not MCAR or MAR, and missingness depends on unobserved data. In other words, there is a well-defined (even though often unknown) cause for missing values. In the case of MNAR, having a missing data in one variable often has some relationship with the observations of other variables. For example, values for variable x1 may be missing/observed when variable x2 has high/low values. Alternatively, values in variable x1 may be missing when values in variable x2 are also missing [34].

In any case, due to the limited number of samples, removal of points with missing data is not a good option. Instead, data imputation algorithms are generally applied, which may be grouped into three categories: methods employing statistical models to essentially estimate the underlying data distribution, methods based on machine learning techniques, and methods based on hybrid combinations of the previous approaches.

Statistical methods replace missing data by estimating their underlying distribution and/or the whole data distribution. The imputed values are drawn from the estimated distribution when a random error may be added to simulate real distributions. Examples of such methods are Hidden Markov Models [36], linear regression models [37], [38], KNN-imputation [39], cold and hot-deck imputation [40], SVD-based imputation [41], or methods that explicitly estimate the underlying distribution by using, for instance, Gaussian mixture models [42]–[44]. These methods are well suited for MAR or MNAR data because they are tied to the estimation of a distribution. However, they are often based on critical parameters having a high impact on the computed values, and setting and evaluating these parameters can be quite difficult because the ground truth (the missing values themselves) is not known.

Machine-learning methods are more recent. They perform imputation by learning the data distribution from the complete samples. For example, Random Forests [19] are particularly appealing because they deal with heterogeneous data whose features can have different data types, do not need any data normalization, and produce explainable values. Other, more complex techniques, are based on neural networks. Among them, several proposals leverage auto-encoder networks [45]–[49] or encoder-decoder Convolutional Neural Networks (CNNs) [50], [51] in order to reconstruct the training samples in their decoding output. Once trained, such decoding networks are able to reconstruct the missing values in test samples.

A completely different imputation approach is used by Generative Adversarial Neural Networks (GANN) [52], which learn to generate "missing" data with the same distribution as the training set. This is done by training a "generative" network, which generates possible imputed values and proposes them to a "discriminative" network, which is trained to accept only those generated values that properly fill the missing ones according to the underlying data distribution. Neural networks may be better able to model MCAR, MAR, and MNAR data because of their inherent non-linearity, but their main disadvantage is the need of a large training set, which is often not available in case of (bio-)medical data. Moreover, neural networks are "black box" models whose predictions are difficult to explain.

Hybrid approaches have been proposed to exploit and merge the advantages of different methods. They are essentially based on the multiple imputation approach initially presented in [53], [54] (see Section IV). Multiple imputation (MI) methods, e.g., MICE [23], [35] (see Section IV),

essentially produce several estimates of the missing data by techniques containing some randomness. Then, two possible approaches are used to obtain the final result: (i) the first approach processes each imputed set in the same way and then combines the computed results through statistical methods [53], [55] (see Section VI-A2); (ii) the second one combines the computed imputations through classical techniques, such as the mean of the imputed values [56] (which may not be appropriate [35]) or by exploiting machine learning methods [57] (which require a lot of training samples).

Although MI techniques are able to produce effective results, their main parameter, i.e., the number $m$ of imputations to be generated and then combined, must be carefully chosen in order to reach a low and stable between-imputation variance (see Section IV-B1). Also, the kind of data missingness (MCAR, MAR, or MNAR) that hybrid techniques are best suited for depends on the merged imputation methods.

### B. AUTOMATED COVID-19 DIAGNOSIS FROM LUNG IMAGES

Since the beginning of 2020, several deep neural models have proven their effectiveness in the diagnosis of COVID-19 infection from either lung CT or CXR images [58]. Although the proposed deep neural networks were developed upon completely different architectures, and exploit different training losses and optimization algorithms, their common trait is the ''Active, Incremental Learning'' approach used for learning [59], [60], which is especially needed when the available datasets are limited in size and only small numbers of new cases can be acquired incrementally.

Thanks to the existence of large open datasets containing either lung CT [61] or CXR [62] images from patients with various diseases other than COVID-19 (e.g., lung cancer, pneumonia, pleural effusion, and others), the problem of COVID-19 diagnosis is commonly addressed by training well-known existing deep neural networks [62]–[64], such as ResNet [65], [66], Inception-Net [67], [68], or VGG [69], [70], on the existing, large datasets. In this way, the network is first trained on a similar task, such as lung cancer or pneumonia diagnosis. Next, the knowledge of the pretrained network is ''incremented'' by applying a training phase where an augmented COVID-dataset is used [12], [71].

Importantly, considering that deep models have been highly criticized in the past for their ''black-box'' explanations, several deep models proposed for COVID-19 diagnosis [12], [14] include a further interpretation step, applied to motivate the computed prediction. Among the various state-of-the-art methods for interpreting the predictions of deep models [72], the mostly used are sensitivity analysis [12], [73], [74], which allows the areas of highest activation to be identified, e.g., in the first hidden layer (since this layer is often considered as the one where base textural and color/gray level features are learned). Another common approach is output back-propagation as used by algorithms such as GRAD-CAM [13], [75] or layer-wise relevance propagation [14], [76], which essentially back-propagate the activation in the output layer to understand which areas are the most relevant in the computation of the final decision.

Deep models for lung CTs and models for CXRs differ in the dimensionality of the input images (CTs are 3D images while CXRs are 2D images), meaning that deep models with 3D convolutional layers are used for processing CTs, whereas models with 2D convolutional layers are used for CXRs. On the other hand, all the methods apply a transfer learning technique and most of them start by a ResNet or an inception-Net.

The work proposed in [13] represents an exception to the above considerations, since the authors eschew the 3D processing generally applied for CT images in favor of the classical 2D processing applied for 2D (CXR) images. The 2D ResNet50 architecture process each 2D slice of the CT and the output of all the ResNet are subsequently used as input to a max pooling layer followed by a dense layer, which computes the final prediction. Another interesting example of a deep learning model for CTs applies transfer learning to ResNet architectures and creates an augmented dataset by applying the usual image transformations to both the original image and the images obtained by wavelet decomposition. More precisely, instead of augmenting the dataset by transforming only the original image, wavelet decomposition is applied and also the images obtained from wavelet decomposition up to 3 levels are added to the training set [12].

In general, deep learning models for CT data obtain higher performance than those trained on CXR data, which presumably reflects the higher sensitivity of CT for diagnosing abnormalities related to COVID-19 as compared to CXR.

Despite this initial enthusiasm for machine learning based on lung CT data, their longer acquisition time and higher costs (when compared to chest CXRs) mean that lung CT are impracticable for the early screening of patients with suspected COVID-19 in EDs, even though CT may be the preferred modality for predicting the disease progression in COVID-19 patients. To this end, a recent study presented a severity score index computed by humans from chest CT, and used it together with other inflammatory indexes and age to form a patient's feature vector input to logistic regression classifiers [77].

A recent approach to feature selection in COVID-19 CXR data used the first convolutional layers of existing networks (e.g. AlexNet, VGGs, GoogleNet, ResNets, InceptionNets, DenseNet) as extractors of ''Deep Features''. The convolutional layers were connected to a dense fully connected layer that transforms the output of the convolutional layers into a 1000 dimensional vector, whose weights are tuned via transfer learning. The 1000-dimensional outputs are then used as input to support vector machines (SVMs). The results showed that a ResNet architecture followed by SVM achieves the best performance [78].

Based on the notion that the residual layers of ResNet are the key for its success, in [11], [14] authors presented CovidNet, a tailored CNN model using residual connections, which is trained to reproduce the scores of lung involvement

(extent and severity, cf. Section III-B) produced by human experts. Given the successful results obtained by such network, we used it to produce two radiological features, which have been added to our dataset.

### C. RISK PREDICTION MODELS FOR COVID-19 PATIENTS

A recent exhaustive survey of the literature on multivariate models and scoring systems for predicting COVID-19 related outcomes revealed 107 studies describing 145 prediction models. Of these, four models aim to identify people at risk in the general population; 60 exploit medical imaging for diagnosing COVID-19 in patients with suspected infection; nine models diagnose disease severity; and 50 propose prognostic models for predicting mortality risk, progression to severe disease, intensive care unit admission, ventilation, intubation, or length of hospital stay.

Besides being a precise report of all the available state-of-the-art works (up to May 5$^{th}$, 2020) for COVID-19-related predictions based on patient data, the method proposed in [15] is very interesting since it highlights all the biases mentioned in Section I and that affect several of the published methods. However, the work in [15] does not describe the different machine learning or statistical approaches used for prediction.

In this work, we sought to update the survey of COVID-19 methods with papers published up to October 7$^{th}$, 2020. We considered all prognosis prediction models for COVID-19 patients, in order to identify their main processing steps. First, we noted that most of the proposed approaches avoid, or do not even mention, any pre-processing phase for data normalization/standardization, missing value imputation, or feature selection, which would surely increase robustness and improve performances. Moreover, while some works only report descriptive statistics obtained by univariate [79] or multivariate [80] analysis, the majority of the approaches exploit logistic regression classifiers [81]–[97]. The remaining methods use RF classifiers [85], [87], [91], [92], [97]–[101] or XGBoost [102], [103], SVMs [87], [91], [97], [100], [101], K-Nearest Neighbor classifiers [87], [91], [100], Cox regression models [104], [105], or artificial neural networks [106].

Unfortunately, except for an approach that was developed and tested based on a private dataset with 929 COVID-19 patients [107], all the published methods were developed with datasets with relatively small sample sizes. This hinders the usage of sophisticated learning models, such as neural networks, which could uncover highly nonlinear relationships. Moreover, since all the datasets are private, an objective comparison between different methods is impossible.

## III. COVID-19 DATASET

In this section we describe our patient dataset and provide a description of the radiological feature computation used by our method.

### A. PATIENT DATASET

This study was performed on clinical, comorbidity, laboratory, and antero-posterior (A-P) or posterior-anterior (P-A) CXR data from patients referred to the ED of an urban multicenter health system, from March, 7, 2020, to April, 10, 2020. All patients in our cohort were RT-PCR positive for COVID-19.

Our inclusion criteria stipulated the availability of five-months of clinical follow-up data, to allow a truthful and reliable risk classification. Additionally, patients were included only if a CXR was performed and evaluated by two experienced radiologists before the availability of the nasopharyngeal swab result.

The five-month follow-up allowed us to accurately classify low-risk patients, who were either not hospitalized or, despite hospitalization, were never intubated and survived with no serious consequences, and patients at high risk, that is patients that either were intubated, experienced serious consequences, or died.

With this setting, the patient set included, 207 and 94 adult men and women with a mean age of $61 \pm 1$ years [min $= 23$, max $= 95$], and with a number of days with symptoms from COVID-19 that were on average $7 \pm 0$ [min $= 1$, max $= 30$]. Among them 214 patients were at low risk, while 87 patients were at high risk.

The data included symptoms (e.g., fever, cough, dyspnea, etc.), clinical history and comorbidities (such as arterial hypertension, chronic obstructive pulmonary disease, cancer, asthma, etc.), laboratory measurements (e.g., LDH, white blood cell count, lymphocyte), saturation/oxygen values, and patient data (age, sex).

Although effective data imputation techniques were applied to fill missing values (see Section IV), two laboratory variables lacking more than 50% of observations (LDH, AST) were removed. Moreover, to avoid singularity, variables having a variance below 0.025 were removed (precisely, logical variables recording the presence/absence of two symptoms, ageusia/anosmia and thoracic pain, and three variables recording comorbidities, that is pulmonary interstitial disease, hepatopathy, and dementia).

The resulting dataset (summarized in Table 1) is composed of 41 variables, whose values were recorded during patients visits at the ED:

- twelve are logical variables representing the presence/absence of a symptom,
- nine are logical variables describing the presence/absence of a comorbidity,
- patient sex is represented with a logical variable (true for men and false for women),
- four integer variables report: patient age, the number of comorbidities, the number of symptoms, and the number of symptomatic days before presentation to the ED,
- two real-valued and two integer-valued variables encode radiological features,
- two integer variables record saturation values,

**TABLE 1.** Variables in the patient dataset.

| Variable name | All sample | Moderate risk | Severe risk | ≈ p-value (chi squared test) |
|---|---|---|---|---|
| **Boolean variables** | | | | |
| Sex | 207 men<br>94 women | 145 men<br>69 women | 62 men<br>25 women | 0.5702 |
| **Symptoms** | | **% presence (no.)** | | |
| Fever | 93 (280) | 92.5 (198) | 94.3 (82) | 0.6382 |
| Cough | 66.8 (201) | 68.7 (147) | 62.1 (54) | 0.2804 |
| Dyspnea | 55.1 (166) | 47.7 (102) | 73.6 (64) | 5E-04 |
| Respiratory Failure (IR) | 13 (39) | 8.9 (19) | 23 (20) | 0.002 |
| Myalgias | 9.3 (28) | 9.3 (20) | 9.2 (8) | 1 |
| Other | 9.6 (29) | 9.8 (21) | 9.2 (8) | 1 |
| Syncope | 4.3 (13) | 5.1 (11) | 2.3 (2) | 0.3598 |
| Asthenia | 12.3 (37) | 11.7 (25) | 13.8 (12) | 0.6932 |
| Vomiting.Nausea | 5 (15) | 4.2 (9) | 6.9 (6) | 0.3963 |
| Diarrhea | 10.3 (31) | 10.7 (23) | 9.2 (8) | 0.8356 |
| Headache | 3 (9) | 3.3 (7) | 2.3 (2) | 0.7346 |
| Pharyngeal.pain | 3 (9) | 3.7 (8) | 1.1 (1) | 0.2894 |
| **Comorbidities** | | **% presence (no.)** | | |
| Pneumo.asthma | 4.7 (14) | 5.1 (11) | 3.4 (3) | 0.5852 |
| Pneumo.BPCO | 5.3 (16) | 4.2 (9) | 8 (7) | 0.2659 |
| Neoplasia (last 5 years) | 10.6 (32) | 7.9 (17) | 17.2 (15) | 0.0235 |
| Smoke | 5.3 (16) | 5.6 (12) | 4.6 (4) | 0.7836 |
| Arterial.hypertension | 29.9 (90) | 26.2 (56) | 39.1 (34) | 0.039 |
| Cardiovascular pathologies | 16.6 (50) | 11.7 (25) | 28.7 (25) | 0.001 |
| Diabetes | 15.9 (48) | 12.1 (26) | 25.3 (22) | 0.0045 |
| Obesity | 6 (18) | 5.6 (12) | 6.9 (6) | 0.7976 |
| Celebral Stroke | 4 (12) | 3.7 (8) | 4.6 (4) | 0.7536 |
| **Integer- and real-valued variables** | | | | |
| | **All sample** | **Moderate risk** | **Severe risk** | **≈ p-value (one-sided Wilcoxon signed-rank test)** |
| **Counts** | | **mean ± s.e. [range]** | | |
| No.Symptoms | 3 ± 0.09 [0, 7] | 3 ± 0.1 [0, 7] | 3 ± 0.17 [0, 6] | class 0 < class 1, 0.0224 |
| No.Comorbidities | 1 ± 0.09 [0, 6] | 1 ± 0.1 [0, 5] | 1 ± 0.19 [0, 6] | class 0 < class 1, 4E-04 |
| Symptoms.No.days | 7 ± 0.29 [1, 30] | 7 ± 0.36 [1, 30] | 7 ± 0.51 [1, 20] | class 0<class 1, 0.4752 |
| Age | 62 ± 1.15 [23, 95] | 58 ± 1.34[23, 92] | 67 ± 1.88 [23, 95] | class 0 < class 1, 0 |
| **Radiological variables** | | **mean ± s.e. [range]** | | |
| usa.radio.score.MAX[1] | 3 ± 0.14 [0, 6] | 3 ± 0.16 [0, 6] | 4 ± 0.24 [0, 6] | class 0 < class 1, 1E-04 |
| radio.SCORE [1] | 9 ± 0.26 [0, 18] | 8 ± 0.31 [0, 16] | 10 ± 0.46 [0, 18] | class 0 < class 1, 0 |
| GEO.extent.score[2] | 4.22 ± 0.07 [1.45, 6.30] | 4.01 ± 0.08 [1.45, 6.22] | 4.74 ± 0.1 [1.73, 6.30] | class 0 < class 1, 0 |
| OPC.extent.score[2] | 3.1 ± 0.06 [1.17, 4.85] | 2.92 ± 0.07 [1.17, 4.85] | 3.55 ± 0.1 [1.24, 4.85] | class 0 < class 1, 0 |
| **Saturation values** | | **mean ± s.e. [range]** | | |
| PaO2.PF | 310 ± 9.18 [40, 733] | 333 ± 9.57 [61, 733] | 231 ± 16.59 [40, 567] | class 0 < class 1, 0 |
| SpO2.in.FA | 93 ± 0.46 [65, 100] | 95 ± 0.37 [82, 100] | 88 ± 1.07 [65, 98] | class 0 < class 1, 0 |
| **Laboratory variables** | | **mean ± s.e. [range]** | | |
| ALT | 35 ± 3.51 [4, 486] | 34 ± 3.69 [4, 378] | 42.5 ± 7.96 [9, 486] | class 0 < class 1, 0.0448 |
| Platelets | 199 ± 6.6 [7, 792] | 196.5 ± 8 [7, 792] | 205 ± 11.69 [34, 513] | class 0 < class 1, 0.4784 |
| White.blood.cells | 8.45 ± 0.71 [1.65, 179.67] | 7.54 ± 0.54 [2.3, 109.77] | 10.66 ± 2.05 [1.65, 179.67] | class 0 < class 1, 0.0041 |
| Red.blood.cells | 4.64 ± 0.04 [2.56, 7.65] | 4.68 ± 0.04[2.56, 7.65] | 4.53 ± 0.07 [2.86, 6.43] | class 0 > class 1, 0.0266 |
| Lymphocyte | 2.49 ± 0.76 [0.11, 172.48] | 2.22 ± 0.65 [0.25, 98] | 3.12 ± 2.04 [0.11, 172.48] | class 0 > class 1, 1E-04 |
| perc.Lymphocyte | 17.94 ± 0.72 [0.6, 96] | 19.7 ± 0.84 [3.3, 85.4] | 13.72 ± 1.29 [0.6, 96] | class 0 > class 1, 0 |
| CRP[3] | 8.92 ± 0.46 [0.05, 34.7] | 7.01 ± 0.46 [0.05, 27.85] | 13.66 ± 0.96 [0.77, 34.7] | class 0 < class 1, 0 |
| Haemoglobin | 13.49±0.11 [7.16, 19.1] | 13.63 ± 0.12 [7.16, 19.1] | 13.14 ± 0.21 [8.6, 17.7] | class 0 > class 1, 0.0124 |
| Haematocrit | 38.88 ± 0.29 [21, 64] | 39.18 ± 0.34 [21, 64] | 38.12 ± 0.54 [25.3, 51.1] | class 0 > class 1, 0.0336 |

[1]Radiological value automatically computed by human, [2] Radiological value automatically computed by CovidNet, [3] CRP = C-Reactive Protein.

- nine real values variables describe laboratory (blood) test results.

Boolean variables (symptoms, comorbidities, and sex) are described through the percentage of true values in all the patient dataset (column ''All samples'' in table 1), in the subset of patients at moderate risk (column ''Moderate Risk'' in table 1), and in the subset of patients at severe risk (column ''Severe Risk'' in Table 1). Integer and real valued variables are represented through their mean $\pm$ standard error (s.e.) of the mean and their range ([minim value, maximum value]) in the entire dataset (column ''All samples''), the subset of patients at moderate risk (column ''Moderate risk''), and the subset of patients at severe risk (column ''Severe Risk'').

To provide a first hint of the class separation provided by each variable, we performed statistical analysis to check whether there are statistically significant differences in the patients distributions. Precisely, for boolean variables we applied the chi-squared test to determine if statistically significant differences were present between patients at low or at high risk. Numerical variables were analyzed to detect statistically significant distribution differences by applying the one-sided Wilcoxon signed-rank test.

### B. CHEST X-RAY ANALYSIS AND AUTOMATED PROCESSING

The Fleischner Society presented three different scenarios and an algorithm for recommending chest imaging that includes CT and/or CXR to direct patient management during the COVID-19 pandemic. Ultimately, the choice of imaging modality is left to the judgement of clinical teams at the point of care, accounting for the differing attributes of CXR and CT, local resources, and expertise [2]. Though CXR shows clear patterns, distinguishable from those of pneumonia [108], when COVID-19 infection becomes serious, it is insensitive in mild or early infection stages [108]. In contrast, lung CT has greater sensitivity for early pneumonic changes, but this advantage is partially diminished by the huge burden placed on radiology departments in terms of staff commitment, CT room workflow, and disinfection procedures [2], [109]. Therefore, many Italian hospitals decided to employ CXR as a first-line triage tool [108]–[111].

Several recent studies on the utility of initial CXR for predicting clinical outcome correlated the presence and the extension of opacities on initial CXR with the need for hospitalization and/or intubation [17], [108], [110], [111].

In light of these considerations, in our study we included four radiological variables expressing the extent and severity of the COVID-19 pattern, visible from the CXR acquired at the time of presentation to the ED.

Two of the four radiological variables, *radio.score* and *usa.radio.score*, were evaluated by expert radiologists, which were blind to the patients' condition; the other two radiological variables, *GEO.extent.score* and *OPC.extent.score*, were computed by a deep neural network trained on a radiological score evaluated by clinical experts [11], [14].

*Radio.score* and *usa.radio.score* were defined by two thoracic radiologists with 23 and 20 years of experience in thoracic imaging, after re-evaluation of the initial CXR that the patients underwent during the admission at the ED. The *radio.score* index was used to assess the severity of pulmonary involvement from both the 156 antero-posterior (A-P) and the 143 postero-anterior (P-A) images. The score is calculated by dividing each lung into three areas (upper, middle, and lower); each area is then scored with an involvement value in the range $\{0, \ldots, 4\}$, where 0 means that no anomaly has been found, while higher scores mean increased presence of severe COVID-19 CXR patterns: 1 = reticular interstitial thickening, 2 = reticular interstitial thickening and ground glass, 3 = ground glass opacities and consolidation with ground glass as the most widespread anomaly, 4 = consolidation as the most widespread anomaly. Summing up the scores assigned to each of the six areas, each lung gets a score in the range $\{0, \ldots, 24\}$. Lin's concordance correlation coefficient [112] between the scores of the two radiologists ($c_{\text{Lin}} = 0.76$, c.i. $= [0.65, 0.76]$, p-value $<$ 1E-58) showed a substantial agreement. Therefore, we averaged the two scores to get a single value.

By binarizing the scores of each lung area, that is by assigning a value of 1 to each area showing at least ground glass opacities and consolidations (area scores greater or equal to 2), an summing up all the binary values, we obtained a simplified version of *radio.score*, falling in the range $\{0, \ldots, 6\}$ and referred to as *usa.radio.score*. Since in this case Lin's correlation coefficient showed a low agreement ($c_{\text{Lin}} = 0.40$, c.i. $= [0.30, 0.49]$, p-value $<$ 1E-11), a pooled score was obtained by taking the maximum value for each patient. This is a conservative way of pooling the results, based on the assumption that a false positive error is less costly than a false negative error. In other words, diagnosing a mild case as severe is better than wrongly considering a severe case to be mild.

To assess the reliability of the scoring system computed through a deep network, we used the state-of-the-art CovidNet deep neural network [11], [14]. Precisely, we automatically preprocessed the CXR images of each patient to first remove positional artifacts, such as rotations and variations in zooming. Subsequently, gray levels were normalized through ACE [113], a spatial color equalization algorithm [114] that has been often used to remove unwanted and adverse illumination conditions [115] and that recently gained importance in the field of medical image processing [116], thanks to its ability to reveal small details without introducing noise and artifacts. The preprocessed images of each patient were used as input to CovidNet in order to get a geographic extent score (*GEO.extent.score*) and an opacity extent score (*OPC.extent.score*).

CovidNet is a deep neural network that was originally developed for recognition of COVID-19 patients [14]. CovidNet was subsequently extended by transfer learning on an augmented dataset composed of only 130 CXRs from Chinese patients [14], which were scored by two

experienced radiologists by adapting the scoring system proposed in [116]. Such scoring method quantifies both the extent of lung involvement by ground glass opacity or consolidation, through the geographic extent score (*GEO.extent.score*), and the kind of COVID-19 patterns seen in the radiographs, through the opacity extent score (*OPC.extent.score*). Both scores are computed separately on each lung, and a final value is then obtained as sum of the left- and right-lung scores. More specifically,

- *GEO.extent.score* takes the following scores on each lung: 0 = no lung involvement; 1 =< 25% of lung involvement; 2 = 25-50%; 3 = 50-75%; 4 => 75% of lung involvement; thus, the final score ranges from 0 to 8.
- *OPC.extent.score* ranges from 0 to 6, and it quantifies the degree of opacity in each lung by using the following values: 0 = no opacity; 1 = ground glass opacity; 2 = consolidation; 3 = white-out [11].

## IV. APPROACHES TO MISSING DATA

The available dataset contains both logical, integer-valued, and real-valued attributes. Both the discrete and continuous variables are affected by missing data; thus, it is appropriate to consider an imputation phase.

### A. UNCOVERING THE MISSING DATA MECHANISM

Since the validity of any imputation method depends on the missing data mechanism, care must be taken to understand whether the involved data are MCAR, MNAR, or MAR [18], [20], [34], [35].

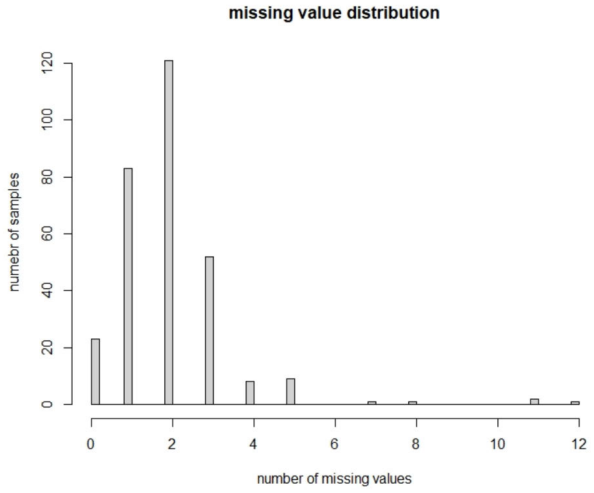


**FIGURE 1. Histogram of missing values for each sample: the maximum number of missing values is 12, corresponding to 25% of the variables. Only one sample has 12 missing values.**

Precisely, to confirm that data is, or is not, MNAR, the missing data pattern is generally observed through visualizations (see Fig. 1 and Fig. 2). We searched for some MNAR pattern by expressing each attribute as a binary variable, whose observations are set to 1 (missing) or 0 (observed). Using this binary representation, we applied the following analysis, which provided no evidence of MNAR data. For each pair of attributes, we found no high correlation between the corresponding binary representations (Pearson correlation coefficient <0.75, with a significant p-value), or we confirmed the independence of their missing/observed data proportions, using the chi-squared test (with Yates's correction). Further, for each variable with a sufficient number of missing values (we set this value to be 25), $x_1$, and each other numeric variable, $x_2$, we used the Wilcoxon signed-rank test to confirm that the difference between the distributions of $x_2$ values for missing and observed values of $x_1$ was not statistically significant.

Next, to determine whether our data are MAR or MCAR, we used the non-parametric test of Jamshidian and Jalal [18], [20], an extension of Little's test [117] that is suited both in case of a high and a low proportion of missing values. Precisely, if homogeneity of covariances (homoscedasticity) between subsets of data having identical missing data patterns is detected, data are supposed to be MCAR. The novelty of the approach relies on the fact that authors test for homoscedasticity using a modification of the statistic proposed by Hawkins [118]. This statistic has the peculiarity of working well for testing homoscedasticity in *complete* data when group sizes are small. In order to process a complete dataset, in case of unknown data distribution authors perform imputation by a method, *distFree*, that only assumes independence of the observations, and the continuity of their cumulative distribution function; no further specific distributional assumptions are required. *distFree* is similar to the imputation technique proposed in [119], which exploits maximum likelihood estimators to compute a linear predictor of the missing observations, and then adds a random error to obtain the final imputations. Although this method implicitly assumes that the variables are linearly related, authors argue that the maximum likelihood technique may indeed provide consistent estimators [120]. In sum, using Jamshidian's and Jalal's test we determined that our data are MCAR.

### B. MISSING DATA IMPUTATION

At the state of the art, several imputation models for MCAR methods have been presented that can deal with "complex" data [45]. Among such methods, we experimented both Multiple Imputations by Chained Equations (MICE [23], [121]), using either predictive mean matching (*micePMM*) or Random Forest classifiers (*miceRF*) as the base imputation model, and *missForest* [19], which also exploits RFs.

More precisely, MI techniques [22] are an effective strategy that exploit randomness for producing unbiased estimates, with a reduced dependency on the normality assumption [22]. MIs are mainly used for estimating the linear or logistic regression coefficients that link predictor variables to a response variable. In this case, given a dataset (with MCAR or MAR values) and an imputation model containing some randomness, $m$ imputed datasets are drawn,

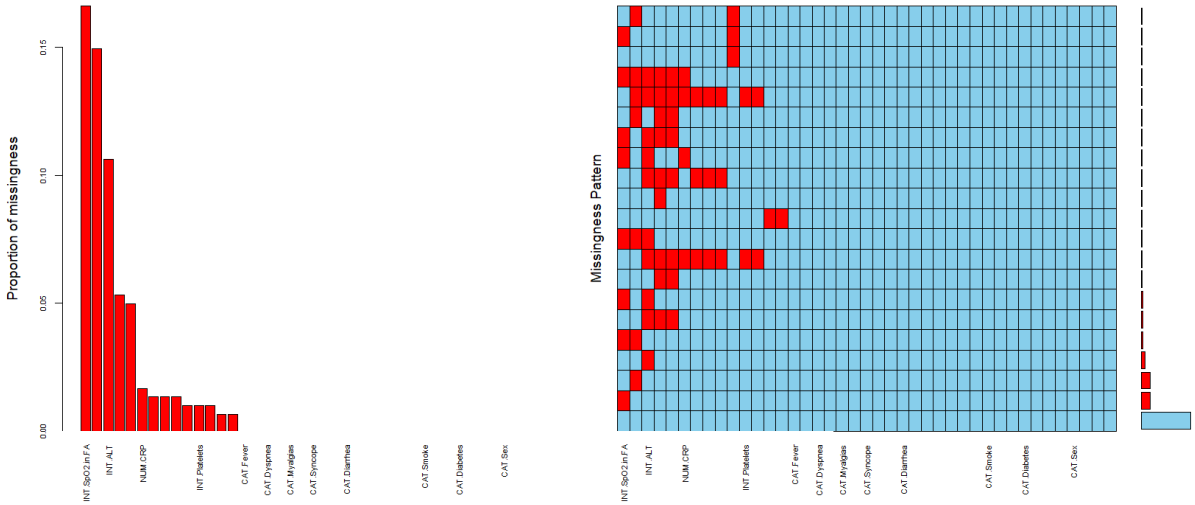


**FIGURE 2.** Missing data patterns. (left) Proportion of missing values for all variables in the dataset, sorted by decreasing order. (right) Combinations of missing values: red squares in a matrix entry denote the presence of missing values for the variable associated to the column in the samples corresponding to the row; the bars on the right show the cardinality of each set of points.

and subsequently processed separately but identically by the chosen estimator. The resulting coefficients are expressed through their mean and (global) variance, computed according to Rubin's Rule [53], [55] (see Section VI-A2), which in turn allows the Wald test to be applied for checking their significance [122]. Note that, although some authors [123], [124] suggest that setting $m = 5$ MIs is enough to produce unbiased estimates, other contributions [35], [125] show that $m > 20$ should be used to obtain reliable estimates for the global variances, so that the simulation error is almost cancelled (in Section IV-B1 we experimentally determine a value for $m$ minimizing the variance).

MICE (aka Fully Conditional Specification, or FCS) is a MI technique that uses a set of conditional densities for each variable with missing data to build a multivariate imputation model on a variable-by-variable basis. Initially, all missing values are replaced by simple random sampling with replacement from the observed values. Subsequently, when using predictive mean matching (PMM [126]) as the base imputation model, the following steps are applied:

- starting from the first variable, $x_1$, a regression model is fit to the observed x1 by using the remaining variables as the independent predictors;
- randomness is introduced by drawing a subset of regression coefficients from the posterior predictive distribution of the computed coefficients; the drawn coefficients are used to predict all (observed and missing) values for $x_1$;
- each missing value in $x_1$ is finally imputed by considering the predicted value of one among $k$ donors, randomly selected among observed elements in $x_1$ whose predicted values are close to the predicted value for the case with missing data.

This process is repeated by using all the variables as independent predictors. When all variables are imputed, a cycle is complete. To stabilize the process, the cycle is repeated $n$ times by using, at each iteration, the previously imputed values as initialization values (authors suggest setting $n$ in the range $\{10, \ldots, 20\}$ for obtaining unbiased results [126]). Note that the variable order used by the iterative univariate imputation may be defined according to different criteria based on missing value proportion, such as decreasing, increasing, or random sorting.

As highlighted in [35], PMM has the advantage of using an implicit data model, thus avoiding the explicit definition of the distribution of missing values, which often brings to model misspecification. Moreover, the values imputed by PMM are actually observed values, therefore avoiding the generation of out-of-range imputations. However, PMM-based MICE (*micePMM*) is a parametric approach that assumes that the observed data have a distribution similar to that of missing data [127]. To avoid any parametric approach, a novel model was presented (*miceRF*), where RFs substitute PMM. More precisely, for each variable, a bootstrap sample is used to impute missing values in the dependent variable by using RFs. The advantage lies here in the usage of a further internal bootstrap sampling, allowing each tree to be fit to a different data sample. Results aggregated by the RF are therefore supported by a source of randomness that is greater than that of PMM; moreover, RFs do not rely on any specific assumption regarding the distribution underlying missing data. Indeed, results shown in [127] suggest that both *miceRF* and the *missForest* algorithm produce more robust results than those computed by *micePMM*.

*missForest* [19] iteratively exploits the ability of RF classifiers to deal with mixed data types without making any assumption about the underlying data distribution. It follows an iterative approach similar to that applied by MICE, that is it iteratively imputes each variable with missing data using the remaining variables. After making an initial guess for the

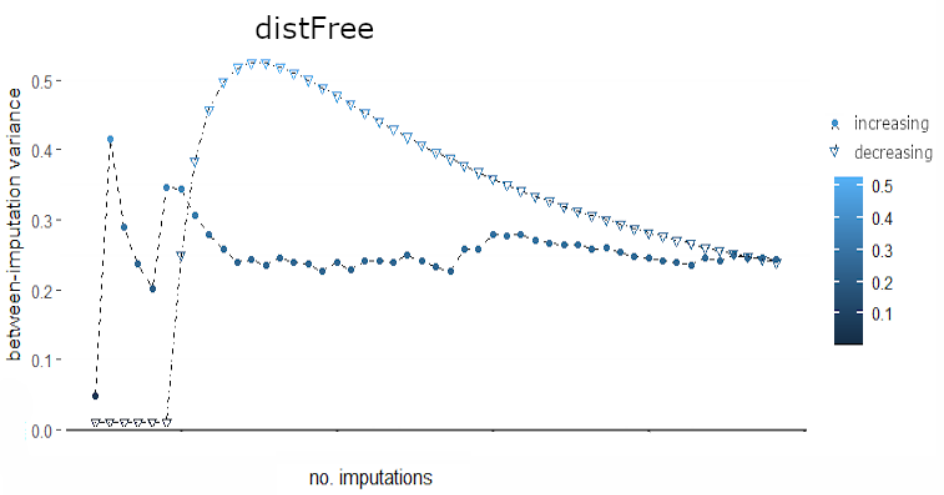


**FIGURE 3.** Between-imputation variances computed on 100 datasets imputed with *distFree*. Dots and triangles mark the variances computed using increasing and decreasing imputation order, respectively.

missing values, e.g., by using the mean of observed values, it considers in turn each variable $x$ with missing entries (by default, variables are considered by increasing missing value proportion, though other sorting criteria can be used). An RF is fit to the observed values of $x$ using the other variables as predictors, and subsequently used to impute the missing values. Such procedure is repeated until the difference between the newly imputed data matrix and the previous one increases for the first time with respect to both continuous and categorical variables (obviously using two separate difference metrics). *missForest* has the same appealing properties of *micePMM* and *miceRF*. Indeed, since RFs are trained on bootstrapped samples, MIs can be computed by using different bootstrap sets, which introduces randomness. Moreover, this method can deal with multivariate data consisting of continuous and categorical variables. Finally, *missForest* does not require assumptions about distributional aspects of the data, nor does it have critical hyper-parameters to be tuned. More precisely, it only requires the number of trees ($n_t$) to be specified; although this value is generally set to a high value, e.g. $n_t = 500$, we set $n_t = 100$ to avoid overfitting and reduce computational time.

The first aim of this work is to provide suggestions about the employment of imputation techniques using different baseline theories. Therefore, we experimented with *distFree*, *micePMM*, *miceRF*, and *missForest*, considering both univariate imputation orders defined by the increasing and decreasing missing values proportion (henceforth referred to as ''increasing imputation'' and ''decreasing imputation'', respectively). To avoid bias, all imputations were performed after discarding the point labels.

Note that, though *distFree* and *missForest* are not MI techniques, they both rely on a randomness source (*distFree* adds random noise to each imputation, while *missForest* trains RFs by using randomly bootstrapped samples) and may be therefore used to produce $m$ different imputations. In all imputation algorithms we set the maximum number of iterations to 11, since values in $\{10, \ldots, 20\}$ allow unbiased imputations to be obtained [126]. Finally, we limited *miceRF* and *missForest* univariate imputations by training RFs with a maximum of $n_t = 100$ trees.

### 1) CHOOSING THE PROPER IMPUTATION ALGORITHMS AND THE VALUE FOR $m$

To compare the imputation algorithms and the univariate imputation order we produce $m = 100$ different imputations and analyze the between-imputation variance. Given the original dataset $D$ with $S$ missing values, $x_{\text{miss}}(s)$ ($s = 1, \ldots, S$) and given an imputation method, imp, specified by an imputation algorithm and an univariate imputation order, let's denote with $D_{\text{imp}}(1), \ldots, D_{\text{imp}}(m)$ the $m$ imputations produced with imp.

To compute the between-imputation variance, we found the normalization coefficients that allow the observed values in each column to be mapped to the range $[0, 1]$ (they depend on the minimum and maximum of the observed values in each column of $D$) and used them to normalize each imputed set, therefore obtaining $D^*_{\text{imp}}(i)$ ($i = 1, \ldots, m$). Next, we computed the between-imputation variance of each missing value $x_{\text{miss}}(n)$ in $D$, $\text{Var}(x_{\text{miss}}(s))$, $s = 1, \ldots, S$, by using its $m$ imputed values in $D_{\text{imp}}(1), \ldots, D_{\text{imp}}(m)$. The global between-imputation variance was finally computed as the mean of the $\text{Var}(x_{\text{miss}}(s))$, $s = 1, \ldots, S$.

Figs. 3–5 show the global between-imputation variances achieved by the four methods (using the increasing (dots) and decreasing (triangles) order of missing values) for $m \in \{2, \ldots, 100\}$. In Table 3 (Appendix A) the ranges of such between-imputation variances are reported.

For both univariate imputation orders, *distFree* achieves the highest between-imputation variances with a mean

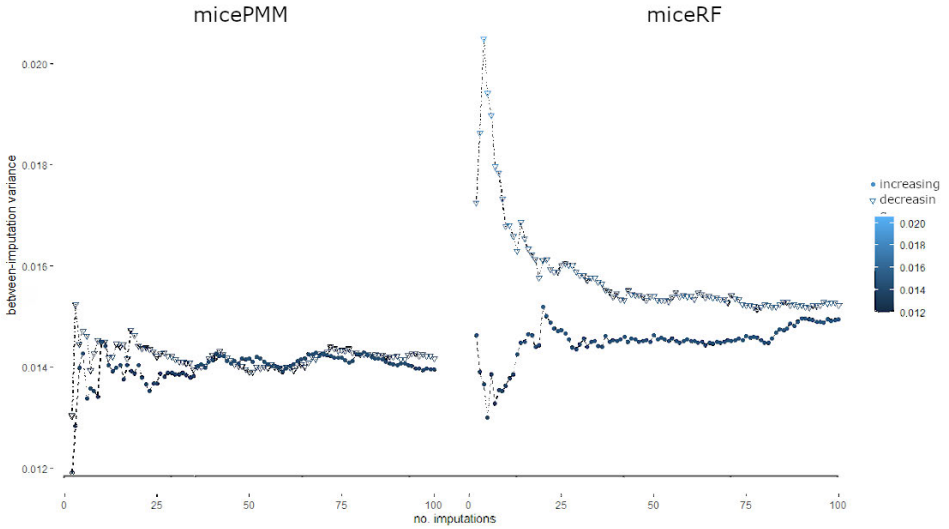


**FIGURE 4.** Between-imputation variances computed on the 100 datasets imputed with *micePMM* (left) and *miceRF* (right), using the same scale for Y axis. Same notations as in Fig. 3.

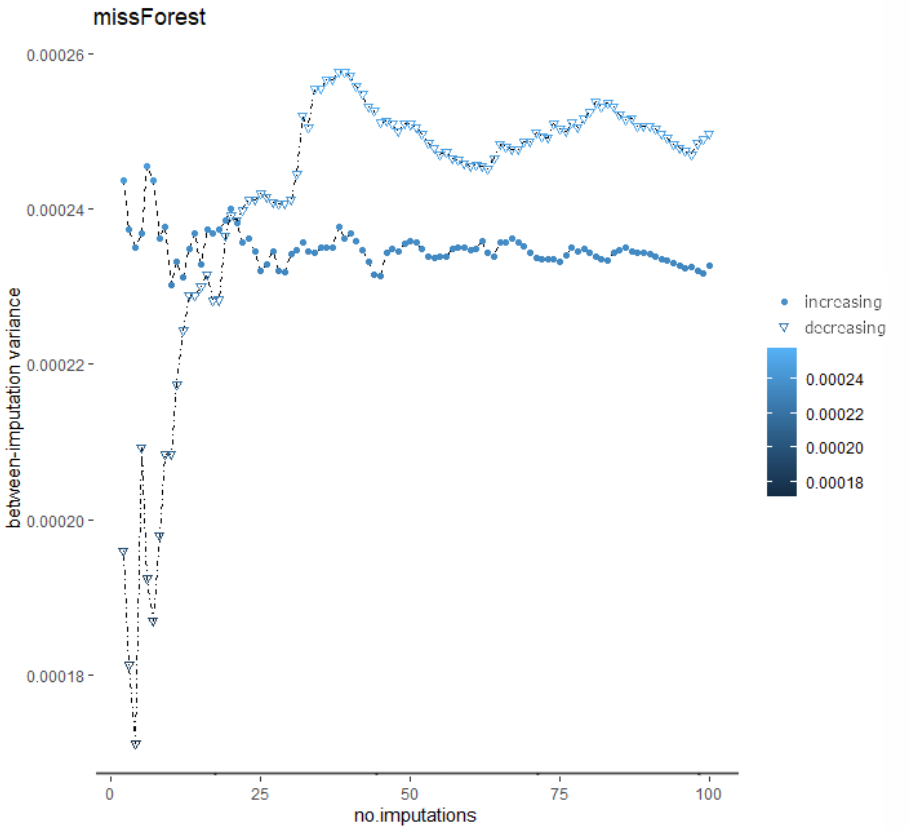


**FIGURE 5.** Between-imputation variances computed on the 100 datasets imputed with *missForest*. The obtained values are negligible, as highlighted by the span of Y axis: this practically means that the imputed values are always similar. Same notations as in Fig. 3.

slightly lower than 0.3 (Fig. 3; cf. also Table 3 in Appendix A); this variance is very high, considering that data are normalized. Moreover, the between-variance ranges of *distFree* are respectively $10^3$ and $10^2$ times bigger than those of *missForest* and of multiple imputations exploiting MICE. The high between-imputation variance computed when using *distFree* practically means that, for each missing value, its imputed values are very noisy. On the contrary, *missForest* has negligible between-imputation variance, meaning that the predicted values are stable.

Each imputation method is characterized by a between-imputation variance whose order of magnitude is independent of the univariate imputation order. However, *distFree* has completely different behaviors when the two orders are used, further suggesting that its results may not be considered as sufficiently robust. In Fig. 4 we show the zoomed between-imputation variances achieved by *micePMM* and *miceRF*, while those computed when using *missForest* are plotted in Fig. 5. When using such algorithms, the between-imputation variances reach a sort of *plateau* after an initial variability involving around 30 imputations. Although the plot based on *missForest* suggests a higher variation of the between-imputation variance, all the values are near to zero. In sum, when using the increasing univariate imputation order, both *micePMM* and *missForest* obtain the lowest between-imputation variance.

Note also that the negligible between-imputation variances produced by *missForest* highlight the fact that the different imputations it computes are very similar. For this reason, this method should not be used to impute missing data when there is the need to test the robustness of subsequent processing steps w.r.t. data variability. On the other hand, *missForest* should be used when the goal is to obtain (almost) reproducible results.

When the underlying data distribution is unknown, we therefore suggest performing imputation with either *missForest*, when negligible between-imputation variances are needed, or *miceRF*, because it combines the advantages provided by working on multiple imputations and therefore allows the robustness of subsequent algorithms to be tested by considering some randomness in the data. Obviously, when the normality assumption holds *micePMM* is also a viable option. Moreover, to achieve stable between-imputation variances, in our problem we suggest using $m > 20$ imputed sets as advised in [35], [125].

In the problem under study we cannot make assumptions about the underlying data distribution; therefore, though *micePMM* achieves low and stable variances, its use would not guarantee a proper imputation of missing values. Anyhow, *miceRF* has similar variances and therefore we can use it to assess MICE-based techniques, comparing it to *missForest*, which obtained the lowest between-imputation variance. With both methods we choose to use the increasing univariate imputation order, which produces more stable results, generating 50 imputed sets. More precisely, after imputing missing data by using these methods, we trained and tested RFs, ATs, and GLMs (see Section V) in order to obtain predicted risk levels, as well as the relevance of each variable in the prediction. For each method, the predictions and relevance computed on the $m$ imputations were pooled by applying Rubin's rule (see Section VI-A2).

## V. RISK PREDICTION APPROACHES

Once missing data have been imputed, we apply two different risk prediction approaches, both described in this section.

Given a training set, the first approach firstly applies a feature selection algorithm,[1] which combines the Boruta algorithm [24], [25] and permutation-based feature selection methods embedded in RFs [27] through a cross-validation strategy (see Section V-A). Secondly, RF classifiers (Section V-A2) are trained on the selected features. To summarize and "explain" the trained RFs, ATs [28], [29] are generated by the former trained RFs (SectionV-A3).

The results computed by RFs are then compared to those obtained by applying GLMs [31], [129] (see Section V-B). GLMs have been chosen since they may be considered as a more powerful extension of logistic regression models, which have been widely used in the medical research field both for their simplicity and for the explainability of their predictions. Since GLMs use a combination of Lasso and Ridge constraints to select the most important features, they were applied to the imputed set without previously applying any feature selection algorithm.

### A. FEATURE SELECTION AND RISK PREDICTION WITH BORUTA, RANDOM FOREST AND ASSOCIATIVE TREES

In this section we describe the overall induction process at the basis of the proposed risk-prediction scheme exploiting RFs and ATs.

#### 1) FEATURE SELECTION

Feature selection is performed on the training set through an internal 5-fold cross-validation (5-cv), where 4 folds are used in each iteration as an "internal" training set to train a RF classifier on the features selected as confirmed or tentative by the Boruta algorithm.

Precisely, Boruta [24]–[26] starts with the complete set of features and applies $n$ iterations that each train a RF on a feature set augmented by "fake features" obtained by random permutations of the actual ones. The features that, for a statistically significant number of iterations, are less/more relevant than all the fake features (relevance is quantified by the mean decrease in accuracy when the feature is permuted), are selected and removed/confirmed. When Boruta has executed $n$ iterations, all features for which a decision has not been taken are returned as tentative. Boruta is a promising feature selection method whose analysis of shuffled, fake features mitigates the impact of false correlations between features and target labels, which sometimes leads to overfitting [25]. However, even when setting a high value for $n$, some features are returned as tentative. Unfortunately, the relevance computed by Boruta cannot be used for selecting/discarding such features because such value is biased by the fake features used by the method. Moreover, Boruta does not account for class imbalance. To remove some uncertainty, Boruta is therefore also internally applied within a 5-fold cross-validation (5-cv), and all the features returned as confirmed are selected,

[1] Feature selection is applied on the training set to avoid incurring a selection bias [128].

together with those selected as tentative at least 3 out of 5 times.

The existence of tentative features and the lack of robustness with respect to class imbalance is the reason why we applied the 5-cv, which trains weighted RFs on the confirmed and tentative features: this approach assigns a ''permutation test importance'' [27], [32], [33] to each of the features, in turn evaluated on the left out fold. Therefore, after the 5-cv iterations, the mean importance for each feature is computed and normalized so that the sum of the normalized importances equals one. The most important features are finally selected by sorting the normalized importance in decreasing order and selecting the features that retain 0.95 of the cumulative sum.

The feature importance can be evaluated by using either the ''mean decrease in node impurity'' (via the Gini criterion), which essentially evaluates how much each feature decreases the mean impurity over all the trees of the forest, or the ''mean decrease in accuracy'' after feature value permutation, which essentially evaluates how the accuracy of the prediction over the training set decreases when the feature values are shuffled. We preferred the ''mean decrease in accuracy'' (also called ''permutation test'') to the Gini criterion, since the latter may lead to biased results [32], [33].

Once the most informative features have been selected, the selected feature set is input to RFs (described in the next subsection V-A2), which are trained to predict the patients' risk. Subsequently, the trained RFs are merged to summarize all their rules through Associative Decision Trees (subsection V-A3), which provides more explainable predictions.

### 2) RISK PREDICTION TROUGH RANDOM FORESTS

The main advantages of RF classifiers are the potential explainability of their decisions, their capability of computing adimensional importance measures (''mean decrease in accuracy'') describing the relevance of each variable in the risk prediction task, and the few number of involved hyper-parameters [27]. The main hyper-parameters of RFs are:

- the number of trees to grow: this parameter was set to 100 since grid search allowed us to discover that higher numbers of trees not only increase computational time, but also tend to produce overfitting
- the number of variables to sample for each split: this number is automatically set in order to maximize the misclassification cost on the training set, by a greedy search algorithm which evaluates all the points in the range $\{n_{\text{feat}}/3, \ldots, n_{\text{feat}}\}$, where $n_{\text{feat}}$ is the number of features obtained after feature selection;
- minimum size of terminal nodes, where the size of a node is the number of training samples falling in that node: low values for this parameter may cause overfitting and tend to grow tall trees; based on this consideration and following the advice of clinical experts, we require that the minimum node size is 10.

Though easy to use, RFs are not robust with respect to class imbalance. Therefore, training was performed by constraining the number of bootstrapped samples per class to be less than or equal to the number of samples of the underrepresented class [130]. Moreover, recalling that RFs are trained and tested by applying a 10-cv, at the end of the latter we have 10 importance measures for each variable. To obtain a single estimate for each variable, we first normalize the importance computed in each cross-validation run so that they sum to one; the global estimate of each variable in the 10-cv run is then computed as the average of the normalized importance for that same variable.

### 3) ASSOCIATIVE TREES GENERATED BY RANDOM FORESTS

As mentioned before, RFs are considered as relatively explainable models since their output is a set of decision trees, each describing a set of classification rules. When a novel sample must be classified, all the trees in the trained RF provide their response and majority voting is used to provide the pooled response. Despite the simplicity of this process, retrieving the rules that led to a specific classification becomes difficult when many trees are grown. For this reason, we translated each trained classifier into a simple associative tree, as described in [28], [29]. Associative classifiers are defined as models made of rules ''whose right-hand side are restricted to the classification class attribute'' [131]. In other words, they are composed by a set of rules which are consecutively evaluated. The first rule that is met provides the classification label. Associative Trees (ATs) are a simple representation of associative classifiers (see Fig. 6), characterized by the fact that each node which is not a terminal node has one child which is a leaf node.

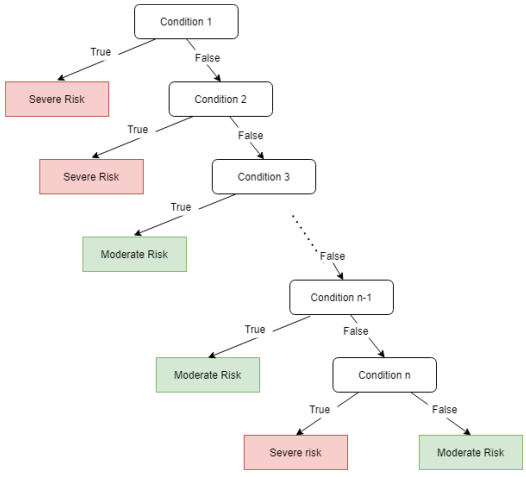


**FIGURE 6. An associative tree. The tree consecutively evaluates all the conditions, until a condition is met, bringing to a decision.**

To generate an AT from a trained RF, the following steps are consecutively applied.

1) All trees are translated into logical expressions, through a process that follows the paths from roots to leaves. Since the most informative splits often occur in the top

level of a tree, the rule extraction process is stopped when it reaches the node at depth 6 [28], [29]. This procedure allows a rule to be extracted from each tree that is composed of a maximum of 6 atomic conditions joined by the logical AND operator. Each atomic rule is expressed as $C \implies T$, where $C$, referred to as the condition of the rule, is a conjunction of variable-value pairs, and $T$ is the outcome of the rule.

2) The trees resulting from RFs are sometimes redundant; the first step after the rule extraction is therefore aimed at applying logical simplification to the rules, discarding redundant duplicates.
3) Next, each rule is pruned by eliminating its atomic conditions whose removal increases the classification error by not more than 0.05. The error of a rule is intended as the proportion of misclassified instances among all those satisfying the rule condition.
4) After pruning, each rule is expressed by a binary vector, whose length is equal to the number of samples. Each element of the vector is set to one if and only if the rule is satisfied for the corresponding sample. This encoding is used in order to apply a simple feature selection algorithm [29], which in turn allows discarding redundant and non-informative rules. A further reduction is done by discarding rules whose frequency is less than 0.01, where frequency is defined as the proportion of training instances satisfying the rule condition.
5) The remaining set of rules is finally used to combine an AT, by using a greedy iterative algorithm; at each iteration, the best rule (intended as that with lowest error, breaking ties by taking the most frequent rule) is added to the tree until no more rules remain. After inserting each best rule, all remaining rules are re-evaluated and those with a frequency lower than 0.01 are removed before the iteration continues.

### B. GENERALIZED LINEAR MODELS

GLMs [31], [129] generalize linear regression by allowing the learnt linear model to be related to the response variable via a link function. Ordinary linear regression estimates the coefficients of a linear model combining a set of variables for predicting the expected value of the response variable, implying normality for the conditional distribution of the response variable given the values of the explanatory variables in the model. GLMs allow this conditional distribution follow different models, e.g., Gaussian for continuous responses or binomial when dealing with a binary response.

In our problem, the binomial function links the linear combination of explanatory values to the response variable; in practice, given a training set $T = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^p$ containing $N$ samples and their labels $\{y_1, \ldots, y_N\} \in [0, 1]$, GLMs find the $p+1$ coefficients $(\beta_0, \beta) \in \mathbb{R}^{p+1}$ by using a penalized logistic regression, whose objective function uses the negative binomial log-likelihood:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} -\Big[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T\beta) - \log\left(1 + e^{(\beta_0 + x_i^T\beta)}\right)\Big] + \lambda\Big[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\Big]$$

Note that the objective contains an (elastic-net) penalty factor, weighted by the tuning parameter $\lambda$ which not only reduces the negative effect of degeneracies when $p > N$ or $p \approx N$, but also regularizes and selects the most important variables. Such penalty factor mixes ridge constraint (when $\alpha = 0$, which tends to select correlated predictors shrinking their coefficients [132]) and lasso constraint ($\alpha = 1$, which selects only one of the correlated predictors [132]).

In our implementation, GLMs work on standardized data, and grid search is applied through an internal 10-cv to automatically choose the most suitable values for $\lambda$ and $\alpha$. The coefficients computed for each variable are often regarded as an (adimensional) measure related to the importance of the variable in the prediction problem. Recalling that, for each imputed set and fold stratification we obtain an unbiased evaluation by applying 10-cv, we averaged the coefficients obtained in the 10 folds to compute a unique coefficient for each feature.

## VI. RESULTS

Our dataset $D$ contains 41 features; 14 (numeric) features (saturation values and laboratory values) have missing values, for a total of 188 missing values. Among features with missing values, those having the highest number of missing values are the two variables related to oxygen (saturation) values (SpO2 in free air, having 50 missing values, and PaO2.PF, having 45 missing values; both values are missing for only 9 patients), followed by lymphocyte values (%lymphocyte has 16 missing values, lymphocyte count has 15 missing values, and all patients with missing lymphocyte values have also %lymphocyte missing); the other 10 features lack at most 5 values.

In this this section we firstly report the experimental setup (Section VI-A); secondly, we report an exhaustive description of the computed results (Section VI-B).

### A. EXPERIMENTAL SETUP

To obtain an unbiased evaluation, all the risk prediction models were trained and tested on each of the $m = 50$ imputed sets, by applying an (external) stratified 10-fold cross-validation (10-cv).

Further, since the results may depend on the specific randomly computed 10-fold stratification, each risk predictor model is applied on each imputed set $n_{cv} = 5$ times by applying $n_{cv}$ different 10-fold stratifications.

In this way, given a performance evaluation measure among those we chose to collect (described in subsection VI-A1), for each imputed set we obtain $m \times n_{cv}$

values, which are combined through Rubin's rule [22], [122], [123] (described in Section VI-A2), and statistically compared with a one-sided Wilcoxon rank-signed test (see subsection VI-A3).

#### 1) PERFORMANCE EVALUATION MEASURES

Several published methods were evaluated by the C-statistic (that is, the area under the ROC curve, or AUC [133]). The C-statistic is the probability that the model predicts a higher risk for positive samples. Moreover, it is adimensional, and thus it allows the comparison of different predictors. However, as highlighted in [133], the C-statistic is not an exhaustive description: for instance, it does not account for uneven class distributions, and hides the method performance on the positive or on the negative samples. To provide an exhaustive description, we therefore decided to record also sensitivity (performance on positive samples), specificity (performance on negative samples), accuracy (ratio of misclassified samples), and F1 score (harmonic mean of precision and recall, which accounts for uneven class distributions). In practice, we used the AUC to select the most promising combinations of imputation method, univariate variable imputation order, and risk prediction method (RFs, ATs, or GLMs). We subsequently selected the most appropriate risk prediction model by analyzing the performance on the positive and negative samples, as described by the remaining performance measures.

#### 2) COMBINING RESULTS THROUGH RUBIN's RULE

Given the imputed set, we obtain a robust comparative evaluation by training and testing each predictor model (RF, AT, or GLM) $n_{cv} = 5$ times on each of the $m$ imputed sets, by using $n_{cv}$ different, randomly generated 10-fold stratifications. For each stratification and each model, we output the previously described performance evaluation measures (namely, AUC, sensitivity, specificity, F1 score, accuracy) and, for each variable, a measure of its relevance in the risk prediction task.

More precisely, given a risk prediction model RM (that is, RF, AT, or GLM) and an imputation method imp (*miceRF* or *missForest*), producing $m$ imputed sets $D_{\text{imp}}(i)$, $i = 1, \ldots, m$, Rubin's rule [53], [55] provides a way to combine the "results" (that is either risk prediction performance or the importance of a single variable) computed by the $n_{cv}$ different runs of each risk prediction model on each of the $m$ imputed sets. Precisely, let $\text{RM}(D_{\text{imp}}(i), \text{fold}(t))$ denote the result computed by RM (e.g., RF importance for a single variable), when using the $t^{\text{th}}$ fold stratification, $\text{fold}(t)$ ($t = 1, \ldots, n_{cv}$), and the $i^{\text{th}}$ imputed set $D_{\text{imp}}(i)$ ($i = 1, \ldots, m$). For the sake of simplicity, we organize all $\text{RM}(D_{\text{imp}}(i), \text{fold}(t))$ values in a matrix $\text{RM}(i, t)$ with $m$ rows and $n_{cv}$ columns. For a fixed imputed set $i$, the mean over the fold stratifications (over the columns of the matrix):

$$\theta(i) = \frac{1}{n_{cv}} \sum_{t=1}^{n_{cv}} \text{RM}(i, t) \tag{1}$$

is the performance over each $D_{\text{imp}}(i)$, and the mean over all such values is the global result computed using RM and imp:

$$\overline{\theta} = \frac{1}{m} \sum_{i=1}^{m} \theta(i). \tag{2}$$

Rubin's rule [53], [55] defines the variance of such result by applying the law of total variance [134] to consider both the uncertainty that comes from the processing method applied to each of the imputed datasets (within-imputation variance) and the added uncertainty that comes from the multiply imputed data (between-imputation variance). Precisely, variances computed along each row (over the $n_{cv}$ values) are the within-imputation variances over each imputed set:

$$\text{Var}(\theta(i)) = \frac{1}{n_{cv} - 1} \sum_{t=1}^{n_{cv}} (\text{RM}(i, t) - \theta(i))^2 \tag{3}$$

while the mean of all the $m$ within-imputation variances is the global within-imputation variance:

$$W = \frac{1}{m} \sum_{i=1}^{m} \text{Var}(\theta(i)) \tag{4}$$

and the between-imputation variance is the variance of the performance measures achieved over all the imputations:

$$B = \frac{1}{m - 1} \sum_{i=1}^{m} \left(\theta(i) - \overline{\theta}\right)^2. \tag{5}$$

Then the global (normalized) variance associated obtained by imp and RM is computed as [122]:

$$\text{Var}\,(\theta) = W + \left(1 + \frac{1}{m}\right) B. \tag{6}$$

At the state of the art, MI is used before linear or logistic regression, to determine the coefficients that link predictor variables to a response variable. As reported in [122], for two-sided hypothesis testing of single regression coefficients after MI, the Wald statistic:

$$\text{Wald} = \frac{\overline{\theta} - \theta_0}{\text{Var}\,(\theta)} \tag{7}$$

can be used to assess the significance of the difference between the computed estimate $\overline{\theta}$, and the value under the null hypothesis, $\theta_0$ (which is generally set to zero), exploiting the fact that Wald follows a chi-square distribution with 1 degree of freedom.

#### 3) STATISTICAL ANALYSIS OF COMPUTED RESULTS

Besides computing a global performance measure by applying Rubin's rule, statistical analysis was applied to compare the performance values computed by different combinations of imputation algorithm and risk prediction approach (Section V). Precisely, we averaged the $n_{cv} = 5$ performance values obtained on each imputed set (we recall that $n_{cv}$ are the different 10-fold stratifications), thus obtaining $m$ mean values for each imputation method + risk

**TABLE 2.** Global performance measures computed by each imputation algorithm + risk prediction model.

| | model | AUC (var) | Sensitivity (var) | Specificity | F1-score | Accuracy |
|---|---|---|---|---|---|---|
| **missForest** | RF | **0.81** (0.00007) | **0.72** (0.00016) | 0.76 (0.00006) | **0.62** (0.00009) | **0.74** (0.00006) |
| | AT | 0.67 (0.00013) | 0.51 (0.00039) | 0.83 (0.00020) | 0.53 (0.00028) | 0.67 (0.00013) |
| | GLM | 0.80 (0.00001) | 0.56 (0.00002) | 0.86 (0.00001) | 0.62 (0.00002) | 0.71 (0.00001) |
| **miceRF** | RF | 0.79 (0.00011) | 0.70 (0.00034) | 0.74 (0.00012) | 0.60 (0.0002) | 0.72 (0.00014) |
| | AT | 0.65 (0.00027) | 0.48 (0.00079) | 0.82 (0.00022) | 0.50 (0.00062) | 0.65 (0.00027) |
| | GLM | 0.78 (0.0005) | 0.53 (0.00025) | 0.85 (0.00004) | 0.59 (0.00014) | 0.69 (0.00009) |

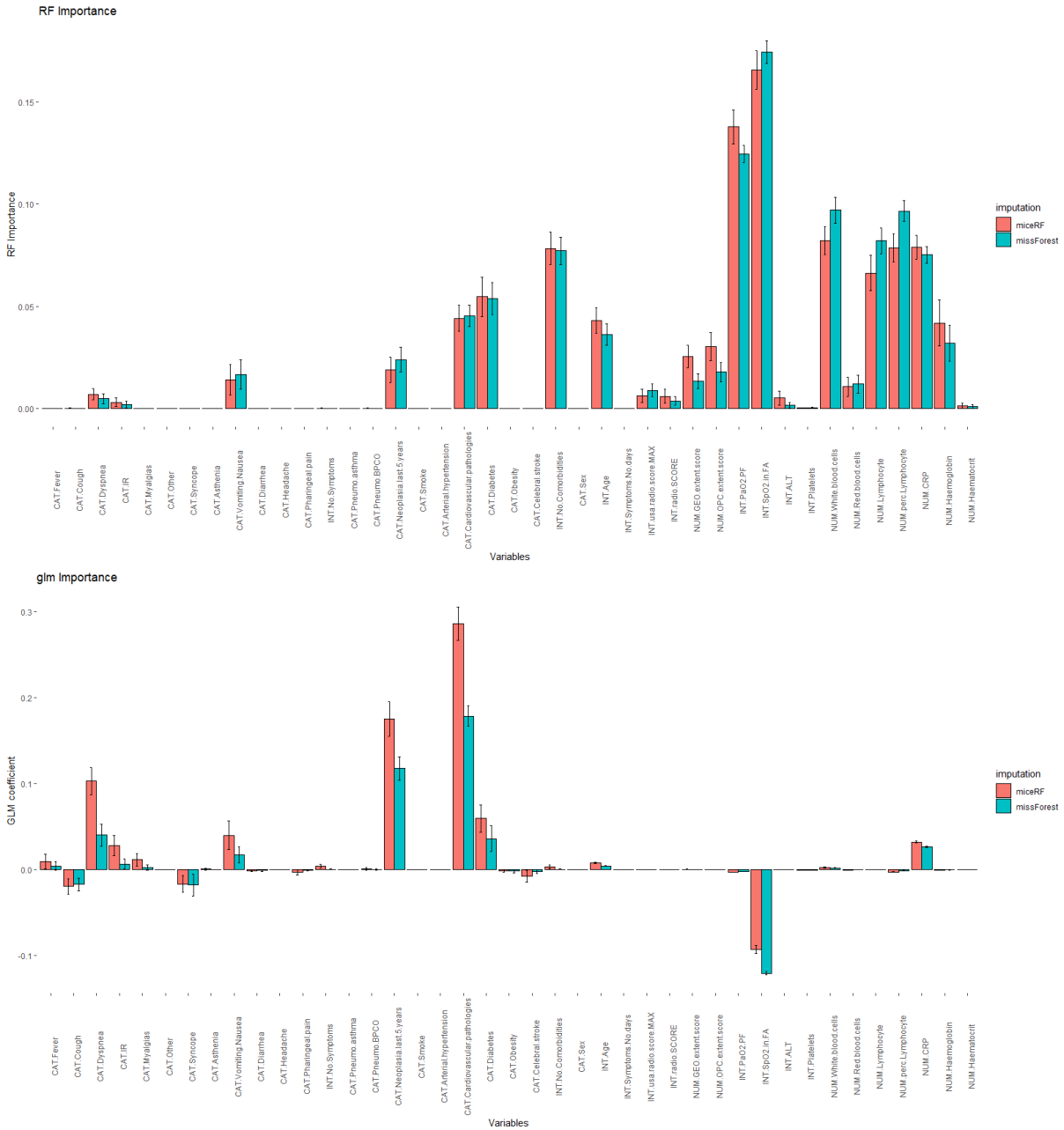


**FIGURE 7.** Top: estimates (and standard errors) of the feature relevance computed by RFs. Bottom: estimates (and standard errors) of the feature coefficients computed by GLMs. Only the significant feature relevances/coefficients are plotted.

prediction approach. At this stage, the one-sided Wilcoxon signed-rank test at the 99% confidence level (p-value $< 0.01$) was applied to perform the statistical comparison between the distributions of the $m$ mean performance values computed by a combination of imputation algorithm and risk prediction model.

| Variable | RF | | | | GLM | | | |
|---|---|---|---|---|---|---|---|---|
| | Global Estimate | Standard Error | p-value | Normalized Estimate | Global Estimate | Standard Error | p-value | Normalized Estimate |
| Fever | | | | | 0.004255 | 0.00501 | 1.7E-07 | 0.01 |
| Cough | | | | | -0.017122 | 0.00769 | 7.3E-22 | -0.03 |
| Dyspnea | 0.00475 | 0.00251 | 1.6E-52 | 0.00 | 0.040332 | 0.01293 | 1.3E-19 | 0.07 |
| Respiratory Failure (IR) | 0.00201 | 0.00162 | 5.3E-48 | 0.00 | 0.006845 | 0.00590 | 7.5E-12 | 0.01 |
| Myalgias | | | | | 0.002577 | 0.00309 | 1.2E-10 | 0.00 |
| Other | | | | | 0.000048 | 0.00008 | 1.4E-88 | 0.00 |
| Syncope | | | | | -0.017777 | 0.01274 | 1.6E-09 | -0.03 |
| Asthenia | | | | | 0.000083 | 0.00013 | 4.3E-74 | 0.00 |
| Vomiting.Nausea | 0.01645 | 0.00716 | 6.5E-23 | 0.02 | 0.017547 | 0.00932 | 2.2E-14 | 0.03 |
| Diarrhea | | | | | -0.000728 | 0.00101 | 1.7E-19 | 0.00 |
| Headache | | | | | | | | |
| Pharingeal.pain | | | | | -0.000498 | 0.00081 | 6.8E-13 | 0.00 |
| No.Symptoms | 0.00003 | 0.00006 | 4.4E-101 | 0.00 | 0.000494 | 0.00055 | 8.9E-58 | 0.00 |
| Pneumo.asthma | | | | | | | | |
| Pneumo.BPCO | | | | | 0.000344 | 0.00053 | 7.8E-24 | 0.00 |
| Neoplasia.last.5.years | 0.02381 | 0.00604 | 2.8E-47 | 0.02 | 0.117632 | 0.01338 | 5.1E-43 | 0.20 |
| Smoke | | | | | | | | |
| Arterial.hypertension | | | | | | | | |
| Cardiovascular.pathologies | 0.04538 | 0.00524 | 2.1E-102 | 0.05 | 0.178751 | 0.01189 | 2.2E-82 | 0.30 |
| Diabetes | 0.05387 | 0.00776 | 7.3E-51 | 0.05 | 0.036098 | 0.01527 | 4.9E-13 | 0.06 |
| Obesity | | | | | -0.001529 | 0.00208 | 2.6E-11 | 0.00 |
| Celebral.stroke | | | | | -0.001899 | 0.00255 | 4.6E-10 | 0.00 |
| No.Comorbidities | 0.07721 | 0.00663 | 4.2E-106 | 0.08 | 0.000504 | 0.00054 | 1.4E-94 | 0.00 |
| Sex | | | | | | | | |
| Age | 0.03605 | 0.00520 | 9.3E-88 | 0.04 | 0.004460 | 0.00049 | 0.0E+00 | 0.01 |
| Symptoms.No.days | 0.00003 | 0.00006 | 2.8E-73 | 0.00 | -0.000004 | 0.00001 | 0.0E+00 | 0.00 |
| usa.radio.score.MAX | 0.00877 | 0.00311 | 1.9E-62 | 0.01 | | | | |
| radio.SCORE | 0.00362 | 0.00217 | 2.0E-49 | 0.00 | | | | |
| GEO.extent.score | 0.01328 | 0.00365 | 2.1E-63 | 0.01 | | | | |
| OPC.extent.score | 0.01774 | 0.00466 | 2.8E-58 | 0.02 | | | | |
| PaO2.PF | 0.12467 | 0.00415 | 0.0E+00 | 0.12 | -0.002341 | 0.00004 | 0.0E+00 | 0.00 |
| SpO2.in.FA | 0.17464 | 0.00557 | 1.5E-277 | 0.17 | -0.120213 | 0.00214 | 0.0E+00 | -0.20 |
| ALT | 0.00149 | 0.00129 | 9.6E-40 | 0.00 | | | | |
| Platelets | 0.00027 | 0.00044 | 4.9E-23 | 0.00 | -0.000166 | 0.00005 | 0.0E+00 | 0.00 |
| White.blood.cells | 0.09714 | 0.00637 | 1.2E-97 | 0.10 | 0.001907 | 0.00042 | 0.0E+00 | 0.00 |
| Red.blood.cells | 0.01191 | 0.00444 | 2.0E-39 | 0.01 | -0.000040 | 0.00007 | 1.7E-92 | 0.00 |
| Lymphocyte | 0.08213 | 0.00638 | 2.0E-97 | 0.08 | 0.000040 | 0.00004 | 0.0E+00 | 0.00 |
| %Lymphocyte | 0.09663 | 0.00510 | 6.3E-207 | 0.10 | -0.001002 | 0.00038 | 0.0E+00 | 0.00 |
| C-Reactive Protein (CRP) | 0.07528 | 0.00406 | 1.3E-268 | 0.08 | 0.026559 | 0.00069 | 0.0E+00 | 0.04 |
| Haemoglobin | 0.03197 | 0.00891 | 3.6E-24 | 0.03 | -0.000118 | 0.00014 | 1.9E-165 | 0.00 |
| Haematocrit | 0.00085 | 0.00098 | 7.9E-34 | 0.00 | -0.000007 | 0.00001 | 0.0E+00 | 0.00 |

**FIGURE 8.** **Relevance/coefficient estimates computed by RFs and GLMs. Only the significant estimates are reported. For GLMs, red bars highlight negative coefficients (that is variables, inversely related to the risk).**

### *B. COMPARATIVE EVALUATION*

We started our comparative evaluation by applying the one-sided signed-rank Wilcoxon to compare the performance measures computed when using *miceRF* or *missForest* as the first step for data imputation (see Table 4 in Appendix B). We firstly compared the risk prediction performance measures achieved by the two imputation methods, irregardless of which risk prediction model is used (column "All risk models" in Table 4, Appendix B). Then, we iterated over all risk prediction models, in turn fixing one of them and comparing the performance distribution when using either *miceRF* or *missForest* followed by the fixed risk prediction model (columns "RF", "AT", and "GLM" in Table 4, Appendix B). Only the specificities obtained with fixed ATs do not show any statistically significant difference; otherwise, *missForest* always achieved the best result. Therefore, we conclude that in this risk prediction task *missForest* is the most suitable imputation method.

In Table 2 we show the performance measures (and variance) computed by using Rubin's rule to combine the results computed on the 50 × 5 10-fold cross-validation runs performed by each of the three risk-prediction models when using either the datasets imputed by using *missForest* or *miceRF* and the increasing univariate imputation order. For each column in Table 2, the highest global mean, confirmed by the one-sided Wilcoxon signed-rank test (p-values reported in Table 5, Appendix B), is highlighted with bold typeface. The results show that, for what regards the AUC, the sensitivity, the F1-score, and the accuracy, RF is the best performing method, especially when combined

with *missForest*. Note that, though no statistically significant difference has been found by one-sided Wilcoxon signed-rank test when comparing the specificity values computed by the three models (see Table 5 in Appendix B), the seemingly lower specificity achieved by RFs both in the comparison with ATs and GLMs is balanced by a higher sensitivity. In practice, both ATs and GLMs are affected by class imbalance, while RFs can cope with such problems by balancing the sampled points during the training phase. Since in our risk prediction model type II errors are worse than type I errors, we can state that the combination of *missForest* + (balanced) RFs is the best performing risk prediction model.

Finally, since our principal aim is the identification of the most important predictors of severe risk, we analyzed the normalized variable importance computed by RFs, when using either *missForest* or *miceRF* as the preliminary imputation steps. For the sake of comparison, we also considered the coefficients computed by GLMs. After applying Rubin's rule (see Section VI-A2) to compute, for each feature, the mean (RF) importance or the mean (GLM) coefficient, and their respective variances and standard errors, we applied the Wald significance test to determine the coefficients that were significantly different from zero. The significant RF importance and GLM coefficients, along with their standard errors, are plotted, in the top and bottom panel of Fig. 7, respectively. Fig. 8 reports the precise values of coefficients resulting as significant (Column "Global Estimate") when using *missForest* followed by RFs (left panel) or GLMs (right panel), together with their standard errors, and the p-values computed by the Wald test. In the visual table in Fig. 8 a column-wise visual comparison of the reported values is allowed by data bars, whose different colors highlight that row-wise comparison is not meaningful. However, to allow a visual comparison of the two global estimates computed by RFs and GLMs, Column "Normalized Estimate" contains the RF variable relevance (left) and GLMs coefficients (right) normalized so that the sum over the column equals one.

Interestingly, the distribution of the feature relevance computed by RFs is very different from that computed by GLMs; generally speaking, RFs mainly consider as relevant all the laboratory variables, the saturation values, and the radiological scores. Even if GLMs predictors selected a similar number of variables (26 variables were selected by GLMs and 25 variables were selected by RFs, see Figs. 7 and 8), and 19 of them are also contained in the subset of variables selected by RFs, the relative importance GLMs attributed to the variables is less balanced. Indeed, GLMs attributed a much higher importance to two comorbidities (cardiovascular pathologies and neoplasia in the last 5 years), followed by only one saturation value (spO2.in.FA), two symptoms (presence of Dyspnea, and Vomiting/Nausea), and only C-Reactive Protein was used among the laboratory variables; the other variables had negligible importance. Such results can be explained by considering that GLMs do not take into account class imbalance; the objective function is easily minimized by decreasing the number of false positives (high specificity), at the expense of a high false negative proportion. Therefore, the features and their relative importance identified by GLMs may be deemed as relevant in the correct identification of patients at low risk. Conversely, the feature selection and importance weighting performed by the proposed RF-based risk prediction system can properly balance sensitivity and specificity.

In sum, we believe that the feature relevance computed through the feature extraction algorithm presented in Section V-A1, followed by the ("balanced") RFs, is the most reliable. Indeed, the relevant features are similar to those extracted by the papers reported in Table 3 in [15], though none of those works sorted features according to their relevance. We identified the following variables as most relevant (in decreasing order): saturation values (spO2 in free air and paO2.PF), white blood cell counts, lymphocyte counts, the number of comorbidities, C-reactive protein, diabetes, cardiovascular pathologies, age, haemoglobin, neoplasia in the past 5 years, the opacity score computed

Correlation between missForest imputed data and LABEL

| Variable | negative correlations | positive correlations |
|---|---|---|
| Fever | | 0.0308 |
| Cough | -0.0637 | |
| Dyspnea | | 0.2361 |
| Respiratory Failure (IR) | | 0.1905 |
| Myalgias | -0.0023 | |
| Other | -0.0095 | |
| Syncope | -0.0634 | |
| Asthenia | | 0.0291 |
| Vomiting.Nausea | | 0.0561 |
| Diarrhea | -0.0232 | |
| Headache | -0.0259 | |
| Pharingeal.pain | -0.0689 | |
| No.Symptoms | | 0.1108 |
| Pneumo.asthma | -0.0364 | |
| Pneumo.BPCO | | 0.0776 |
| Neoplasia.last.5.years | | 0.1367 |
| Smoke | -0.0204 | |
| Arterial.hypertension | | 0.1279 |
| Cardiovascular.pathologies | | 0.2077 |
| Diabetes | | 0.1627 |
| Obesity | | 0.0246 |
| Celebral.stroke | | 0.0199 |
| No.Comorbidities | | 0.1947 |
| Sex | -0.0343 | |
| Age | | 0.2900 |
| Symptoms.No.days | -0.0029 | |
| usa.radio.score.MAX | | 0.2095 |
| radio.SCORE | | 0.2330 |
| GEO.extent.score | | 0.2724 |
| OPC.extent.score | | 0.2596 |
| PaO2.PF | -0.4070 | |
| SpO2.in.FA | -0.4593 | |
| ALT | | 0.0966 |
| Platelets | -0.0009 | |
| White.blood.cells | | 0.1307 |
| Red.blood.cells | -0.1049 | |
| Lymphocyte | -0.1258 | |
| perc.Lymphocyte | -0.2548 | |
| CRP | | 0.3513 |
| Haemoglobin | -0.1210 | |
| Haematocrit | -0.0976 | |

**FIGURE 9. Global, significant estimates of pooled correlation coefficients between each feature and the label computed on the 50 sets imputed by *missForest*.**

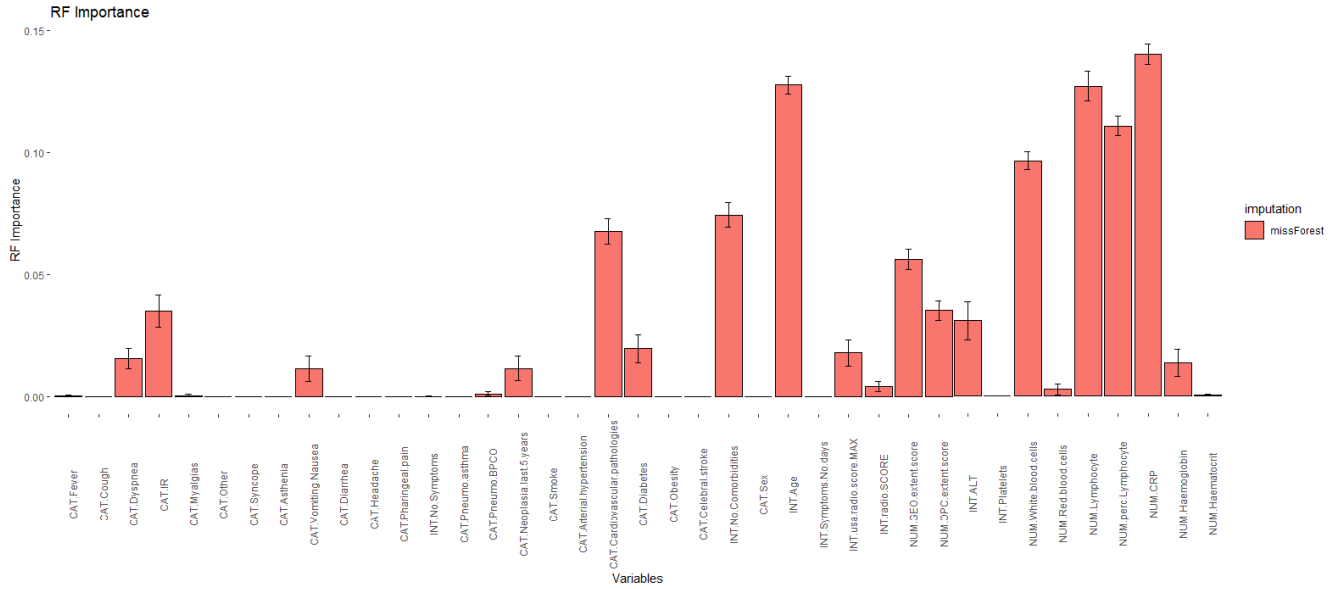


**FIGURE 10.** **Pooled significant estimates of feature importance computed by RFs on the 50 sets (imputed by *missForest*) when saturation variables are removed.**

on CXR by the deep network, nausea, the extent of COVID-19 pattern computed on CXR by the deep network (*OPC.extent.score* and *GEO.extent.score*), the red blood cell count, *usa.radio.score*, dyspnea, *radio.score*, respiratory failure (IR), and haematocrit.

Interestingly, the relevance attributed by RFs to radiological features are quite low; moreover, the radiological score computed by deep networks is higher than that computed by experts. To understand such results, we considered the 50 sets imputed with *missForest* and, for each set, we computed the pairwise Pearson, Spearman, and Kendall correlations between features. Subsequently, we used Rubin's rule in order to pool the mean correlation estimates and to verify their significance (see Fig. 11 in Appendix C). The same procedure was used to compute an estimate of the correlation between each feature and the label (see Fig. 9).

By observing the computed pairwise correlations (in Fig. 11, Appendix C), we note that radiological features are positively correlated (as expected); moreover, they also correlate with C-Reactive Protein (CRP), and have an inverse correlation with the saturation values. Concerning the correlations with the label (Fig. 9), we note that saturation values have the highest (absolute) correlation with the label, followed by CRP, the radiological scores computed by CovidNet, and the radiological scores computed by experts. The obtained correlation results explain the computed relevance; indeed, among a set of correlated features, RFs tend to choose the variables with the highest discriminative power, neglecting the other ones.

As expected, oxygen saturation values are inversely correlated with some symptoms (dyspnea and respiratory failure - IR) and comorbidities (cardiovascular pathologies or arterial hypertension) (Fig. 11, Appendix C). Therefore we performed a test by running all the algorithms without using the two saturation variables (SpO2 in free air - SpO2.in.FA - and PaO2.PF); we retrained RFs (on the features selected as described in Section V-A1) by using 50 MI sets imputed by *missForest*.

With this setting the pooled risk prediction estimates displayed a reduction in accuracy by a mean of 0.06, over the five performance measures (AUC from 0.81 to 0.76, sensitivity from 0.72 to 0.66, specificity from 0.76 to 0.71, F1 score from 0.62 to 0.55, accuracy from 0.74 to 0.68).

The pooled, significant feature-importance estimates are shown in Fig. 10. In this case, CRP is attributed a much higher relative relevance, together with patient's age. The importance of lymphocyte values, and of all the laboratory variables, is confirmed and radiological features (particularly those computed by CovidNet) have an increased relevance. As expected, those symptoms and comorbidities that are related to saturation values have a significant importance.

## VII. LIMITATIONS OF THE STUDY

Though promising results were obtained with the proposed risk-prediction system, our study has some limitations.

At first, though we use RF classifiers for the high explainability of their decisions, the complexity of RFs explanations grows with the number of trained trees. For this reason, we propose using ATs, which are derived by the trained RFs to produce a unique, simple, explainable predictor summarising the RF rules. Unfortunately, ATs are not robust with respect to class imbalance. This is because the greedy procedure used to generate ATs iteratively adds the best rule from the RFs, where rule evaluation is measured on all the training set, without normalization with respect to the between-class proportions.

**TABLE 3.** Ranges of between-imputation variances achieved by the four imputation methods when using the increasing and decreasing univariate imputation order.

| Imputation method | Increasing order | Decreasing Order |
|---|---|---|
| *missForest* | 2.3E-04 ± 0 [2.3E-04, 2.5E-04] | 2.4E-04 ± 0 [1.7E-04, 2.6E-04] |
| *miceRF* | 1.5E-02 ± 4E-05 [1.3E-02, 1.5E-02] | 1.6E-02 ± 1.E-04 [1.5E-02, 2.1E-02] |
| *micePMM* | 1.4E-02 ± 3E-05 [1.2E-02, 1.4E-02] | 1.4E-02 ± 2E-05 [1.3E-02, 1.5E-02] |
| *distFree* | 3.0E-01 ± 6.1E-03 [4.7E-02, 4.2E-01] | 2.5E-01 ± 1.3E-02 [1.0E-02,5.2E-01] |

**TABLE 4.** p-values resulting from the one-sided Wilcoxon signed-rank tests applied to compare the performance values computed when *miceRF* or *missForest* are used for imputation.

| | All risk models | | RF | | AT | | GLM | |
|---|---|---|---|---|---|---|---|---|
| **Alternative** | **lower** | **greater** | **lower** | **greater** | **lower** | **greater** | **lower** | **greater** |
| **AUC** | 1.42E-04 | 1.00E+00 | 9.96E-08 | 1.00E+00 | 4.40E-03 | 9.96E-01 | 4.94E-09 | 1.00E+00 |
| **Sensitivity** | 2.19E-02 | 9.78E-01 | 2.37E-07 | 1.00E+00 | 1.18E-02 | 9.89E-01 | 7.60E-03 | 9.93E-01 |
| **Specificity** | 2.67E-02 | 9.74E-01 | 1.28E-07 | 1.00E+00 | 1.91E-01 | 8.14E-01 | 1.43E-04 | 1.00E+00 |
| **F1** | 7.98E-06 | 1.00E+00 | 1.39E-09 | 1.00E+00 | 3.53E-03 | 9.97E-01 | 6.64E-06 | 1.00E+00 |
| **Accuracy** | 7.75E-03 | 9.92E-01 | 8.06E-08 | 1.00E+00 | 4.40E-03 | 9.96E-01 | 1.10E-04 | 1.00E+00 |

Therefore, even if ATs can provide simple and human understandable decision rules, a limitation of this approach is that the resulting model does not exactly fit the original RFs, and the accuracy is significantly worsened. To deal with this issue, our future work will be therefore aimed at modifying the procedure proposed in [28], [29] in order to obtain ATs robust w.r.t. class imbalance.

With this setting, the features that were considered as most relevant during training were: saturation values, laboratory values (lymphocyte counts, C-Reactive Protein, white blood cells counts, haemoglobin), variables related to comorbidities (number of comorbidities, presence of cardiovascular pathologies and/or arterial hypertension), radiological values computed through CovidNet, and presence of symptoms (vomiting/nausea or dyspnea or respiratory failure).

Another limitation of our study is that the dataset contains only 300 patients and is not public due to privacy restrictions. Since no public dataset with a larger sample size is available yet, the importance of the selected feature set was confirmed by clinical experts, but it has yet to be validated on a larger and more diverse population.

Finally, the limit of the review in [15] and of our work, which stems from the lack of a shared dataset, is that an objective comparative evaluation with state-of-the-art models is not possible. The opportunity for the scientific community to use common datasets is one of the main and important goals to simplify and speed up research activity. In summary, it is necessary to create a deidentified, shareable database to enable an objective comparative evaluation of more rigorous and exhaustively tested prediction models.

## VIII. CONCLUSION

In this article we pursued the development of a prediction model able to process clinical, radiological, and laboratory data of COVID19-related patients in order to predict their risk of severe outcomes.

The clinical and laboratory values were collected at the time of each patient's presentation to the ED, while the four radiological values were retrospectively evaluated from the patients CXR, by either pooling radiological experts' evaluations or by applying CovidNet [11], [14].

The collected variables contain missing values. Therefore, as advocated in [15], we firstly conducted a thorough analysis for identifying both the missingness pattern and the most stable missing data imputation algorithm, among two different MI techniques (*micePMM* and *miceRF*), an RF-based technique (*missForest*), and a maximum-likelihood estimator (*distFree*).

Our evaluation shows that: (i) though the maximum-likelihood imputation method is effective when used for statistically determining whether the data are MCAR or MAR [34], [35], it produces too noisy estimations; (ii) MI techniques reach stability after at least $m = 25$ multiple imputed datasets; (iii) the only method showing negligible between-imputation variance is *missForest*. Our results confirm that, at least $m = 20$ imputed sets should be used for MI to reduce between-imputation variance [35], [125].

Our results demonstrate that stable feature-selection may be obtained by combining the Boruta algorithm and permutation-based feature selection embedded in RFs. When the selected feature set is input to RFs constrained to work on balanced bootstrapped samples, the effect of class imbalance is reduced and improved results are obtained, better than those achieved by either ATs or GLMs. Additionally, we showed that all the risk prediction approaches obtain the best results when using missForest as the previous imputation model.

In conclusion, our analysis demonstrates that the best results are obtained when: (i) imputing the missing data with *missForest*, where the univariate imputation order is based on the increasing amount of missing values, (ii) selecting the most discriminative features by combining Boruta and permutation-based feature selection through an internal

**TABLE 5.** p-values obtained by one-sided Wilcoxon signed-rank test when comparing the three risk prediction models.

| Imputation method | alternative | AUC | Sensitivity | Specificity | F1 | Accuracy |
|---|---|---|---|---|---|---|
| **missForest** | RF vs AT | 7.07E-10 | 6.32E-10 | 1.00E+00 | 7.91E-15 | 7.05E-10 |
| | RF vs GLM | 7.07E-10 | 6.41E-10 | 1.00E+00 | 1.53E-05 | 7.05E-10 |
| **miceRF** | RF vs AT | 7.07E-10 | 6.32E-10 | 1.00E+00 | 7.91E-15 | 7.05E-10 |
| | RF vs GLM | 7.07E-10 | 6.41E-10 | 1.00E+00 | 1.53E-05 | 7.05E-10 |

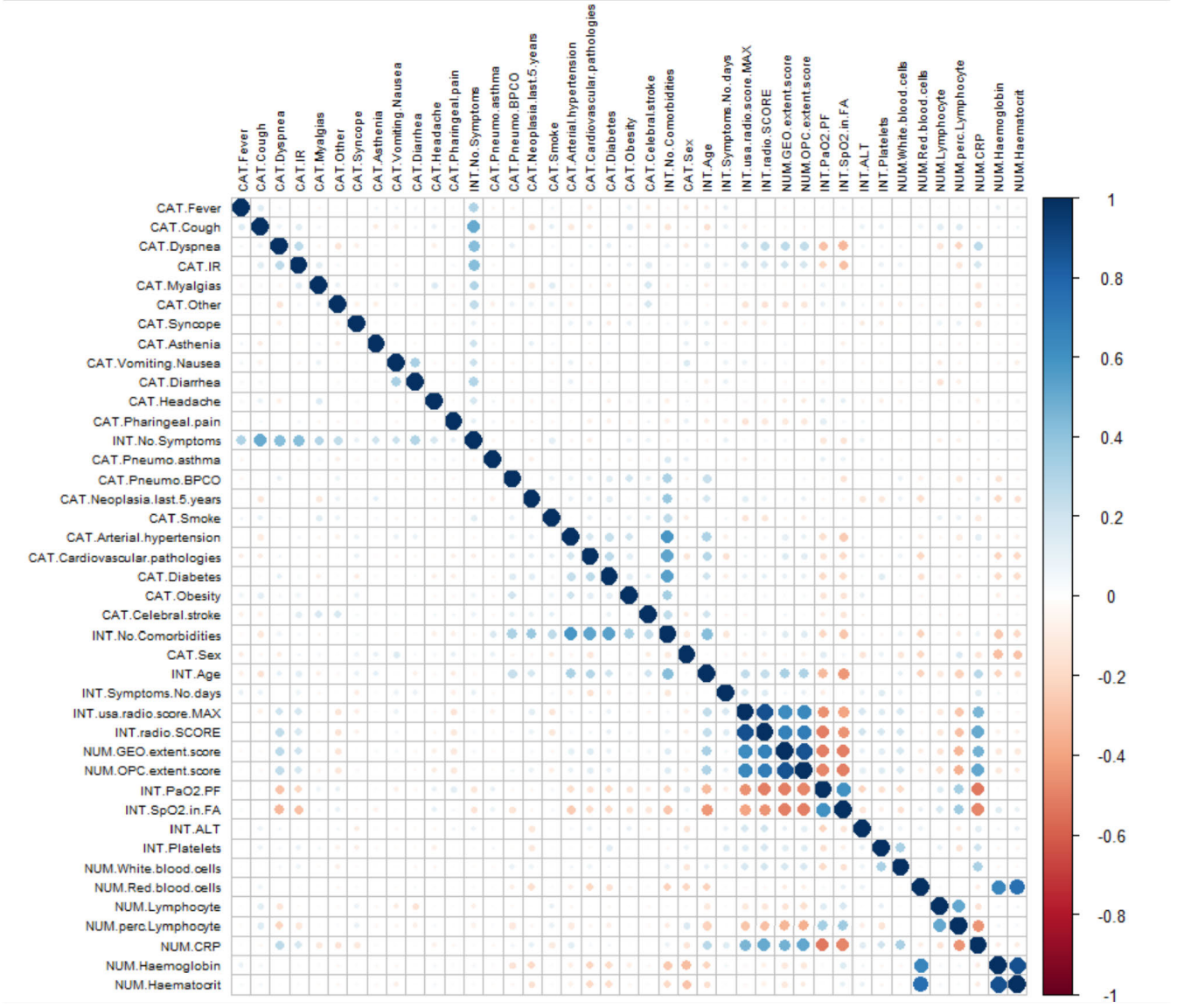


**FIGURE 11.** Pooled pairwise (Perason, spearman, and Kendall's) correlation coefficients between pair of variables computed over the 50 datasets imputed by *missForest*.

cross-validation, and (iii) training RFs on the selected features.

## APPENDIX A
## BETWEEN-IMPUTATION VARIANCES

In Table 3, the between-imputation variances obtained by the imputation methods *missForest*, *miceRF*, *micePMM*, and *distFree* are reported. As also illustrated in Fig. 5, *missForest* has negligible between-imputation variance, meaning that similar imputations are computed for each missing value. Conversely, *distFree* produces noisier imputations.

## APPENDIX B
## COMPARATIVE EVALUATION THROUGH ONE-SIDED WILCOXON SIGNED-RANK TESTS

In this appendix we firstly report the p-values for comparisons of the performance evaluation measures computed by using *miceRF* or *missForest* as imputation methods (see Table 4).

The column "All risk models" shows the p-values computed when neglecting the separation given by the employed risk prediction models. Columns "RF", "AT", and "GLM" report the p-values achieved for RFs, ATs, and GLMs as risk prediction models.

Columns "lower" report the p-value of the one-sided test where the alternative is: "*miceRF* < *missForest*"; columns "greater" report the p-value of the one-sided test where the alternative is: "*missForest* < *miceRF*". Note that only the specificities obtained with fixed ATs do not show a statistically significant difference; otherwise *missForest* always achieves the best result.

Next, in Table 5 we report the result of the one-sided Wilcoxon signed-rank test comparing RF vs AT and RF vs GLM, when either *missForest* or *miceRF* are fixed. The p-values express the probability of the null hypothesis when the alternative is "AT < *missForest*" and "GLM < *missForest*".

## APPENDIX C
## PAIRWISE CORRELATION COEFFICIENTS BETWEEN VARIABLES

To visualize the pairwise similarities/dissimilarities between variables distributions, in Fig. 11 we show the pooled correlation coefficients between pairs of variables. These pooled coefficients were computed on the 50 datasets imputed by *missForest* by calculating three pairwise correlation indices (Pearson, Spearman, and Kendall's coefficients), and by applying Rubin's rule to pool the 50 × 3 correlations computed for each pair of variables.

Note that the radiological variables have a relevant and statistically significant (inverse) correlation with saturation values and a high direct correlation with CRP. Such high correlation may be the reason why radiological features obtain a unexpectedly low importance; if two variables have similar distributions, once the RF has used the most discriminating for a split, it will never use the other one for the next splits.

In the plot, each variable name has a prefix that reminds its type; boolean variables have prefix "CAT", integer variables have prefix "INT", real variables have prefix "NUM".

## ACKNOWLEDGMENT

The authors would like to thank Prof. Alberto Bertoni for his constant and invaluable support throughout his academic career.

## REFERENCES

[1] C. Huang *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[2] Z. Huang, W. Dong, L. Ji, and H. Duan, "Outcome prediction in clinical treatment processes," *J. Med. Syst.*, vol. 40, no. 1, p. 8, Jan. 2016, doi: 10.1007/s10916-015-0380-6.

[3] J. Gliozzo, P. Perlasca, M. Mesiti, E. Casiraghi, V. Vallacchi, E. Vergani, M. Frasca, G. Grossi, A. Petrini, M. Re, A. Paccanaro, and G. Valentini, "Network modeling of patients' biomolecular profiles for clinical phenotype/outcome prediction," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, Art. no. 3612, doi: 10.1038/s41598-020-60235-8.

[4] B. R. Barricelli, E. Casiraghi, J. Gliozzo, A. Petrini, and S. Valtolina, "Human digital twin for fitness management," *IEEE Access*, vol. 8, pp. 26637–26664, 2020.

[5] S. Fong, N. Dey, and J. Chaki, *Artificial Intelligence for Coronavirus Outbreak* (Springer Briefs in Computational Intelligence). Singapore: Springer, 2020.

[6] Q. Pham, D. C. Nguyen, T. Huynh-The, W. Hwang, and P. N. Pathirana, "Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts," *IEEE Access*, vol. 8, pp. 13083–30820, 2020.

[7] S. Fong, G. Li, N. Dey, R. G. Crespo, and E. Herrera-Viedma, "Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction," *Appl. Soft Comput.*, vol. 93, pp. 1–21, Apr. 2020.

[8] A. Joshi, N. Dey, and K. Santosh, *Intelligent Systems and Methods to Combat COVID-19* (Springer Briefs in Computational Intelligence). Singapore: Springer, 2020.

[9] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial intelligence (AI) applications for COVID-19 pandemic," *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 14, no. 4, pp. 337–339, Jul. 2020, doi: 10.1016/j.dsx.2020.04.012.

[10] E. Neri, V. Miele, F. Coppola, and R. Grassi, "Use of CT and artificial intelligence in suspected or COVID-19 positive patients: Statement of the Italian society of medical and interventional radiology," *La Radiol. Med.*, vol. 125, no. 5, pp. 505–508, May 2020.

[11] A. Wong, Z. Q. Lin, L. Wang, A. G. Chung, B. Shen, A. Abbasi, M. Hoshmand-Kochi, and T. Q. Duong, "COVIDNet-S: Towards computer-aided severity assessment via training and validation of deep neural networks for geographic extent and opacity extent scoring of chest X-rays for SARS-CoV-2 lung disease severity," 2020, *arXiv:2005.12855*. [Online]. Available: http://arxiv.org/abs/2005.12855

[12] S. Ahuja, B. Panigrahi, N. Dey, T. Gandhi, and V. Rajinikanth, "Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices," *Techrxiv*, 2020, doi: 10.36227/techrxiv.12334265.v2.

[13] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, K. Cao, D. Liu, G. Wang, Q. Xu, X. Fang, S. Zhang, J. Xia, and J. Xia, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, vol. 296, Mar. 2020, Art. no. 200905.

[14] L. Wang and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," 2020, *arXiv:2003.09871*. [Online]. Available: http://arxiv.org/abs/2003.09871

[15] L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal," *BMJ*, vol. 369, no. 369, 2020. [Online]. Available: https://www.bmj.com/content/369/bmj.m1328

[16] J.-M. Kwon, K.-H. Kim, K.-H. Jeon, S. E. Lee, H.-Y. Lee, H.-J. Cho, J. O. Choi, E.-S. Jeon, M.-S. Kim, J.-J. Kim, K.-K. Hwang, S. C. Chae, S. H. Baek, S.-M. Kang, D.-J. Choi, B.-S. Yoo, K. H. Kim, H.-Y. Park, M.-C. Cho, and B.-H. Oh, "Artificial intelligence algorithm for predicting mortality of patients with acute heart failure," *PLoS ONE*, vol. 14, no. 7, Jul. 2019, Art. no. e0219302, doi: 10.1371/journal.pone.0219302.

[17] D. Toussie, N. Voutsinas, M. Finkelstein, M. A. Cedillo, S. Manna, S. Z. Maron, A. Jacobi, M. Chung, A. Bernheim, C. Eber, J. Concepcion, Z. Fayad, and Y. S. Gupta, "Clinical and chest radiography features determine patient outcomes in young and middle age adults with COVID-19," *Radiological*, vol. 297, no. 1, 2020, Art. no. 201754, doi: 10.1148/radiol.2020201754.

[18] M. Jamshidian and S. Jalal, "Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data," *Psychometrika*, vol. 75, no. 4, pp. 649–674, 2010.

[19] D. J. Stekhoven and P. Bühlmann, "MissForest-non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, Jan. 2012.

[20] M. Jamshidian, S. Jalal, and C. Jansen, "MissMech: An R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR)," *J. Stat. Softw.*, vol. 56, no. 6, pp. 1–31, 2014. [Online]. Available: https://escholarship.org/uc/item/51x8q0nn

[21] R. J. A. Little, "Missing-data adjustments in large surveys," *J. Bus. Econ. Statist.*, vol. 6, no. 3, pp. 287–296, Jul. 1988.

[22] J. Barnard, "Miscellanea. Small-sample degrees of freedom with multiple imputation," *Biometrika*, vol. 86, no. 4, pp. 948–955, Dec. 1999.

[23] S. V. Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.

[24] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010. [Online]. Available: http://www.jstatsoft.org/v36/i11/paper

[25] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta—A system for feature selection," *Fundamenta Informaticae*, vol. 101, no. 4, pp. 271–285, 2010.

[26] J.-M. Chang, H. Zeng, R. Han, Y.-M. Chang, R. Shah, C. M. Salafia, C. Newschaffer, R. K. Miller, P. Katzman, J. Moye, M. Fallin, C. K. Walker, and L. Croen, "Autism risk classification using placental chorionic surface vascular network features," *BMC Med. Informat. Decis. Making*, vol. 17, no. 1, p. 162, Dec. 2017, doi: 10.1186/s12911-017-0564-8.

[27] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[28] H. Deng, "Interpreting tree ensembles with in trees," *Int. J. Data Sci. Anal.*, vol. 7, no. 4, pp. 277–287, Jun. 2019.

[29] H. Deng, G. Runger, E. Tuv, and W. Bannister, "CBC: An associative classifier with a small number of rules," *Decis. Support Syst.*, vol. 59, pp. 163–170, Mar. 2014.

[30] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[31] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *J. Roy. Stat. Soc., A, Gen.*, vol. 135, no. 3, pp. 370–384, 1972.

[32] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinf.*, vol. 8, no. 1, p. 25, Dec. 2007, doi: 10.1186/1471-2105-8-25.

[33] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, May 2010.

[34] T. P. Morris, I. R. White, and P. Royston, "Tuning multiple imputation by predictive mean matching and local residual draws," *BMC Med. Res. Methodol.*, vol. 14, no. 1, p. 75, Dec. 2014, doi: 10.1186/1471-2288-14-75.

[35] S. Van Buuren, *Flexible Imputation of Missing Data*. Boca Raton, FL, USA: CRC Press, 2018. [Online]. Available: https://stefvanbuuren.name/fimd/

[36] P. Scheet and M. Stephens, "A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase," *Amer. J. Hum. Genet.*, vol. 78, no. 4, pp. 629–644, Apr. 2006.

[37] C. Dimauro, R. Steri, M. A. Pintus, G. Gaspa, and N. P. P. Macciotta, "Use of partial least squares regression to predict single nucleotide polymorphism marker genotypes when some animals are genotyped with a low-density panel," *Animal*, vol. 5, no. 6, pp. 833–837, Jun. 2011.

[38] C. Dimauro, M. Cellesi, G. Gaspa, P. Ajmone-Marsan, R. Steri, G. Marras, and N. P. Macciotta, "Use of partial least squares regression to impute SNP genotypes in italian cattle breeds," *Genet. Selection Evol.*, vol. 45, no. 1, p. 15, Dec. 2013.

[39] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, Jun. 2001.

[40] G. Kalton, *Compensating for Missing Survey Data* (Research Report Series). Ann Arbor, MI, USA: The Univ. of Michigan, Institute for Social Research, Survey Research Center, 1983.

[41] A. B. Owen and P. O. Perry, "Bi-cross-validation of the SVD and the nonnegative matrix factorization," *Ann. Appl. Statist.*, vol. 3, no. 2, pp. 564–594, Jun. 2009.

[42] L. Hunt and M. Jorgensen, "Mixture model clustering for mixed data with missing information," *Comput. Statist. Data Anal.*, vol. 41, nos. 3–4, pp. 429–440, Jan. 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167947302001901

[43] T. I. Lin, J. C. Lee, and H. J. Ho, "On fast supervised learning for normal mixture models with missing information," *Pattern Recognit.*, vol. 39, no. 6, pp. 1177–1187, Jun. 2006.

[44] R. J. Steele, N. Wang, and A. E. Raftery, "Inference from multiple imputation for missing data using mixtures of normals," *Stat. Methodol.*, vol. 7, no. 3, pp. 351–365, May 2010.

[45] L. Cappelletti, T. Fontana, G. W. Di Donato, L. Di Tucci, E. Casiraghi, and G. Valentini, "Complex data imputation by auto-encoders and convolutional neural networks—A case study on genome gap-filling," *Computers*, vol. 9, no. 2, p. 37, May 2020, doi: 10.3390/computers9020037.

[46] T. Marwala and S. Chakraverty, "Fault classification in structures with incomplete measured data using autoassociative neural networks and genetic algorithm," *Current Sci.*, vol. 90, pp. 542–548, 2006.

[47] W. Qiao, Z. Gao, R. G. Harley, and G. K. Venayagamoorthy, "Robust neuro-identification of nonlinear plants in electric power systems with missing sensor measurements," *Eng. Appl. Artif. Intell.*, vol. 21, no. 4, pp. 604–618, Jun. 2008.

[48] V. Miranda, J. Krstulovic, H. Keko, C. Moreira, and J. Pereira, "Reconstructing missing data in state estimation with autoencoders," *IEEE Trans. Power Syst.*, vol. 27, no. 2, pp. 604–611, May 2012.

[49] S. J. Choudhury and N. R. Pal, "Imputation of missing data with neural networks for classification," *Knowl.-Based Syst.*, vol. 182, Oct. 2019, Art. no. 104838.

[50] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[51] Y. Zhuang, R. Ke, and Y. Wang, "Innovative method for traffic data imputation based on convolutional neural network," *IET Intell. Transp. Syst.*, vol. 13, no. 4, pp. 605–613, Apr. 2019.

[52] J. Yoon, J. Jordon, and M. van der Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proc. 35th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 80, J. Dy and A. Krause, Eds. Stockholm, Sweden: Stockholmsmässan, Jul. 2018, pp. 5689–5698. [Online]. Available: http://proceedings.mlr.press/v80/yoon18a.html

[53] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ, USA: Wiley, 2004.

[54] D. B. Rubin, "Formalizing subjective notions about the effect of nonrespondents in sample surveys," *J. Amer. Stat. Assoc.*, vol. 72, no. 359, pp. 538–543, Sep. 1977.

[55] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Statist. Med.*, vol. 30, no. 4, pp. 377–399, Feb. 2011.

[56] P. Zhang, "Multiple imputation: Theory and method," *Int. Stat. Rev.*, vol. 71, no. 3, pp. 581–592, Jan. 2007. [Online]. Available: https://projecteuclid.org:443/euclid.isr/1066768709

[57] D. Sovilj, E. Eirola, Y. Miche, K.-M. Björk, R. Nian, A. Akusok, and A. Lendasse, "Extreme learning machine for missing data using multiple imputations," *Neurocomputing*, vol. 174, pp. 220–231, Jan. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231215011182

[58] K. C. Santosh, "AI-driven tools for coronavirus outbreak: Need of active learning and cross-population train/test models on multitudinal/multimodal data," *J. Med. Syst.*, vol. 44, no. 5, pp. 1–5, May 2020.

[59] G. Webb and C. Sammut, Eds., *Encyclopedia of Machine Learning and Data Mining*. Boston, MA, USA: Springer, 2017.

[60] M.-R. Bouguelia, S. Nowaczyk, K. C. Santosh, and A. Verikas, "Agreeing to disagree: Active learning with noisy labels without crowdsourcing," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 8, pp. 1307–1319, Aug. 2018.

[61] J. Yang, G. Sharp, H. Veeraraghavan, W. van Elmpt, A. Dekker, T. Lustberg, and M. Gooding, "Data from lung CT segmentation challenge," *Cancer Imag. Arch.*, 2017. [Online]. Available: https://wiki.cancerimagingarchive.net/display/Public/Lung+CT+Segmentation+Challenge+2017, doi: 10.7937/K9/TCIA.2017.3r3fvz08.

[62] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 590–597.

[63] J. Ma, Y. Song, X. Tian, Y. Hua, R. Zhang, and J. Wu, "Survey on deep learning for pulmonary medical imaging," *Frontiers Med.*, vol. 14, pp. 1–20, Dec. 2019.

[64] K. K. Bressem, L. Adams, C. Erxleben, B. Hamm, S. Niehues, and J. Vahldiek, "Comparing different deep learning architectures for classification of chest radiographs," 2020, *arXiv:2002.08991*. [Online]. Available: http://arxiv.org/abs/2002.08991

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[66] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," 2020, *arXiv:2003.10849*. [Online]. Available: http://arxiv.org/abs/2003.10849

[67] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*. [Online]. Available: http://arxiv.org/abs/1602.07261

[68] D. Das, K. C. Santosh, and U. Pal, "Truncated inception net: COVID-19 outbreak screening using chest X-rays," *Phys. Eng. Sci. Med.*, vol. 43, no. 3, pp. 915–925, Sep. 2020.

[69] I. D. Apostolopoulos and T. A. Mpesiana, "COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, Jun. 2020.

[70] K. H. Shibly, S. K. Dey, M. T.-U. Islam, and M. M. Rahman, "COVID faster R-CNN: A novel framework to diagnose novel coronavirus disease (COVID-19) in X-ray images," *Informat. Med. Unlocked*, vol. 20, 2020, Art. no. 100405. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352914820305554

[71] M. D. Li, N. T. Arun, M. Gidwani, K. Chang, F. Deng, B. P. Little, D. P. Mendoza, M. Lang, S. I. Lee, A. O'Shea, A. Parakh, P. Singh, and J. Kalpathy-Cramer, "Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks," *Radiol., Artif. Intell.*, vol. 2, no. 4, Jul. 2020, Art. no. e200079.

[72] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200417302385

[73] J. M. Zurada, A. Malinowski, and I. Cloete, "Sensitivity analysis for minimization of input data dimension for feedforward neural network," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, vol. 6, May/Jun. 1994, pp. 447–450.

[74] A. H. Sung, "Ranking importance of input parameters of neural networks," *Expert Syst. Appl.*, vol. 15, nos. 3–4, pp. 405–411, Oct. 1998.

[75] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[76] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[77] Z. Feng *et al.*, "Early prediction of disease progression in COVID-19 pneumonia patients with chest CT and clinical characteristics," *Nature Commun.*, vol. 11, no. 1, p. 4968, Dec. 2020.

[78] P. K. Sethy and S. K. Behera, "Detection of coronavirus disease (COVID-19) based on deep features," 2020, doi: 10.20944/preprints202003.0300.v1.

[79] A. Pervaiz, U. Pasha, S. Bashir, R. Arshad, M. Waseem, and O. Qasim, "Neutrophil to lymphocyte ratio (NLR) can be a predictor of the outcome and the need for mechanical ventilation in patients with COVID-19 in Pakistan," *Pakistan J. Pathol.*, vol. 31, no. 2, pp. 38–41, 2020.

[80] H. Yildiz, J. C. Yombi, and D. Castanares-Zapatero, "Validation of a risk score to predict patients at risk of critical illness with COVID-19," *Infectious Diseases*, pp. 1–3, Oct. 2020.

[81] S. Schalekamp, M. Huisman, R. A. van Dijk, M. F. Boomsma, P. J. Freire Jorge, W. S. de Boer, G. J. M. Herder, M. Bonarius, O. A. Groot, E. Jong, A. Schreuder, and C. M. Schaefer-Prokop, "Model-based prediction of critical illness in hospitalized patients with COVID-19," *Radiology*, to be published, doi: 10.1148/radiol.2020202723.

[82] X. Bai, C. Fang, Y. Zhou, S. Bai, Z. Liu, Q. Chen, Y. Xu, T. Xia, S. Gong, X. Xie, D. Song, R. Du, C. Zhou, C. Chen, D. Nie, D. Tu, C. Zhang, X. Liu, L. Qin, and W. Chen, "Predicting COVID-19 malignant progression with Ai techniques," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.03.20.20037325v3

[83] F. Caramelo, N. Ferreira, and B. Oliveiros, "Estimation of risk factors for COVID-19 mortality-prelim-i-nary results," *medRxiv*, 2020, doi: 10.1101/2020.02.24.20027268.

[84] J. Xie, D. Hungerford, H. Chen, S. Abrams, S. Li, G. Wang, Y. Wang, H. Kang, L. Bonnett, R. Zheng, X. Li, Z. Tong, B. Du, H. Qiu, and C.-H. Toh, "Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19," *medRxiv*, 2020, doi: 10.1101/2020.03.28.20045997.

[85] X. Qi, Z. Jiang, Q. Yu, C. Shao, H. Zhang, H. Yue, B. Ma, Y. Wang, C. Liu, X. Meng, and S. Huang, "Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.02.29.20029603v1

[86] H. Huang, S. Cai, Y. Li, Y. Li, Y. Fan, L. Li, C. Lei, X. Tang, F. Hu, F. Li, and X. Deng, "Prognostic factors for COVID-19 pneumonia progression to severe symptoms based on earlier clinical features: A retrospective analysis," *Frontiers Med.*, vol. 7, p. 643, Oct. 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmed.2020.557453

[87] M. Pourhomayoun and M. Shakibi, "Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making," *medRxiv*, 2020. [Online]. Available: https://europepmc.org/article/ppr/ppr137985

[88] O. Y. Bello-Chavolla, J. P. Bahena-López, N. E. Antonio-Villa, A. Vargas-Vázquez, A. González-Díaz, A. Márquez-Salinas, C. A. Fermín-Martínez, J. J. Naveja, and C. A. Aguilar-Salinas, "Predicting mortality due to SARS-CoV-2: A mechanistic score relating obesity and diabetes to COVID-19 outcomes in mexico," *J. Clin. Endocrinol. Metabolism*, vol. 105, no. 8, pp. 2752–2761, Aug. 2020.

[89] E. Carr *et al.*, "Supplementing the national early warning score (news2) for anticipating early deterioration among patients with COVID-19 infection," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.04.24.20078006v4

[90] D. Colombi, F. C. Bodini, M. Petrini, G. Maffi, N. Morelli, G. Milanese, M. Silva, N. Sverzellati, and E. Michieletti, "Well-aerated lung on admitting chest CT to predict adverse outcome in COVID-19 pneumonia," *Radiology*, vol. 296, no. 2, pp. E86–E96, Aug. 2020.

[91] A. K. Das, S. Mishra, and S. S. Gopalan, "Predicting COVID-19 community mortality risk using machine learning and development of an online prognostic tool," *PeerJ*, vol. 8, Sep. 2020, Art. no. e10083.

[92] X. Chen and Z. Liu, "Early prediction of mortality risk among severe COVID-19 patients using machine learning," *medRxiv*, 2020, doi: 10.1101/2020.04.13.20064329.

[93] Q. Liu, X. Fang, S. Tokuno, U. Chung, X. Chen, X. Dai, X. Liu, F. Xu, B. Wang, and P. Peng, "Prediction of the clinical outcome of COVID-19 patients using T lymphocyte subsets with 340 cases from Wuhan, China: A retrospective cohort study and a Web visualization tool," 2020, doi: 10.2139/ssrn.3557995.

[94] M. P. McRae, G. W. Simmons, N. J. Christodoulides, Z. Lu, S. K. Kang, D. Fenyo, T. Alcorn, I. P. Dapkins, I. Sharif, D. Vurmaz, S. S. Modak, K. Srinivasan, S. Warhadpande, R. Shrivastav, and J. T. McDevitt, "Clinical decision support tool and rapid point-of-care platform for determining disease severity in patients with COVID-19," *Lab Chip*, vol. 20, no. 12, pp. 2075–2085, 2020.

[95] C. V. Guillamet, R. V. Guillamet, A. A. Kramer, P. M. Maurer, G. A. Menke, C. L. Hill, and W. A. Knaus, "Toward a COVID-19 score-risk assessments and registry," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.04.15.20066860v1

[96] H. Zhang, T. Shi, X. Wu, X. Zhang, K. Wang, D. Bean, R. Dobson, J. T. Teo, J. Sun, P. Zhao, C. Li, K. Dhaliwal, H. Wu, Q. Li, and B. Guthrie, "Risk prediction for poor outcome and death in hospital in-patients with COVID-19: Derivation in Wuhan, China and external validation in London, UK," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/05/03/2020.04.28.20082222

[97] J. Gong, J. Ou, X. Qiu, Y. Jie, Y. Chen, L. Yuan, J. Cao, M. Tan, W. Xu, F. Zheng, and Y. Shi, "A tool to early predict severe corona virus disease 2019 (COVID-19): A multicenter study using the risk nomogram in Wuhan and Guangdong, China," *Clin. Infectious Diseases*, vol. 71, pp. 833–840, Apr. 2020.

[98] C. Iwendi, A. K. Bashir, A. Peshkar, R. Sujatha, J. M. Chatterjee, S. Pasupuleti, R. Mishra, S. Pillai, and O. Jo, "COVID-19 patient health prediction using boosted random forest algorithm," *Frontiers Public Health*, vol. 8, p. 357, Jul. 2020.

[99] J. Sarkar and P. Chakrabarti, "A machine learning model reveals older age and delayed hospitalization as predictors of mortality in patients with COVID-19," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.03.25.20043331v1

[100] G. Chassagnon *et al.*, "AI-driven CT-based quantification, staging and short-term outcome prediction of COVID-19 pneumonia," 2020, *arXiv:2004.12852*. [Online]. Available: http://arxiv.org/abs/2004.12852

[101] X. Jiang, M. Coffee, A. Bari, J. Wang, X. Jiang, J. Huang, J. Shi, J. Dai, J. Cai, T. Zhang, and Z. Wu, "Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity," *Comput., Mater. Continua*, vol. 63, pp. 537–551, May 2020.

[102] L. Yan, H.-T. Zhang, Y. Xiao, M. Wang, C. Sun, J. Liang, S. Li, M. Zhang, Y. Guo, Y. Xiao, and X. Tang, "Prediction of criticality in patients with severe COVID-19 infection using three clinical features: A machine learning-based prognostic model with clinical data in Wuhan," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.02.27.20028027v2

[103] A. Vaid *et al.*, "Machine learning to predict mortality and critical events in COVID-19 positive new york city patients: A cohort study (preprint)," *J. Med. Internet Res.*, to be published, doi: 10.2196/24018.

[104] J. Lu *et al.*, "ACP risk grade: A simple mortality index for patients with confirmed or suspected severe acute respiratory syndrome coronavirus 2 disease (COVID-19) during the early stage of outbreak in Wuhan, China," 2020, doi: 10.2139/ssrn.3543603.

[105] D. Ji, D. Zhang, J. Xu, Z. Chen, T. Yang, P. Zhao, G. Chen, G. Cheng, Y. Wang, J. Bi, L. Tan, G. Lau, and E. Qin, "Prediction for progression risk in patients with COVID-19 pneumonia: The CALL score," *Clin. Infectious Diseases*, vol. 71, no. 6, pp. 1393–1399, Sep. 2020.

[106] H. Al-Najjar and N. Al-Rousan, "A classifier prediction model to predict the status of coronavirus COVID-19 patients in South Korea," *Eur. Rev. Med. Pharmacol. Sci.*, vol. 24, no. 6, pp. 3400–3403, 2020.

[107] K. Hajifathalian, R. Z. Sharaiha, S. Kumar, T. Krisko, D. Skaf, B. Ang, W. D. Redd, J. C. Zhou, K. E. Hathorn, T. R. McCarty, A. N. Bazarbashi, C. Njie, D. Wong, L. Shen, E. Sholle, D. E. Cohen, R. S. Brown, W. W. Chan, and B. E. Fortune, "Development and external validation of a prediction risk model for short-term mortality among hospitalized U.S. COVID-19 patients: A proposal for the COVID-AID risk tool," *PLoS ONE*, vol. 15, no. 9, Sep. 2020, Art. no. e0239536.

[108] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Comput. Biol. Med.*, vol. 121, Jun. 2020, Art. no. 103792.

[109] H. Y. F. Wong, H. Y. S. Lam, A. H.-T. Fong, S. T. Leung, T. W.-Y. Chin, C. S. Y. Lo, M. M.-S. Lui, J. C. Y. Lee, K. W.-H. Chiu, T. W.-H. Chung, E. Y. P. Lee, E. Y. F. Wan, I. F. N. Hung, T. P. W. Lam, M. D. Kuo, and M.-Y. Ng, "Frequency and distribution of chest radiographic findings in patients positive for COVID-19," *Radiology*, vol. 296, no. 2, pp. E72–E78, Aug. 2020, doi: 10.1148/radiol.2020201160.

[110] S. S. Hare, J. Rodrigues, A. Nair, and G. Robinson, "Lessons from the frontline of the COVID-19 outbreak," *BMJ Opinion*. Accessed: Oct. 3, 2020. [Online]. Available: https://blogs.bmj.com/bmj/2020/03/20/lessons-from-the-frontline-of-the- covid-19-outbreak

[111] Q. Yang, Q. Liu, H. Xu, H. Lu, S. Liu, and H. Li, "Imaging of coronavirus disease 2019: A chinese expert consensus statement," *Eur. J. Radiol.*, vol. 127, Jun. 2020, Art. no. 109008.

[112] L. I.-K. Lin, "Assay validation using the concordance correlation coefficient," *Biometrics*, vol. 48, no. 2, pp. 599–604, 1992.

[113] A. Rizzi, C. Gatta, and D. Marini, "A new algorithm for unsupervised global and local color correction," *Pattern Recognit. Lett.*, vol. 24, no. 11, pp. 1663–1677, Jul. 2003.

[114] A. Rizzi and J. J. McCann, "On the behavior of spatial models of color," *Proc. SPIE*, vol. 6493, pp. 11–24, Jan. 2007.

[115] G. Schaefer, M. I. Rajab, M. E. Celebi, and H. Iyatomi, "Colour and contrast enhancement for improved skin lesion segmentation," *Comput. Med. Imag. Graph.*, vol. 35, no. 2, pp. 99–104, Mar. 2011.

[116] K.-B. Park, S. H. Choi, and J. Y. Lee, "M-GAN: Retinal blood vessel segmentation by balancing losses through stacked deep fully convolutional networks," *IEEE Access*, vol. 8, pp. 146308–146322, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9162010

[117] R. J. A. Little, "A test of missing completely at random for multivariate data with missing values," *J. Amer. Stat. Assoc.*, vol. 83, no. 404, pp. 1198–1202, Dec. 1988.

[118] D. M. Hawkins, "A new test for multivariate normality and homoscedasticity," *Technometrics*, vol. 23, no. 1, pp. 105–110, Feb. 1981.

[119] M. S. Srivastava and M. Dolatabadi, "Multiple imputation and other resampling schemes for imputing missing observations," *J. Multivariate Anal.*, vol. 100, no. 9, pp. 1919–1937, 2009.

[120] K.-H. Yuan, "Normal distribution based pseudo ML for missing data: With applications to mean and covariance structure analysis," *J. Multivariate Anal.*, vol. 100, no. 9, pp. 1900–1918, Oct. 2009.

[121] S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction," *BMC Med. Res. Methodol.*, vol. 20, no. 1, pp. 1–12, Jul. 2020, doi: 10.1186/s12874-020-01080-1.

[122] I. Eekhout, M. A. van de Wiel, and M. W. Heymans, "Methods for significance testing of categorical covariates in logistic regression models after multiple imputation: Power and applicability analysis," *BMC Med. Res. Methodol.*, vol. 17, no. 1, Dec. 2017, Art. no. 129, doi: 10.1186/s12874-017-0404-7.

[123] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. New York, NY, USA: Chapman & Hall, 1997.

[124] T. D. Pigott, "A review of methods for missing data," *Educ. Res. Eval.*, vol. 7, no. 4, pp. 353–383, 2001.

[125] P. Royston, "Multiple imputation of missing values," *Stata J.*, vol. 4, no. 3, pp. 227–241, 2004.

[126] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer, "High-dimensional variable selection for survival data," *J. Amer. Stat. Assoc.*, vol. 105, no. 489, pp. 205–217, 2010.

[127] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, "Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study," *Amer. J. Epidemiol.*, vol. 179, no. 6, pp. 764–774, Mar. 2014.

[128] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 10, pp. 6562–6566, May 2002.

[129] J. Fox, *Applied Regression Analysis and Generalized Linear Models*. Thousand Oaks, CA, USA: Sage, 2015.

[130] M. Schubach, M. Re, P. N. Robinson, and G. Valentini, "Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants," *Sci. Rep.*, vol. 7, no. 1, Dec. 2017, Art. no. 2959, doi: 10.1038/s41598-017-03011-5.

[131] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in *Proc. Int. Conf. Knowl. Discovery Data Mining*, 1998, pp. 80–86.

[132] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, 2nd ed. Berlin, Germany: Springer, 2009.

[133] B. Grund and C. Sabin, "Analysis of biomarker data: Logs, odds ratios, and receiver operating characteristic curves," *Current Opinion HIV AIDS*, vol. 5, no. 6, pp. 473–479, Nov. 2010.

[134] J. K. Blitzstein and J. Hwang, *Introduction to Probability*. Boca Raton, FL, USA: CRC Press, 2019.

**ELENA CASIRAGHI** (Member, IEEE) received the M.Sc. degree in computing and the Ph.D. degree in computer Science from the Università degli Studi di Milano, Italy, in 2001 and 2005, respectively.

In 2000, she started her research activity with the Information Technology Department, Valtion Teknillinen Tutkimuskeskus (VTT), Helsinki, Finland. In 2005, she was hired as an Assistant Professor with the Università degli Studi di Milano, where she became an Associate Professor, in March 2020. She also investigated novel learning algorithms for pattern recognition, manifold learning, and intrinsic dimensionality estimation to develop novel theories and automatic algorithms dealing with high-dimensional datasets characterized by a small cardinality (small sample size problem). Her cooperation with several Italian and foreign hospitals has resulted in more than 50 papers in international journals and conferences. Her research interests include artificial intelligence to develop automatic systems for medical and biomedical image processing and pattern recognition.

**DARIO MALCHIODI** received the M.Sc. degree in computing and the Ph.D. degree in computational mathematics and operations research from the Università degli Studi di Milano, Italy, in 1997 and 2000, respectively.

Since 2002, he has been an Assistant Professor with the Department of Computer Science, Università degli Studi di Milano, where he was appointed as an Associate Professor, in 2011. He teaches statistics and data analysis and algorithms for massive datasets. He is the author of about 100 scientific publications. He is also actively involved in the popularization of computing. His research interests include the treatment of uncertainty in machine learning, with a particular focus to data-driven induction of fuzzy sets, compression of machine learning models, mining of knowledge bases in semantic web, negative example selection in bioinformatics, and application of machine learning to the medical, veterinary, and forensics fields.

**GABRIELLA TRUCCO** received the M.Sc. and Ph.D. degrees in computer science from the Università degli Studi di Milano, Italy, in 2002 and 2005, respectively.

Since 2006, she has been an Assistant Professor with the Università degli Studi di Milano. She is the author of about 40 papers in international conferences and journals. Her research interests include bioinformatics and the development of algorithms for phylogenetic analysis, the analysis of noncoding DNA, and the application of machine learning techniques.

**MARCO FRASCA** received the Ph.D. degree in computer science from the Università degli Studi di Milano, Italy, in 2012.

He was a Postdoctoral Researcher with the Department of Biosciences and the Department of Computer Science, Università degli Studi di Milano. Since 2017, he has been an Assistant Professor with the Department of Computer Science, Università degli Studi di Milano. He is a member of AnacletoLab, whose research activities regard the field of machine learning applied in biology and medicine, with numerous collaborations with international research groups, including the Institute for Medical and Human Genetics Charité, Berlin; The Jackson Laboratory for Genomic Medicine, Farmington, USA; and the Department of Computer Science, Royal Holloway University of London. He has been an Invited Research Visitor at several universities, including the Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, and the Institute of Molecular Biology, Johannes Gutenberg University of Mainz. He contributed to consolidate the application of Hopfield networks to classification and ranking problems with the development of single- and multi-task parametric Hopfield models. His research interest includes the design and analysis of new machine learning methods, with applications in bioinformatics, computational biology, and medicine.

**LUCA CAPPELLETTI** received the B.Sc. degree in computer engineering from the Polytechnic University of Milan, Italy, in 2017, and the M.Sc. degree in computer science from the Università degli Studi di Milano, Italy, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Computer Science.

The main application of his research is on heterogeneous data within bioinformatics. His main research interests include graph embedding and multitask machine learning.

**TOMMASO FONTANA** received the B.Sc. degree in computer engineering from the Polytechnic University of Milan, Italy, where he is currently pursuing the master's degree in computer science and engineering.

He is a part of the mHackeroni Computer Security Team. His main research interests include graph embedding and machine learning.

**ALESSANDRO ANDREA ESPOSITO** received the Medical degree in surgery and medicine and a specialization in radiology from the Università degli Studi di Milano, Italy, in 1999 and 2003, respectively.

He currently works as a Consultant of Radiology with the Fondazione IRCCS Ca Granda Ospedale Maggiore Policlinico, Milan, Italy. He is also a Teacher of radiation protection in the courses organized by hospitals to employees. He is also responsible for the quality and satisfaction service of the Radiology Department. He is also a Tutor for medical students. His research interests include emergency radiology, computed tomography (neck, chest, and abdomen), ultrasonography, magnetic resonance imaging, conventional radiology, and musculoskeletal imaging.

Dr. Esposito has been a member of the Italian Society of Radiology (SIRM), since 2000.

**EMANUELE AVOLA** received the degree in medicine and surgery from the Sapienza Università di Rome, Italy, in 2016.

He is currently a Second-Year Resident, attending the Postgraduate School in Radiodiagnostics, Università degli Studi di Milano, Italy. His research interests include toracic and abdominal radiology, artificial intelligence, and radiomics.

**ALESSANDRO JACHETTI** was born in Milan. He received the Medical degree from The University of Pavia, Italy, in 2011, and a specialization in emergency medicine from the Università degli Studi di Parma, Italy, in 2017.

He actually works as an Emergency Physician with the Milano University Hospital, Italy. He had humanitarian experience with doctors without borders, prior to focus on research and teaching. His primary research interests include emergency medicine, decision-making processes, and new technologies.

**JUSTIN REESE** received the Ph.D. degree in computational biology from the University of Virginia, USA, in 2004.

He has been a Computational Biologist with the Lawrence Berkeley National Laboratory, since 2019. His research interests include integration of heterogeneous data using knowledge graphs and machine learning techniques for extracting actionable information from biomedical data.

**ALESSANDRO RIZZI** (Member, IEEE) received the degree in information science from the Università degli Studi di Milano, Italy, in 1992, and the Ph.D. degree in information engineering from Università degli Studi di Brescia, Italy, in 1999.

Since 1990, he has been doing research in the field of digital imaging with a particular interest on color, visualization, photography, HDR, VR, and on the perceptual issues related to digital imaging, interfaces, and lighting. Since 2015, he has been a Full Professor with the Department of Computer Science, Università degli Studi di Milano, where he is currently teaching multimedia, colorimetry, and film restoration. He is also the Head of the MIPS Laboratory, Department of Computer Science. He is the author of about 400 scientific works.

Dr. Rizzi has been a Fellow of the Society for Imaging Science and Technology (IS&T). He is a member of several program committees of conferences related to color and digital imaging. In 2015, he received the Davies Medal from the Royal Photographic Society. He has been one of the founders of the Italian Color Group, the Secretary of the CIE Division 8, and the Vice President. He is the Co-Chair of the IS&T Conference Color Imaging: Displaying, Processing, Hardcopy and Applications, the Topical Editor of Applied Color Science of the *Journal of the Optical Society of America A*, and an Associate Editor of the *Journal of Electronic Imaging*.

**PETER N. ROBINSON** received the degree in mathematics and computer science from Columbia University, USA, and the degree in medicine from the University of Pennsylvania, USA.

He completed his training as a Pediatrician with the Charité University Hospital, Berlin, Germany. He has been a Full Professor of Computational Biology with The Jackson Laboratory for Genomic Medicine, since 2016. His group developed the Human Phenotype Ontology (HPO), which is currently an international standard for computation over human disease that is used by the Sanger Institute, several NIH-funded groups, including the Undiagnosed Diseases Program, Genome Canada, the rare diseases section of the U.K.'s 100 000 Genomes Project, and many others. His group develops algorithms and software for the analysis of exome and genome sequences and has used whole-exome sequencing and other methods to identify a number of novel disease genes, including CA8, PIGV, PIGO, PGAP3, IL-21R, PIGT, and PGAP2.

**GIORGIO VALENTINI** received the degrees in biology and computer science and the Ph.D. degree in computer science from the University of Genoa, Italy, in 1981, 1999, and 2003, respectively.

Since 2019, he has been a Full Professor with the Department of Computer Science, Università degli Studi di Milano, Italy, where he is currently teaching bioinformatics and machine learning methods for precision medicine. He is also the Director of AnacletoLab and the Computational Biology and Bioinformatics Laboratory, Department of Computer Science. He is the author of over 150 scientific publications with peer-review in journals, book chapters, and international conferences in the field of machine learning, bioinformatics, and computational biology. His main research interests include the development and application of artificial intelligence methods to bio-medical problems, with a special focus on machine learning methods for personalized and precision medicine, network medicine, and systems biology. He is a member of the editorial board of the *Scientific Reports*, *Computers*, and other bioinformatics journals.