IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 September 2023, accepted 22 October 2023, date of publication 26 October 2023, date of current version 3 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3327808

RESEARCH ARTICLE

Explainable Machine-Learning Models for COVID-19 Prognosis Prediction Using Clinical, Laboratory and Radiomic Features

FRANCESCO PRINZI[1], CARMELO MILITELLO[2], NICOLA SCICHILONE[3], SALVATORE GAGLIO[2,4], (Life Member, IEEE), AND SALVATORE VITABILE[1]

[1]Department of Biomedicine, Neuroscience and Advanced Diagnostics (BiND), University of Palermo, 90127 Palermo, Italy
[2]Institute for High-Performance Computing and Networking (ICAR-CNR), Italian National Research Council, 90146 Palermo, Italy
[3]Division of Respiratory Diseases, Department of Health Promotion Sciences, Maternal and Infant Care, Internal Medicine and Medical Specialties (PROMISE), University of Palermo, 90127 Palermo, Italy
[4]Department of Engineering, University of Palermo, 90128 Palermo, Italy

Corresponding author: Carmelo Militello (carmelo.militello@cnr.it)

This work was partially funded by the European Union – NextGenerationEU – MUR D.M. funds 737/2021 – research project "A Trustworthy Clinical Decision Support System for non-communicable diseases detection and prediction."

ABSTRACT The SARS-CoV-2 virus pandemic had devastating effects on various aspects of life: clinical cases, ranging from mild to severe, can lead to lung failure and to death. Due to the high incidence, data-driven models can support physicians in patient management. The explainability and interpretability of machine-learning models are mandatory in clinical scenarios. In this work, clinical, laboratory and radiomic features were used to train machine-learning models for COVID-19 prognosis prediction. Using Explainable AI algorithms, a multi-level explainable method was proposed taking into account the developer and the involved stakeholder (physician, and patient) perspectives. A total of 1023 radiomic features were extracted from 1589 Chest X-Ray images (CXR), combined with 38 clinical/laboratory features. After the pre-processing and selection phases, 40 CXR radiomic features and 23 clinical/laboratory features were used to train Support Vector Machine and Random Forest classifiers exploring three feature selection strategies. The combination of both radiomic, and clinical/laboratory features enabled higher performance in the resulting models. The intelligibility of the used features allowed us to validate the models' clinical findings. According to the medical literature, LDH, PaO2 and CRP were the most predictive laboratory features. Instead, ZoneEntropy and HighGrayLevelZoneEmphasis - indicative of the heterogeneity/uniformity of lung texture - were the most discriminating radiomic features. Our best predictive model, exploiting the Random Forest classifier and a signature composed of clinical, laboratory and radiomic features, achieved AUC=0.819, accuracy=0.733, specificity=0.705, and sensitivity=0.761 in the test set. The model, including a multi-level explainability, allows us to make strong clinical assumptions, confirmed by the literature insights.

INDEX TERMS Chest X-ray images, clinical and laboratory features, COVID-19 prognosis, explainable AI, machine learning classifiers, predictive models, radiomic features.

I. INTRODUCTION

The worldwide diffusion of the SARS-CoV-2 virus had devastating effects on various aspects of life, including the economy, society, and public health. Despite the spread of non-invasive variants and the availability of vaccines have reduced mortality rates, early prediction of health-threatening symptoms still remains a critical task [6], [9], [26].

While chest CT scans have demonstrated high sensitivity in detecting COVID-19 [1], [29], CXRs have proven to

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

be a more sustainable and efficient means of handling the high volume of daily cases [24]. Furthermore, when CXR images are combined with clinical and laboratory features, their prognostic performance increases [4], [52], [64]. In fact, several works have proposed Machine Learning (ML) models to improve the COVID-19 prognosis prediction process. However, model accuracy optimization is not the only requirement. In critical contexts, such as in clinical settings, it is also essential to ensure the explainability of the trained models. The models have to be validated technically by engineers to enhance their robustness and reliability, and clinically by physicians to verify any existing clinical evidence and confirm their effectiveness. One of the main methods of inferring model explainability is through the use of inherently interpretable inputs. While clinical and laboratory features are intelligible (i.e., understandable by humans), imaging features can be uninterpretable depending on their extraction process. For example, despite several methods being proposed to explain the features extracted *via* deep neural networks, their nature is inherently unintelligible. These methods are focused on saliency map computation to highlight the areas that most influence the model decision [34], [35]. As an illustration, the Grad-CAM method demonstrated a limitation in differentiating multifocal lesions, as evidenced by recent studies [40], [56]. Furthermore, it has been shown that different methods can produce conflicting results, as reported in a recent study [69]. In addition, these methods enable just a local explanation for a specific instance (i.e., patient), not allowing a global validation of the systems. For these reasons, saliency maps have still to demonstrate to be an objective tool for validating clinical findings.

In recent years, especially in the radiology field, Radiomics has emerged as a very powerful tool for feature extraction. Radiomics aims to extract highly informative quantitative features from radiological images. The extracted radiomic features provide a complementary point of view to the qualitative analysis performed by the radiologist. In fact, through the use of radiomic features, it is possible to train data-driven models to predict a clinical outcome. It has been shown that doctors' diagnostic performance improves when supported by quantitative features [15], [27]. The main advantage of radiomic features lies in their inherent interpretability: it is well known the meaning each feature expresses. In recent years, there has been a notable surge in radiomic research endeavours geared towards enhancing diagnostic capabilities. However, these efforts are often hampered by deficiencies in terms of repeatability and explainability. To guarantee the repeatability of radiomic analyses, it is imperative to meticulously address each facet of the workflow, encompassing image acquisition, image reconstruction, segmentation, feature definition, feature extraction, feature selection, and model setup [7], [50]. In addition, many studies in the literature, provide only an informative radiomic signature without exploring in depth any clinical explanation or interpretation.

Intelligible inputs are the first step toward an interpretable model. However, the ML algorithms have also to be explained. Although Support Vector Machine (SVM) and Tree Ensemble are defined as black boxes, many techniques were developed [18], [31] for their *post-hoc* explanation [32]. These techniques provide global and local explanations, enabling the findings introspection from the healthcare process stakeholders (clinicians, technicians, nurses, general practitioners, healthcare makers, and patients) [10].

In this work the conducted analysis aimed at defining predictive models for COVID-19 prognosis prediction. Utilizing clinical, laboratory, and radiomic features as inputs, we implemented two distinct machine learning classifiers, namely Support Vector Machine (SVM) and Random Forest (RF), alongside employing several feature selection strategies. First of all, unimodal models were evaluated, exploiting only clinical and laboratory data, and only CXR radiomic features. Successively, multimodal models were considered, combining both clinical and CXR radiomics. The above-listed ML algorithms and the use of intrinsically interpretable features, allowed us to propose the *multi-level explanation*, depicted in Figure 1. In particular, the global explanation is used for model introspection to assess the significance of individual features, identify phenomena such as distributional drift, and validate any pre-existing clinically proven evidence. The local explanation is used to explain the predictions for each patient. Intrinsically explainable inputs combined with global and local explanations lay the basis for the development of an eXplainable Clinical Decision Support System (X-CDSS).

The main contributions of this work are:

- an in-depth analysis of two ML classifiers (i.e., Support Vector Machine and Random Forest) to define predictive models for the prognosis (i.e., MILD *vs.* SEVERE) of COVID-19 patients;
- the implementation of different feature selection strategies for the identification of the optimal signature composed of radiomic and clinical/laboratory features;
- a multi-level explainability taking into account the developer, physician, and patient perspectives, assessing the role of each feature and quantifying their contribution to the final decision.

The remaining of the paper is organized as follows: Section II analyzes the literature works to support the clinical decision in the management of COVID-19 disease; Section III describes the conducted study, detailing each step of the processing pipeline, to set up the ML predictive models; Section IV illustrates the obtained experimental results, concerning each specific step and the performance of the built-up models, both in the training/validation phase and in the testing phase; Section V discuss the obtained results, highlighting the clinical viewpoint of the findings; finally, conclusions are provided in Section VI.

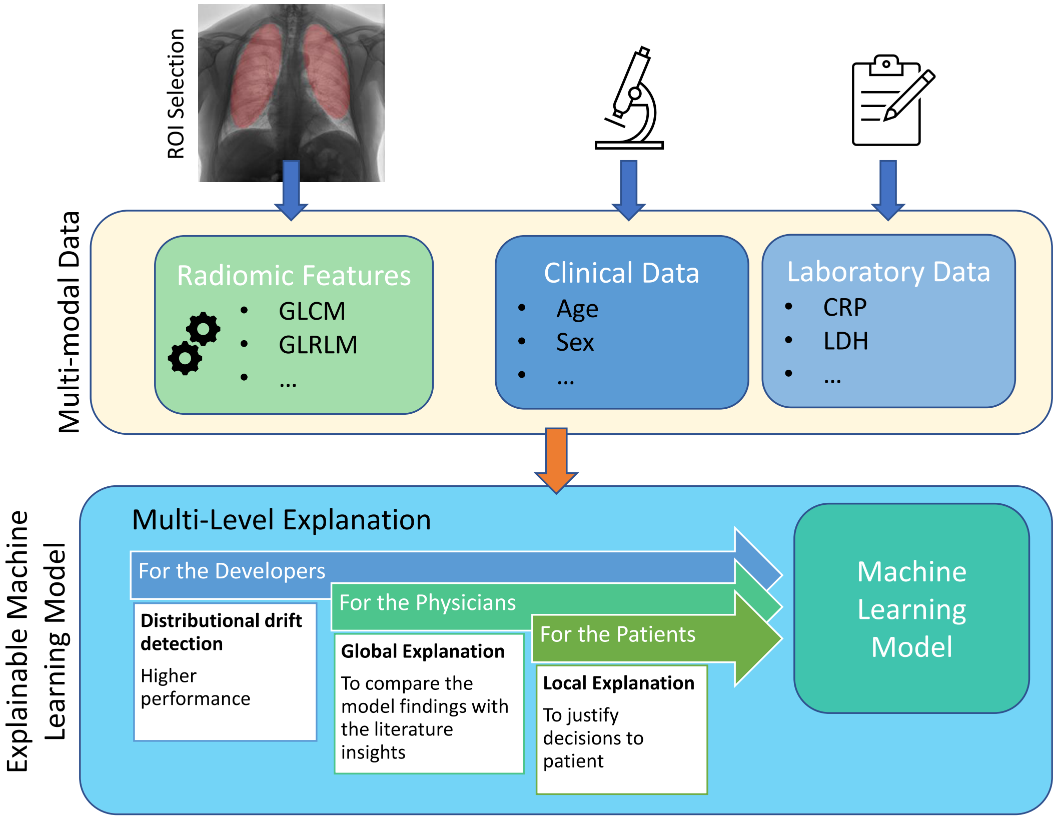


FIGURE 1. The proposed multilevel explainability makes it possible to focus on the needs of key stakeholders involved in the healthcare process.

II. RELATED WORK

With the proliferation of the COVID-19 pandemic, there has been a growing surge of interest in radiomic analysis concerning the ability to infer further knowledge about diagnosis, severity [52], [57], [68] or prognosis [4], [8], [54], [56], [64], [67] of the disease. These studies were conducted considering both unimodal data (i.e. Computer Tomography (CT) or CXR imaging) [54], [67], [68] and multimodal data (i.e. imaging and clinical) [4], [52], [57], [64]. Regarding the imaging modality, while CT scans offer high-quality images, CXRs enable the NHS to perform numerous daily examinations rapidly and efficiently, ensuring sustainability.

Angeli et al. [4] evaluated the prognostic value of CT integrated with clinical and laboratory data. The Pulmonary Involvement (PI) score and the Pulmonary Consolidation (PC) score were extracted from 301 CT images. Univariate and Multivariate Logistic Regressions were used for feature selection and model training to predict the improvement/recovery *vs.* ICU admission or death. There was no notable association observed for the PC score, yielding an area under the curve (AUC) of 0.722 when considering only the PI score. However, when integrating the PI and PC visual-imaging features with demographic, comorbidities, and laboratory features, the AUC improved to 0.841. In [54], 14339 CT images were used to predict overall survival outcomes (alive *vs.* deceased). Texture, intensity, and shape radiomic features were extracted considering the lungs segmentation computed through COLI-Net [53] and selected considering ANOVA, Kruskal-Wallis, Recursive Feature Elimination and Relief methods. Logistic regression, LASSO, Linear discriminant Analysis, Random Forest, AdaBoost, Naive Bayes, and Multi-Layer Perceptron (MLP) were used as classifiers. The combination of ANOVA features selector and Random Forest resulted in the highest performance with an AUC, sensitivity, and specificity of 0.83 (CI 95%: 0.81–0.85), 0.81, and 0.72, respectively. In [64], a court of 188 patients was used to predict aggravation or improvement of disease progression. Regions of Interest (ROIs) encompassing lesions were automatically generated from the 188 CT scans and subsequently subjected to manual revision. A total of 1218 radiomic features were extracted considering original images, Laplacian of Gaussian Filters, and Wavelet Transforms. They integrated radiomic with demographic and laboratory features, selected *via* ICC and F-test methods. Linear regression, SVM, Decision Tree, RF, and XGBoost were used to test only clinical signature, only radiomic signature, and their union. For the radiomics, clinical, and combined features, 0.843 *vs.* 0.813 *vs.* 0.865 of AUC were obtained in the test set, respectively. In [68], 284 CT images were used to classify the progression of COVID-19 into four groups: early, progressive, severe, and absorption. The ROIs were manually segmented, and a total of 1688 radiomic features were extracted including original features and considering logarithm, wavelet, exponent, gradient, square, square root, and local binary pattern. Thirty-eight radiomic features were selected using the select K-best method and the ElasticNet algorithms. The SVM was trained, obtaining a microaverage AUC of 0.89 and a macroaverage AUC of 0.90 on the test dataset. The study described in [52] utilized a primary cohort of 156 COVID-19 patients and a validation cohort of 104 COVID-19 patients from three different hospitals. A radiomic nomogram was developed to identify the severity of infection in COVID-19 patients (Mild/Moderate *vs.* Severe/Critical). They used clinical information and laboratory examinations combined with radiomic features. In particular, a Multi-Task U-Net 2D with a single encoder and two decoders was used for lesion and lung segmentation. Finally, the LASSO regression was applied to select the radiomic signature and compute the Rad-Score as a linear combination between feature values and their regression coefficients. A multivariate Logistic Regression was trained with the radiomic signature, Rad-Score, comorbidity, and abnormal White Blood Cell counts obtaining an AUC of 0.978 in the validation cohort.

In [57], 820 CXR images were used for prognosis prediction (MILD *vs.* SEVERE). They explored the predictivity of only clinical/laboratory features, only radiomic, and their combination. For clinical features they evaluated shallow learning and Deep Learning (DL) approaches, SVM and MLP respectively. For CXR images alone and the combination, they evaluated three approaches: handcrafted, hybrid, and end-to-end deep learning. In particular, for the handcrafted approach, radiomic features were extracted using a pixel-based approach [44] and exploiting the lungs segmented through U-Net and manually refined. For the hybrid approach, several Convolutional Neural Networks (CNNs) (such as AlexNet, VGG, ResNet, DensNet, SqueezeNet, MobileNet and their different variants) were trained for deep features extraction. Then the deep features were concatenated with the clinical and laboratory ones and selected *via* Mutual Information and Recursive Features Elimination. SVM, Logistic Regression, and RF were used as classifiers. For the end-to-end DL approach, deep features extracted using CNNs (with ResNet50 performing the best) were concatenated with clinical features. These deep features from the CNN were processed through a dense structure, and similarly, the clinical features underwent processing through a dense structure before being combined. This architecture was then trained with Stochastic Gradient Descent. Considering only CXR imaging, deep features provide higher accuracy than radiomic features (0.705 *vs.* 0.65). Performance was higher for all three approaches when clinical and imaging features were combined. By far, higher accuracy was obtained when considering the hybrid approach, using GoogleNet and Logistic Regression. An explanation of the results was provided through Grad-CAM. Also in [5] several deep architectures based on ResNet-18 and DenseNet-121 were proposed. Among the proposed approaches, the best one was DenseNet-121 — pre-trained on CheXpert dataset (224k CXR images) and tested on CORDA-SLG dataset (451 CXR images) — able to classify between positive and negative COVID-19 patients. Sensitivity=0.79, specificity=0.82, and AUC=0.84 was computed with the best classifier. To implement explainability Grad-CAM was used.

In [17] several convolutional architectures and dense networks for prognosis prediction (MILD *vs.* SEVERE) were explored. They used the same dataset of [57], and an additional 283 CXRs were considered as external validation. Among the various trained networks, the ensemble between three CNNs (GoogleNet-based, VGG-based, and ResNet-based) and one MLP for clinical data was used. An accuracy of 77.90 ± 1.27 was obtained, considering both imaging and clinical features. They also used Grad-CAM for saliency maps computation of the three CNNs, and Integrated Gradient for MLP. In [8] the authors proposed the Brixia score to assess COVID-19 infection. Each image was divided into 6 zones considering the upper, middle and lower parts of the right and left lungs. Each zone is assigned a score from 0 to 3 to indicate the impairment of the zone, and finally, each score is summed to form a score between 0 and 18. For 100 CXRs, the Brixia score was manually assigned by an experienced thoracic radiologist and used to distinguish between recovery or death patients. Weighted Kappa (k_w) and Mann-Whitney U-test were used to compare CXR scores with the final outcome in selected patients (k_w, 0.82; 95% CI, 0.79–0.86). In [56], the BS-net was proposed for Brixia score prediction, using 5000 CXR. The authors exploited a semi-quantitative approach in order to leverage the sensitivity of CXRs and the ability of radiologists to identify COVID-19 pneumonia. In particular, the BS-Net was implemented as an end-to-end scheme, to segment, align, and predict the Brixia score. The ResNet-18 was used for features extraction; the Neasted U-Net [71] for segmentation; the alignment introducing synthetics transformation such as rotation, scale, shift, elastic transformation, and grid and optical distortion; an optional hard self-attention; the ROI pooling to obtain the 3×2 matrix representing the Brixia score values; the Feature Pyramid Network [30] to combine multi-scale feature maps. They used a sparse categorical cross entropy (SCCE) with a Mean Absolute Error contribution, to solve the Brixia score prediction as a joint-multi-class classification and regression. The BS-Net demonstrates a high degree of accuracy for the Brixia score and other scores (e.g., Toussie Score and GE-LO Score). To improve the explainability of the Grad-CAM algorithm, they proposed a method (inspired by LIME [49]) using the concept of super-pixels [63]. In summary, they computed the difference between the probability maps produced by the i^{th} replica (in which a single super-pixel is masked to zero) and the probability map produced by the model. The explainability maps generated help the understanding of the network activity in the lung areas, improving the poor localization capability of Grad-CAM.

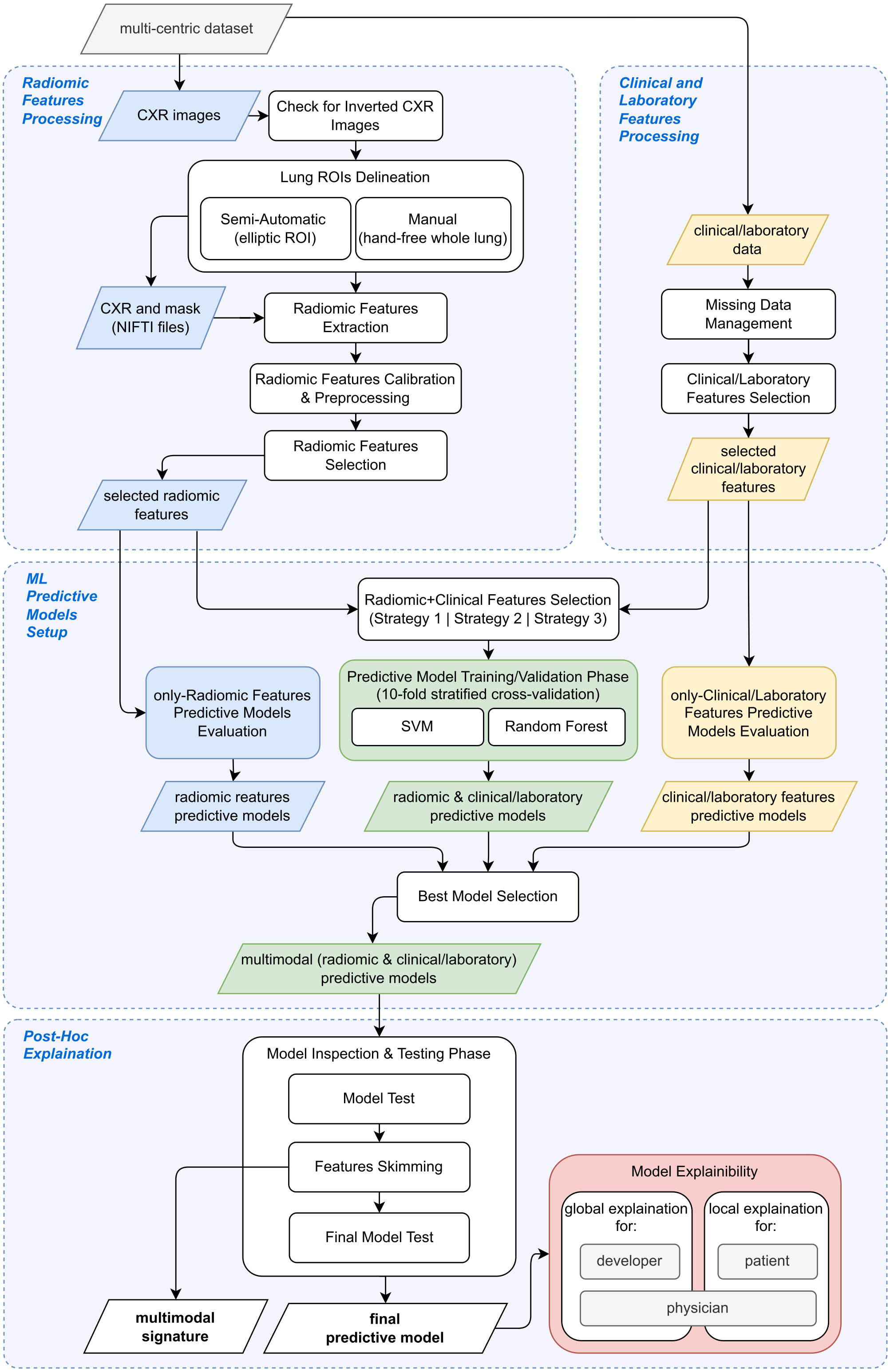

multi-centric dataset
Radiomic Features Processing
CXR images
Check for Inverted CXR Images
Lung ROIs Delineation
Semi-Automatic (elliptic ROI)
Manual (hand-free whole lung)
CXR and mask (NIFTI files)
Radiomic Features Extraction
Radiomic Features Calibration & Preprocessing
Radiomic Features Selection
selected radiomic features
Clinical and Laboratory Features Processing
clinical/laboratory data
Missing Data Management
Clinical/Laboratory Features Selection
selected clinical/laboratory features
ML Predictive Models Setup
Radiomic+Clinical Features Selection (Strategy 1 | Strategy 2 | Strategy 3)
only-Radiomic Features Predictive Models Evaluation
Predictive Model Training/Validation Phase (10-fold stratified cross-validation)
SVM
Random Forest
only-Clinical/Laboratory Features Predictive Models Evaluation
radiomic reatures predictive models
radiomic & clinical/laboratory predictive models
clinical/laboratory features predictive models
Best Model Selection
multimodal (radiomic & clinical/laboratory) predictive models
Post-Hoc Explaination
Model Inspection & Testing Phase
Model Test
Features Skimming
Final Model Test
Model Explainibility
global explaination for:
local explaination for:
developer
patient
physician
multimodal signature
final predictive model

FIGURE 2. **Overall flow diagram depicting the whole processing pipeline implemented, where it is possible to individualize four high-level blocks: *i)* radiomic features processing, *ii)* clinical and laboratory features processing, *iii)* setup of the machine learning predictive models, and, finally, *iv)* implementation of the *post-hoc* explanation.**

Related studies indicate that only a limited number of works have tackled the issue of interpretability. Furthermore, when attempts have been made in this regard, the resulting saliency maps have often produced unsatisfactory and inconsistent interpretations [34], [40]. This aspect is highlighted in [56], which proposed an *ad-hoc* method to overcome the critical issues of Grad-CAM. In our proposed work, our primary objective is to emphasize the importance of explainability. We aim to achieve this by presenting a methodology that effectively addresses the needs of developers, clinicians, and patients.

III. MATERIALS AND METHODS

As introduced in Section I, the article proposes ML models for prognosis prediction of COVID-19 patients, focusing on model explainability. Figure 2 shows the overall workflow. The next subsections describe each block of the processing pipeline in detail.

A. MULTI-CENTRIC DATASET DESCRIPTION

The dataset is composed of clinical, laboratory, and CXR data coming from 1589 COVID-19 patients, classified as 'SEVERE', 'MILD' and 'LIEVE' prognosis. Patients were classified according to the hospital support they received. Specifically, the SEVERE class included patients who required non-invasive ventilation support, intensive care unit (ICU), or who died. The others were considered as MILD [57]. The dataset was collected from 6 different hospitals. This dataset was divided into 1103 patients used for the training/validation of the predictive models and 486 patients used for the testing phase. The division into 1103 and 486 cases was proposed by the organizing committee of the Covid CXR Hackathon competition [19], who made the dataset available. Unfortunately, these are the only samples available for this multicenter study. In addition, the use of only the samples in this dataset is justified by the need to fairly compare our approach with other approaches in the literature (i.e., Soda et al. [57]), which use the same dataset. Informed consent was waived because the dataset analysed in this study is publicly available to members of the scientific community upon request at the Covid CXR Hackathon competition [19]. Table 1 shows the class distribution for each hospital.

1) CXR IMAGES DETAILS

The CXR images were provided in. PNG format (16-bit depth) and no metadata related to acquisition details (e.g., X-Ray imaging parameters, allocated bits, pixel spacing, etc.) are available. From a mere qualitative assessment, the dataset appeared highly heterogeneous in terms of both size and overall quality. Table 2 shows the size distribution among the various centres.

The images were acquired at 6 different centres (indicated as A, B, C, D, E, and F), leading to CXR images with great variability in terms of intrinsic image quality and acquisition conditions. Concerning the quality, it seems that — in addition to natively digital images — there are also images obtained by scanning X-Ray plates, resulting in poor-quality images. Moreover, some images exhibit an inverted pattern, in contrast to the typical representation where bones are depicted as hyperintense regions (indicative of high density) and lung hypointense areas (indicative of low density). For this reason, we inverted the images with an opposite pattern to the conventional one (where bones are represented with hyperintense regions). Instead, concerning the environment, the dataset contains images of patients with both permanent life-support devices (such as pacemakers) and temporary ones (e.g., tubes for forced ventilation, thoracic electrodes, and monitoring wires).

2) CLINICAL AND LABORATORY FEATURES SELECTION AND DATA IMPUTATION

For each patient, clinical and laboratory data were associated with the CXR image (the complete features list is provided in A). The *prognosis* feature was used as a label for supervised training. Then, only 23 features were used as input. In particular:

- 3 features (i.e., *Hospital* and *Position*, *Death*) were excluded *a-priori*;
- 5 features (i.e., *Fibrinogen*, *PCT*, *dDimer*, *SaO2*, *Obesity*) were not considered because a missing data percentage over 50%;
- 6 features (i.e., *OxPercentage*, *CardiovascularDisease*, *IschemicHeartDisease*, *AtrialFibrillation*, *HeartFailure*, *Ictus*) were excluded because not present on the test set.

Univariate and multivariate data imputation techniques were implemented to handle missing values of the remaining 23 clinical features. In the first case, mean and median values were used. In the second case, a regressor was considered. In particular, at each step, the feature column to impute is designated as output and the other feature columns are treated as inputs for a regressor. The regressor is then used to predict the missing values of the feature considered.

B. LUNG ROIS DELINEATION ASSESMENT

A MatLab-coded custom tool was implemented to delineate the lung ROIs for radiomic feature extraction. In particular, the following two segmentation modalities were implemented:

1) *hand-free whole lung delineation*: manual modality to detect the entire left and right lungs. These segmentations were performed by a radiologist with more than 3 years of experience in X-Ray annotation, in consensus with a consultant senior radiologist.
2) *semi-automated elliptic ROI delineation*: semi-automatic modality employed to identify the maximum elliptical region that is fully contained within the lungs. The operator only needs to centre the bounding box on the lung, and automatically the implemented software locates the ellipse. This delineation modality was implemented to focus the attention only on the central area, excluding peripheral zones.

The GUI allows 1) interactive selection of the two selection modes; 2) execution of segmentation; and 3) final saving. Specifically, the image and its mask were saved in NIFTI format.

TABLE 1. Multi-centric dataset characteristics used for the predictive models training/validation and the testing phases.

Hospital	Phase	Image Number	'SEVERE' cases (%)	'MILD' cases (%)	'LIEVE' cases (%)
A	train/validation	120	85 (70.83)	35 (29.17)	n.a.
B	train/validation	104	45 (43.27)	59 (56.73)	n.a.
C	train/validation	151	81 (46.36)	70 (53.64)	n.a.
D	train/validation	139	63 (45.33)	76 (54.67)	n.a.
E	train/validation	101	46 (45.55)	55 (54.45)	n.a.
F	train/validation	488	248 (50.81)	151 (30.94)	89 (18.25)
All	train/validation	1103	568 (50.36)	446 (46.6)	89 (3.04)
F	test	486	180 (37.04)	306 (62.96)	n.a.

TABLE 2. Variability in CXR image size across the different hospitals. Only the top 3 most frequent sizes (along with the number of images and percentage) are reported for each centre.

Hospital	1^{st} most frequent size $(r \times c) : \#imgs - \%imgs$	2^{nd} most frequent size $(r \times c) : \#imgs - \%imgs$	3^{rd} most frequent size $(r \times c) : \#imgs - \%imgs$
A	$(4280 \times 3520) : 75 - 62.5\%$	$(2500 \times 2048) : 20 - 16.6\%$	$(2772 \times 2771) : 10 - 8.3\%$
B	$(4240 \times 3480) : 90 - 86.5\%$	$(2846 \times 2330) : 2 - 1.9\%$	$(2836 \times 2336) : 2 - 1.9\%$
C	$(2866 \times 2350) : 66 - 43.7\%$	$(3000 \times 3000) : 6 - 3.9\%$	$(2917 \times 2402) : 6 - 3.9\%$
D	$(2648 \times 2208) : 33 - 23.7\%$	$(2140 \times 1760) : 21 - 15.1\%$	$(2648 \times 2176) : 21 - 15.1\%$
E	$(4280 \times 3520) : 33 - 32.6\%$	$(2880 \times 2880) : 24 - 23.7\%$	$(2936 \times 3080) : 8 - 7.9\%$
F	$(2836 \times 2336) : 392 - 80.3\%$	$(2336 \times 2836) : 17 - 3.4\%$	$(2012 \times 2012) : 7 - 1.4\%$

TABLE 3. Number of patients with health-supporting on the CXR image.

Subset	Prognosis	Health-supporting devices patients (*)
training/validation (1103)	MILD (535) SEVERE (568)	23 (4.30%) 51 (8.98%)
testing (486)	MILD (306) SEVERE (180)	16 (5.23%) 37 (20.55%)

(*) Value in parenthesis represents the percentage of the samples calculated with respect to the number of images in the corresponding prognosis class.

The radiomic features quantify the distribution/texture of the lung tissues. Considering that the image presents a lot of external health-supporting devices (as summarized in Table 3, the presence of external health-supporting devices (e.g., pacemakers, monitoring wires, respirator pipes, etc.) would alter the extracted feature values. For this reason, areas containing external health-supporting devices were excluded in both delineation modalities (as in the SEVERE case shown in Figure 3).

C. RADIOMIC FEATURES EXTRACTION

A total of 1023 features were extracted by means of the PyRadiomics [62], [72] toolkit. In particular, 93 original features were extracted, considering:

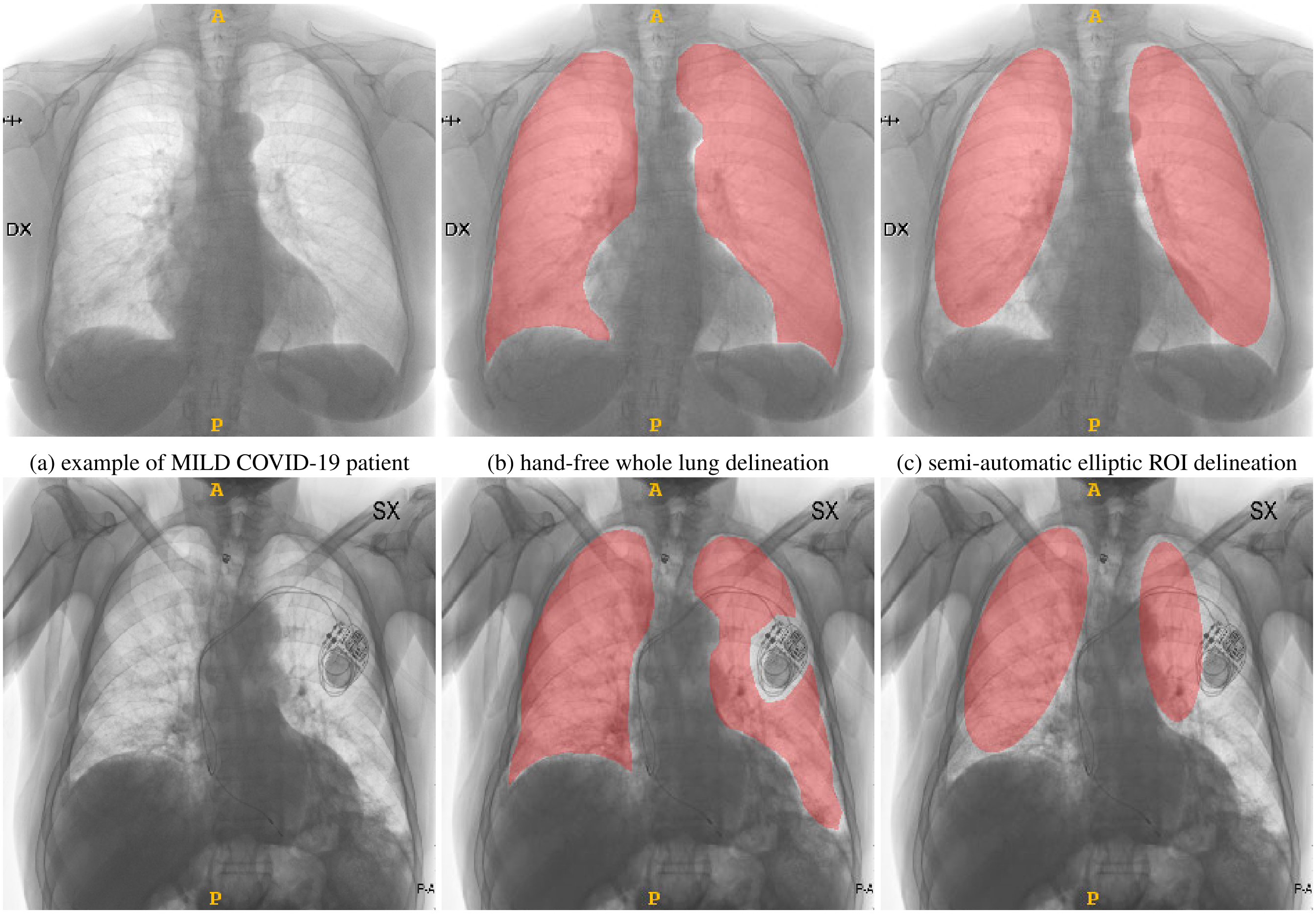
- first-order intensity histogram statistics;
- Gray Level Co-occurrence Matrix features (GLCM) [20], [21];
- Gray Level Run Length Matrix (GLRLM) [13];
- Gray Level Size Zone Matrix (GLSZM) [60];
- Gray Level Dependence Matrix (GLDM) [59];
- Neighboring Gray Tone Difference Matrix (NGTDM) [3].

Then the same features were extracted considering Laplacian of Gaussian (LoG) and Wavelets filtered images. For LoG filtering three different values of σ were considered ($\sigma \in \{1, 3, 5\}$), collecting 279 features ($279 = 93 \times 3$); for Wavelets transform, the Haar kernel [47] and two decomposition levels ($levels \in \{1, 2\}$) were considered, obtaining 651 features ($651 = 93 \times 7$). Finally, 930 features were extracted from the filtered images. Moreover, to determine the optimal quantization level, the features were extracted considering different $binWidth$ values ($binWidth \in \{8, 16, 32, 64, 128, 256\}$).

D. RADIOMIC FEATURES CALIBRATION AND PRE-PROCESSING

Features calibration and pre-processing were performed by following the steps [42]:

1) **quantization level analysis**: the quantization [62] level was established considering the highest number of radiomic features according to the Intraclass Correlation Coefficient (ICC); this analysis allowed to determine the optimal width of bins, which maximized the number of robust (in terms of ICC) features. In this study, the two-way random-effects model, consistency, single rater/measurement ICC, named ICC(3,1), was considered [37], [55];
2) **near-zero variance analysis**: the objective of the near-zero variance analysis was to eliminate features with near-zero variance values. The features exhibiting variance less than or equal to 0.01 were discarded;
3) **redundant features analysis**: this step was involved in removing highly correlated features, using the Spearman correlation for pairwise feature comparison. Considering that values greater than 0.80 are commonly used for Spearman correlation [28], [39], [41], [70], a threshold of 0.85 was chosen.
4) **statistical analysis**: the Mann-Whitney U test was used to test the difference between MILD and SEVERE distribution computing the p-value for each feature selected in the previous steps. The p-value threshold



(a) example of MILD COVID-19 patient
(b) hand-free whole lung delineation
(c) semi-automatic elliptic ROI delineation
(d) example of SEVERE COVID-19 patient
(e) hand-free whole lung delineation
(f) semi-automatic elliptic ROI delineation

FIGURE 3. Two examples related to the implemented annotation modalities: in (a) and (d) original CXR images of MILD and SEVERE patients, respectively; in (b) and (e) the *hand-free whole lung* delineations; in (c) and (f) the *semi-automatic elliptic ROI* delineations. To avoid altering the radiomic features extracted from lungs, areas containing external health-supporting devices were excluded in the segmentation, such as in the SEVERE case above (subfigures (d), (e), (f)).

was 0.05. Obtained p-values were adjusted using the Bonferroni–Holm method [66].

E. FEATURES SELECTION AND PREDICTIVE MODEL SETUP

1) ELLIPTIC VS. HANDCRAFTED SEGMENTATION EVALUATION

Feature selection was preceded by the evaluation of the most predictive segmentation technique (i.e., *hand-free whole lung* and *automated elliptic ROI*), using the Sequential Feature Selector [51] algorithm. Sequential Feature Selector was set in *forward* (SFS) and *floating mode* (SFFS) mode. A stratified 10-fold cross-validation (CV) for performance evaluation was considered, using SVM and RF as algorithms. Data were normalized for all experiments involving SVM.

2) RADIOMIC AND CLINICAL/LABORATORY FEATURE SELECTION

A preliminary feature selection was performed to evaluate the performance of each unimodal model separately (i.e., clinical/laboratory, radiomic). SFFS was used to evaluate how performance varies as the feature number increases, and, consequently, to choose the optimal number of features to maximize accuracy. For the clinical/laboratory features, SFFS was set to show the accuracy trend considering all the features (23). Instead, for the radiomic features SFFS was set to select the best 30 features. All experiments involving feature selection with SFFS were performed through a stratified 10-fold CV. Subsequently, three strategies were applied:

- *Selection Strategy 1*: SFFS was applied considering the clinical/laboratory and radiomic features selected in the preliminary selection step.
- *Selection Strategy 2*: SFFS was applied to all the clinical/laboratory and all radiomics features.
- *Selection Strategy 3*: SFFS was applied to the optimal number of radiomic features selected in the preliminary selection step and all the clinical/laboratory features. This strategy was implemented to balance the ratio between clinical/laboratory and radiomic features.

3) MODEL TRAINING AND TEST

Before the training/validation phase, imputation for clinical and laboratory data was performed, as well as calibration, and pre-processing of radiomic features. Successively, for each feature selection strategy, a stratified 10-fold CV was repeated 20 times for hyperparameters tuning. In terms of accuracy, the best model obtained in the CV procedure was selected and used to evaluate the performance on the test set.

F. MULTI-LEVEL EXPLAINABILITY

The development and integration of a Clinical Decision Support System (CDSS) into real clinical practice require that the system is explainable and the decision understandable by users. Therefore, the proposed multi-level explanation takes into account the perspectives of the developer and the stakeholders involved in the care process (e.g., physician and patient)

1) DEVELOPER PERSPECTIVE

The developer aims to train models able to generalize on unseen data. Although stingy validation protocols may allow detecting the overfitting problem on the training distribution, the model may perform poorly when a distribution encountered during deployment is slightly different [65]. This occurrence, commonly known as *distribution drift*, can be addressed through the use of explainable AI algorithms.

To detect and avoid the distributional drift issue, the Mean Decrease Accuracy (MDA) method available in ELI5 framework [12] was used. Features' importance was calculated through the Leave One Center Out (LOCO) procedure, where each centre represents one of the six hospitals (i.e., A, B, C, D, E, and F). The LOCO evaluation consists of splitting the dataset samples for each centre and assigning, to each iteration, five of the six centres to the training and the remaining one to the test. The following methodology was used to drop center-dependent features and select only the descriptive features of the COVID-19 prognosis. In particular:

- the MDA method was used to calculate the features' importance of each centre. For example, to compute feature importance for hospital A, all hospitals were used for the training, and A for test and MDA computation. This procedure is repeated for each hospital.
- according with MDA method, a positive weight is representative of significant features, *vice versa* a negative weight is representative of unstable features.
- to define a feature as stable, 3 different criteria were established, selecting features that in at least 3, 4, or 5 centres (out of 6) obtained positive weights; features with less than 3 positive weights (across all hospitals) were considered dependent on the acquisition centre and then discarded.

Applying this procedure, three different skimmed subsets of features were obtained (features having at least 3, 4, or 5 positive weights on the six centres). To evaluate the approach improvement, model performance was computed without this debugging step and considering only the features in the 3 subsets. The rationale is that removing features that lack stability across multiple centres (i.e., features that are not informative of the phenomenon but rather dependent on the originating hospital) can enhance the generalization capabilities of the trained models.

TABLE 4. Quantization level analysis results.

Bin Width Value	Robust Features Number
8	319
16	407
32	**573**
64	488
128	416
256	381

2) PHYSICIAN PERSPECTIVE

Physicians need to ensure that the learned patterns by the model are supported by clinical evidence. Through the use of inherently interpretable features, such as clinical, laboratory, and radiomic, results can be compared with clinical practice, and inconsistent behaviour with medical literature can be detected. The SHapley Additive exPlanations (SHAP) analysis [33] was used to provide a global explanation. In particular, was employed to identify the features that drive the system's output towards either a SEVERE or MILD prediction. This step was addressed with a medical team, which was able to compare and verify the results obtained with the medical literature.

3) PATIENT PERSPECTIVE

Finally, the General Data Protection Regulation (GDPR) [43] imposes an explanation on the users who receive the systems' decisions: the patients. For this purpose, a local explanation is performed for each specific instance. The SHAP analysis was also used to obtain a local explanation and to evaluate the features pushing the model toward a SEVERE or MILD decision.

IV. EXPERIMENTAL RESULTS

A. RADIOMIC FEATURES PRE-PROCESSING AND LUNG DELINEATION SELECTION

Starting from the initial set of 1023 radiomic features, calibration and pre-processing steps were performed to select robust, informative, and non-redundant features. The ICC analysis was used to establish the best quantization level considering $binWidth \in \{8, 16, 32, 64, 128, 256\}$. Table 4 shows the number of robust features for each bin width. A $binWidth = 32$ was chosen (considering $ICC \geq 0.85$) and used in all subsequent steps of the processing pipeline.

Then, the number of radiomic features was progressively reduced within each pre-processing step, and a final set of 40 features was obtained (see Table 5).

Finally, Table 6 reports the accuracy values (computed through SFFS, during the 10-fold CV procedure) obtained

TABLE 5. Calibration and pre-processing of radiomic features.

Pre-processing Step	Analysis Method	Remaining Features
initial features	n.a.	1023
near-zero variance analysis	$variance \leq 0.01$	354
redundant features analysis	Spearman correlation ($cutoff = 0.85$)	57
statistical analysis	Mann-Whitney U rank test ($p < 0.05$)	40

TABLE 6. Evaluation and choice of the best lung delineation approach. With both classifiers (i.e., SVM and RF), the automated elliptic ROI modality shows slightly better behaviour than the hand-free whole lung modality. Radiomic features considered here belong to both types (original and filtered).

Classifier	Delineation Approach	Selected Features	Accuracy
SVM	whole lung	19	0.673 ± 0.050
	elliptic ROI	22	0.710 ± 0.036
RF	whole lung	12	0.687 ± 0.071
	elliptic ROI	16	0.703 ± 0.057

in the evaluation of the best lung delineation approach (i.e., *hand-free whole lung* and *automated elliptic ROI* delineations). Both SVM and RF showed a higher accuracy using the features extracted from elliptic ROIs. Automatic elliptical ROI segmentation reaches better results than hand-free modality. This result was justified by the clinicians considering that elliptic modality focuses on the central area of the lung, the most representative with respect to the peripheral ones. For this reason, features extracted from the elliptic ROIs were used in the following experiments.

B. IMPUTATION OF MISSING VALUES IN CLINICAL DATA

SFFS was used also to select the best imputation method. Table 7 shows the results of SVM and RF using the three clinical data imputation approaches. No significant differences were calculated between the used approaches. However, a smaller standard deviation was obtained with the mean. Therefore, according to [57], the mean was used for data imputation.

TABLE 7. Findings obtained with the different imputation approaches used to manage missing data in clinical features.

Classifier	Imputation Approach	Selected Features	Accuracy
SVM	Mean	11	0.750 ± 0.031
	Median	11	0.753 ± 0.046
	LR	10	0.748 ± 0.042
RF	Mean	15	0.728 ± 0.031
	Median	15	0.746 ± 0.038
	LR	14	0.737 ± 0.042

C. FEATURE SELECTION AND MODEL TRAINING

As introduced previously, SFFS [51] was used to select the best features subset maximizing accuracy, within a stratified 10-fold CV. More details are provided in B, where Figure 7 shows the preliminary feature selection results to evaluate the unimodal models. Considering the features number maximize accuracy, Table 8 shows the performance obtained by SVM and RF, considering only clinical/laboratory features, and only radiomic.

After the preliminary selection, three feature selection strategies combining both clinical/laboratory and radiomic features were implemented to evaluate the multimodal model:

- *Selection Strategy 1*: in this case, 22 radiomic and 11 clinical/laboratory were considered for SVM and 16 radiomic and 15 clinical/laboratory for RF.
- *Selection Strategy 2*: SFFS was applied on all the clinical/laboratory (23) and all radiomic (40) features.
- *Selection Strategy 3*: In this case, 22 radiomic and 23 clinical/laboratory for SVM, 16 radiomic and 23 clinical/laboratory for RF.

Table 9 summarizes the training/validation performance computed for the three selection strategies, considering the 20-repeated stratified 10-fold CV. As expected, when clinical laboratory and radiomic features are used simultaneously, model performance improves compared with unimodal models [4], [64].

D. PREDICTIVE MODELS TEST

Considering the higher performance, the multimodal models were used for the testing phase. The AUC is the most widely used index of global diagnostic accuracy since higher values correspond to a better selective ability of the biomarkers [58]. Therefore, AUC was used to select the best predictive model. As a matter of fact, AUC values obtained in the test had a minimal decrease compared with the training/validation phase, demonstrating promising generalization capability. Table 10 summarizes the results obtained in the testing phase for SVM and RF. RF achieved better performance compared to SVM. The *Selection Strategy 1* was the best strategy, obtaining an AUROC of 0.800. It can be concluded that the Random Forest + *Selection Strategy 1* guarantees the highest performance.

E. MODEL INSPECTION AND FINAL TEST PERFORMANCE

To further improve the generalization capabilities of the model, additional feature skimming was performed. In particular, MDA [12] analysis, calculated in LOCO modality, was used to remove the features subject to distributional drift. Starting from the most performing model (Random Forest + *Selection Strategy 1*) and using the weights provided by MDA across the 6 centres (i.e., hospitals), we reduced the set of input features. Table 11 shows the features with positive weight in more than 3, 4, and 5 centers, simultaneously. Three feature subsets composed of 17, 11, and 6 features were obtained and used to retrain the models and recalculate the test performance. Table 12 shows the obtained improvement in discarding features susceptible to the distributional drift phenomenon. In particular, accuracy=0.733 and AUROC=0.819 were obtained using the features with positive weight in 4 centres simultaneously,

TABLE 8. Preliminary feature selection result obtained by SVM and RF. This selection is used to determine the optimal signature, considering radiomic and clinical/laboratory features separately. For each metric, the *mean value* ± *standard deviation* and the confidence interval are reported.

	SVM		RF	
Evaluation Metrics	**Radiomic Features**	**Clinical / Laboratory Features**	**Radiomic Features**	**Clinical / Laboratory Features**
Accuracy	0.694 ± 0.039 [0.686, 0.701]	0.750 ± 0.041 [0.742, 0.758]	0.672 ± 0.044 [0.664, 0.680]	0.721 ± 0.038 [0.714, 0.728]
Sensitivity	0.668 ± 0.056 [0.658, 0.678]	0.772 ± 0.050 [0.763, 0.781]	0.659 ± 0.065 [0.647, 0.671]	0.736 ± 0.054 [0.726, 0.746]
Specificity	0.720 ± 0.062 [0.708, 0.731]	0.724 ± 0.064 [0.712, 0.736]	0.685 ± 0.062 [0.673, 0.697]	0.703 ± 0.061 [0.692, 0.714]
AUC	0.741 ± 0.044 [0.732, 0.749]	0.804 ± 0.041 [0.796, 0.812]	0.719 ± 0.049 [0.710, 0.728]	0.794 ± 0.039 [0.787, 0.801]
Selected Features	**22**	**11**	**16**	**15**

TABLE 9. Performance obtained in the training/validation phase by the SVM and RF classifiers, with the 10-fold stratified CV procedure (20 repetitions were performed). For each metric, the *mean value* ± *standard deviation* and the confidence interval are reported.

	SVM			RF		
Evaluation Metrics	**Selection Strategy 1**	**Selection Strategy 2**	**Selection Strategy 3**	**Selection Strategy 1**	**Selection Strategy 2**	**Selection Strategy 3**
Accuracy	0.748 ± 0.040 [0.741, 0.755]	0.755 ± 0.039 [0.748, 0.762]	0.760 ± 0.036 [0.753, 0.768]	0.741 ± 0.040 [0.733, 0.748]	0.746 ± 0.041 [0.738, 0.754]	0.746 ± 0.042 [0.738, 0.754]
Sensitivity	0.741 ± 0.060 [0.730, 0.752]	0.768 ± 0.051 [0.758, 0.778]	0.781 ± 0.052 [0.771, 0.790]	0.743 ± 0.054 [0.733, 0.753]	0.753 ± 0.059 [0.742, 0.764]	0.747 ± 0.057 [0.736, 0.758]
Specificity	0.755 ± 0.053 [0.745, 0.765]	0.742 ± 0.060 [0.731, 0.753]	0.738 ± 0.056 [0.728, 0.748]	0.738 ± 0.064 [0.726, 0.750]	0.740 ± 0.061 [0.729, 0.751]	0.745 ± 0.061 [0.734, 0.756]
AUC	0.803 ± 0.041 [0.795, 0.811]	0.816 ± 0.041 [0.808, 0.824]	0.827 ± 0.035 [0.820, 0.834]	0.812 ± 0.040 [0.805, 0.819]	0.813 ± 0.042 [0.805, 0.821]	0.815 ± 0.039 [0.808, 0.822]
Selected Features (clinical/laboratory, radiomic)	**18** (5, 13)	**38** (16, 22)	**21** (14, 7)	**21** (12, 9)	**38** (17, 21)	**21** (14, 7)

TABLE 10. Performance obtained in the testing phase by the SVM and RF classifiers.

	SVM			RF		
Evaluation Metrics	**Selection Strategy 1**	**Selection Strategy 2**	**Selection Strategy 3**	**Selection Strategy 1**	**Selection Strategy 2**	**Selection Strategy 3**
Accuracy	0.707	0.720	0.709	0.705	0.697	0.706
Sensitivity	0.700	0.794	0.794	0.727	0.688	0.755
Specificity	0.712	0.647	0.624	0.683	0.702	0.656
AUC	0.775	0.783	**0.778**	**0.800**	0.795	0.796

vs. accuracy=0.705 and AUROC=0.800 without the distributional drift management.

V. DISCUSSION AND ANALYSIS

In this work, a ML model was proposed to provide an explainable output for COVID-19 prognosis prediction. The model aims to support physicians in discriminating among different disease evolution. The clinical scenario needs an explanation of the predictions to justify the decision-making process. For this reason, we proposed a multi-level explanation to address the needs of the stakeholders involved in the model development and in the clinical decision process (i.e., developer, physician, patient). In fact, intrinsically interpretable clinical, laboratory, and radiomic features were used to allow model introspection and a global and local explanation.

A. CLINICAL VALIDATION

The SHAP Tree Explainer [33] allowed interpretation and clinical validation of the model findings. Figure 4 shows the selected features with the highest impact on the trained model. Clinical and laboratory features show an important correspondence with findings applied in clinical practice:

- patients with high values of *Lactate DeHydrogenase Concentration* (LDH) in the blood are generally predisposed to SEVERE diseases, while low values seem to have greater resistance and are limited to MILD diseases [22];
- low values of *Partial Pressure of oxygen* (PaO2) in arterial blood are indicative of SEVERE disease, while high values indicate a MILD level disease [61];
- the clinical evidence confirmed that the high values of *C-Reactive Protein* (CRP) are an indicator of SEVERE disease, while parameter low values indicate a MILD level disease [2];
- male subjects [38], and, naturally, older subjects are more exposed to severe disease (*Sex, Age*).

Laboratory parameters, such as LDH and CRP, are associated with the most severe forms of COVID-19 disease.

TABLE 11. Starting from the 21 features selected *via* the *Selection Strategy 1* (see Subsection III-E2) with the RF classifier, several selection criteria were applied (i.e., th=3+, 4+, and 5+, respectively) to remove distributional drift-affected features.

Feature Name	Feature Category	Image Type	Positive Weight (th=3+)	Positive Weight (th=4+)	Positive Weight (th=5+)
Age	n.a.	n.a.	X	X	X
Sex	n.a.	n.a.	X	X	-
DaysFever	n.a.	n.a.	-	-	-
DifficultyInBreathing	n.a.	n.a.	X	X	X
WBC	n.a.	n.a.	X	-	-
RBC	n.a.	n.a.	X	-	-
CRP	n.a.	n.a.	X	X	-
LDH	n.a.	n.a.	X	X	X
PaO2	n.a.	n.a.	X	X	X
Diabetes	n.a.	n.a.	-	-	-
Cancer	n.a.	n.a.	X	-	-
RespiratoryFailure	n.a.	n.a.	-	-	-
Kurtosis	first order	original	X	-	-
DependenceNonUniformity	gldm	original	X	X	-
HighGrayLevelZoneEmphasis	glszm	LoG ($\sigma = 1.0mm$)	X	X	-
Maximum	first order	LoG ($\sigma = 3.0mm$)	X	-	-
Skewness	first order	LoG ($\sigma = 5.0mm$)	X	X	-
Kurtosis	first order	wavelet HL	X	X	X
ZoneEntropy	glszm	wavelet HL	X	X	X
HighGrayLevelZoneEmphasis	glszm	wavelet HH	-	-	-
ZoneEntropy	glszm	wavelet HH	X	-	-
Remaining Features			**17**	**11**	**6**

TABLE 12. Performance obtained in the testing phase by the RF classifier after MDA features skimming.

RF	Positive Weights (th=3+)	Positive Weights (th=4+)	Positive Weights (th=5+)
Accuracy	0.710	**0.733**	0.681
Sensitivity	0.711	**0.761**	0.722
Specificity	**0.709**	0.705	0.640
AUC	0.800	**0.819**	0.765

This is not surprising, given that they are an expression of the inflammatory cascade. Previous studies have already confirmed that high levels of serum LDH are among the best predictors of clinical worsening. Similarly, high blood values of CRP predict the worst clinical evolution of patients who are diagnosed with COVID-19 diseases. In a recent study [45], both LDH and CRP were found to increase the accuracy of the COVID-19 diagnosis in suspected patients with respiratory symptoms. In another study [11], both laboratory parameters were demonstrated to be able to predict — in combination with radiological features — the need for invasive ventilation in patients with COVID-19 pneumonia. In line with this, it is logical to predict that low levels of PaO2 are associated with severe disease because they describe a condition of lung failure. Since the first wave of the pandemic, it became evident that hypoxemia at the time of diagnosis identified the most severe cases of COVID-19 disease, who experienced the highest risk of severe respiratory distress and death. This lung functional parameter is included in the most used predictive scores for the identification of subjects with acute respiratory failure by COVID-19 at risk of mortality [23]. The current findings confirm the importance of adding in the discriminating approach the lung functional and laboratory parameters that are an expression of hyper-inflammatory processes and lung involvement.

As concerning the radiomic features, the wavelet-derived and LoG-derived features showed high discriminatory properties. Figure 5 shows an example of wavelet and LoG images.

The most important features belong to the GLSZM category, quantifying grey level zones in an image (where a grey level zone is defined as the number of connected pixels sharing the same grey level intensity). In particular:

- for *HighGrayLevelZoneEmphasis*, a higher value indicates a greater proportion of higher grey-level values and size zones in the image. In our case, high values mean that the lung is more uniform (with large uniform regions) and no lesions are present;
- for *ZoneEntropy*, SEVERE patients show a more heterogeneous texture. Hence, the behaviour of ZoneEntropy is analogous to HighGrayLevelZoneEmphasis in the classification process.

The behaviour of the two radiomic features is also proved by the connected components analysis of the processed images. In fact, the mean area of the connected components is significantly larger in MILD patients than in SEVERE patients. This means that MILD patients show a more regular pattern (with larger and more uniform regions), while SEVERE patients show a more inhomogeneous pattern (with smaller irregular regions). This analysis introduces and assesses a method for lung impairment degree estimation. Both zoneEntropy and HighGrayLevelZoneEmphasis low values lead to SEVERE lung impairment predictions, while high values lead to MILD lung impairment predictions. The

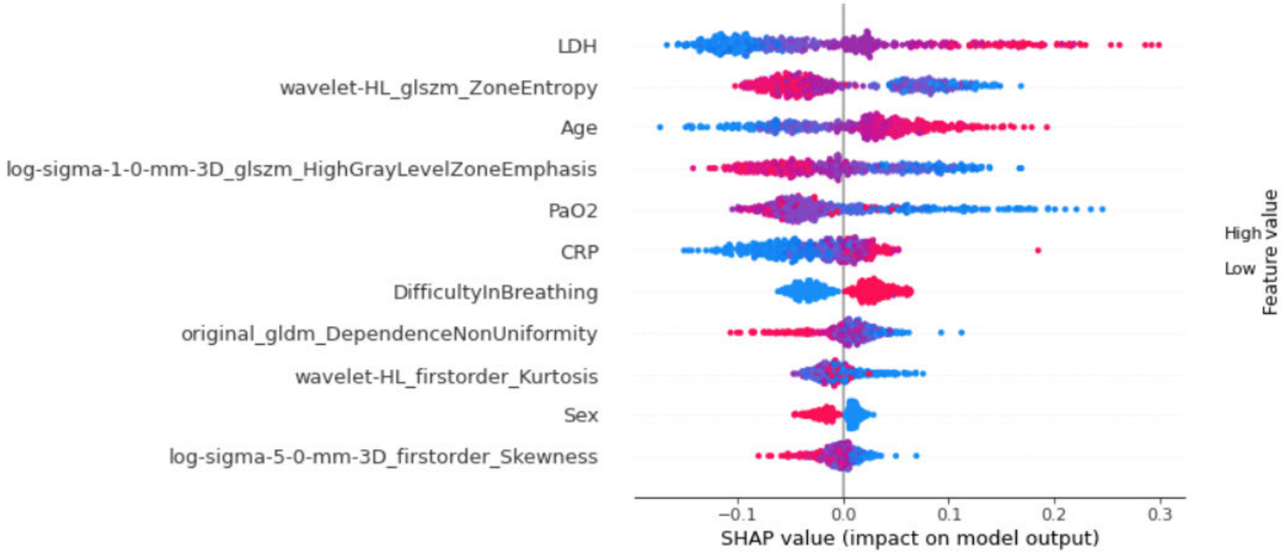
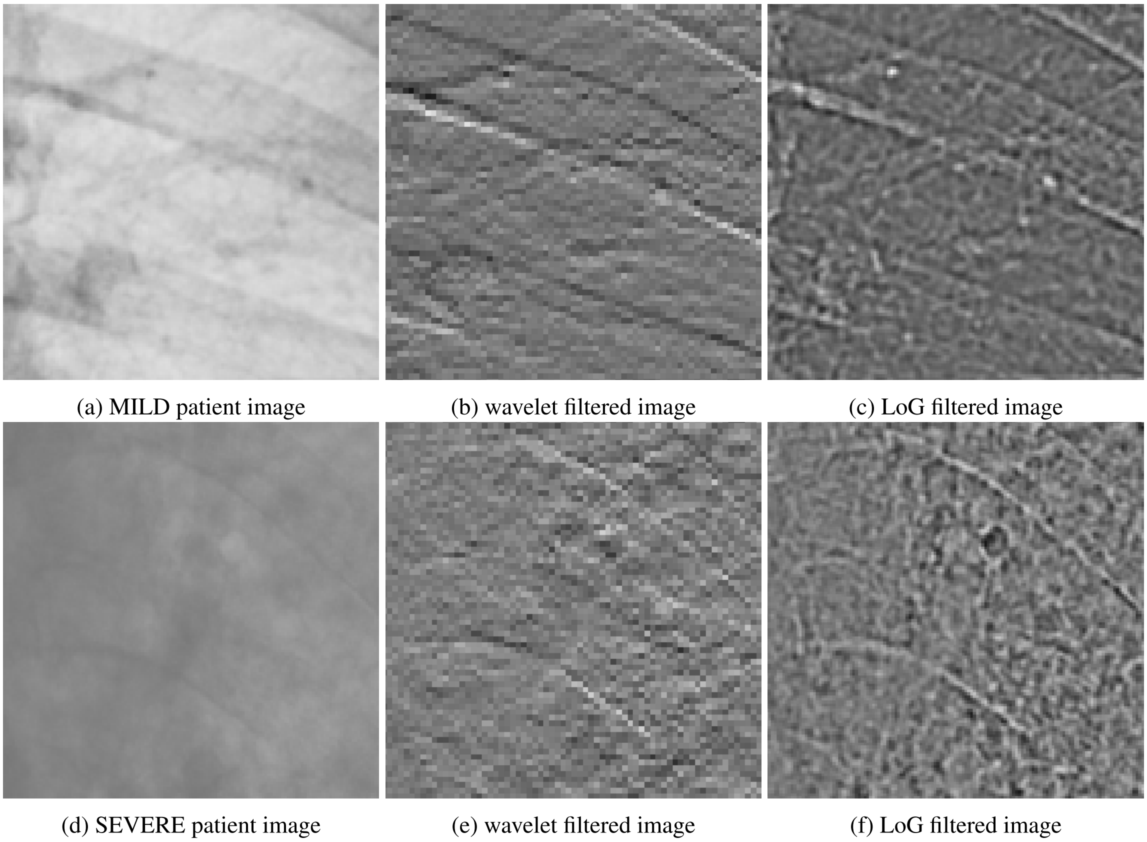


FIGURE 4. **The beeswarm plot proposed by SHAP was used to evaluate the importance of each feature of the trained model. For each test sample, Shapley values were calculated and aggregated in the graph. In particular, the importance of the features is ordered in a decreasing way, so LDH is the most significant feature for the classification, followed by ZoneEntropy (Wavelet HL, GLSZM), etc. The colour of the dots is representative of low (blue) or high (red) feature values. The presence of dots on the left or right side of the vertical line (Shapley value equal to 0.0) means that this specific feature leads the model to go towards a MILD or SEVERE prediction, respectively.**


(a) MILD patient image

(b) wavelet filtered image

(c) LoG filtered image

(d) SEVERE patient image

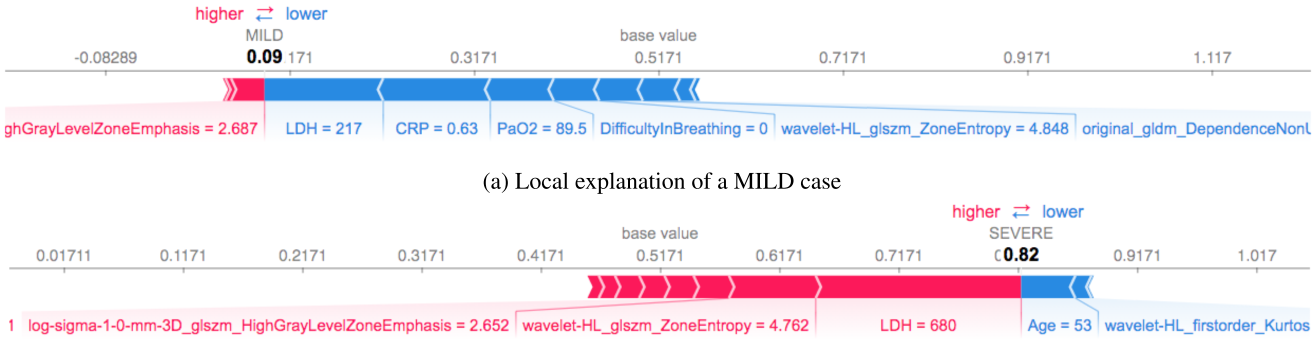
(e) wavelet filtered image

(f) LoG filtered image

FIGURE 5. **In (a) and (d) the lung area of MILD and SEVERE patients, respectively; in (b) and (e) images after filtering with wavelet Haar HL; in (c) and (f) images after filtering with LoG with $\sigma = 1.0$ *mm*.**

result suggests higher variability between connected regions due to a partial lung region impairment.

Using the SHapley values, it was also possible to obtain a local explanation to assess the predictions for each



(a) Local explanation of a MILD case

(b) Local explanation of a SEVERE case

FIGURE 6. The graphs show a local explanation of two clinical cases predicted as MILD (a) and SEVERE (b). Features leading the model's prediction toward a MILD outcome are represented in blue, while features leading the model to predict SEVERE are represented in red.

specific patient. Figure 6 shows an example of two patients predicted as MILD and SEVERE, respectively. For the first case (Figure 6a), the normal LDH (217), CRP (0.63), PaO2 (89.5), and absence of respiratory distress, push the prediction towards a MILD prognosis. For the second case (Figure 6b), a high LDH value (680), and medium/low value of ZoneEntropy lead to the prediction of SEVERE prognosis. These examples demonstrate how explainability makes a predictive model also a CDSS for the physician, allowing an exhaustive explanation.

B. PERFORMANCE DISCUSSION AND LITERATURE COMPARISON

In the testing phase, RF trained using the skimmed signature with 4+ positive weights in the LOCO modality achieved an accuracy=0.733 and AUC=0.819. It demonstrates promising generalization capabilities and minimal performance degradation with respect to training/validation performance.

Our work can be compared fairly with [5], [17], and [57], the only literature works using a subset of our same dataset. In particular, considering only CXR images, in [57] the accuracy obtained (on a subset of 820 cases) with radiomics was 65.8 ± 1.50 against 74.2 ± 1.0 with deep features; in [17] the best model yielded an accuracy of 73.36 ± 1.95 using deep features. These results improved when clinical features were also considered: [57] obtained an accuracy of 76.9 ± 5.4, while [17] got 77.90 ± 1.27. The accuracy values reported by [57] are those obtained in the training/validation phase. Also in [5] deep architectures were proposed, with the best one giving sensitivity=0.79, specificity=0.82, and AUC=0.84. In summary, our results are promising and in line with the literature on the same dataset or on a subset [5], [17], [57].

Nevertheless, it is important to highlight the explainability and accuracy trade-off of our solution. In fact, in [57] and [17] deep learning approaches slightly improve performances compared with our model. However, deep features extracted by CNN do not guarantee a high level of explainability. From a clinical point of view is difficult to correlate the deep features learned with morpho-functional characteristics of a disease found by physicians. Through the use of intrinsically interpretable clinical and radiomic features, the proposed multi-level explanation improves the model's clinical validation.

In addition, as shown in Table 13, few works focus on explainable solutions. In [5], [17], and [57], saliency maps are used to realize explanation. We believe that the decrease in performance (compared with deep approaches proposed in [5], [17], [57]) obtained in our study is reasonable and justifies the choice of improving the explainability for a clinically compliant solution.

C. SHALLOW AND DEEP STRATEGIES

The main purpose of the study is to provide a globally and locally explainable model, able to clinically validate the model findings with the physician's support. This is achievable through the use of intrinsically intelligible input features and XAI methods for global and local explanation, such as the SHAP method. For this reason, we preferred to use radiomic features, rather than deep features, because it is well-known the meaning each feature expresses. In addition, radiomic features, combined with clinical and laboratory features (e.g., tabular data) allow us to draw clinical conclusions and interpret model findings [36]. In this scenario, shallow learning techniques are more appealing than deep learning methods. First, because they require fewer computational resources and data samples. Moreover, the explanation proves to be reliable and complete (both from a global and local point of view). Shallow learning algorithms - like RF and SVM - have been the standard for processing tabular data [14], [25], [48]. Indeed, in [57] it was demonstrated that shallow learning methods like SVM outperform their deep learning counterparts when tabular data are used (Tables 4 and 5).

The main key points that lead us to use shallow learning techniques instead of deep ones, are here summarized:

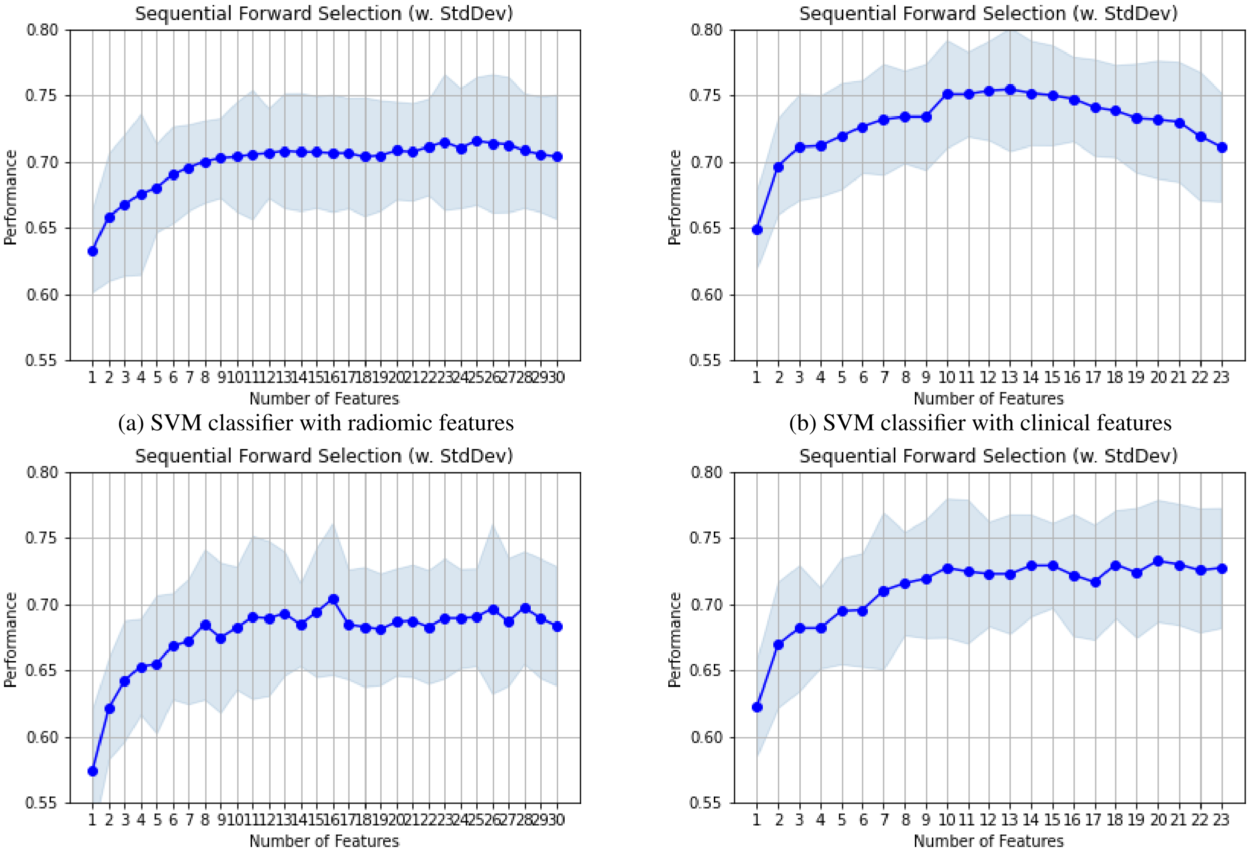
- *features interpretability and predictive model explainability:* typically in deep learning approaches, images are the input of the deep neural network architectures,

TABLE 13. Literature approaches and comparison. (*) In the *Dataset/Modality/Centers* column, the values between round parenthesis represent the number of images used for training, validation, and testing phases, respectively.

Reference	Task	Dataset (*) / Modality / Centers	Features	Method	Results	Explainability
Angeli *et al.* [4]	recovery *vs.* ICU or Death	301 / CT / 1	imaging, demographic, laboratory	LR	AUC=0.841	n.a.
Shiri *et al.* [54]	survival prediction	14339 / CT / 19	radiomic	LR, LASSO, LDA, RF, AdaBoost, Naïve Bayes, MLP	AUC=0.83, sens=0.81, spec=0.72	n.a.
Wang *et al.* [64]	aggravation *vs.* improvement	188 / CT / 1	radiomic, clinical	LR, SVM, DT, RF, XGBoost	AUC=0.843 (radiomic), AUC=0.813 (clinical), AUC=0.865 (combined)	n.a.
Xu *et al.* [68]	early, progressive, severe, or absorption stages	284 / CT / 1	radiomic	SVM	AUC=0.90	n.a.
Shi *et al.* [52]	infection severity	260 / CT / 3	clinical, laboratory, radiomic	LR multivariate	AUC=0.978	n.a.
Borghesi *et al.* [8]	recovery *vs.* death	100 / CXR / 1	Brixia score	weighted Kappa, Mann-Whitney U-test	kw=0.82	n.a.
Signoroni *et al.* [56]	Brixia score prediction	5000 / CXR / n.a.	deep	BS-Net	MAE=0.441	super-pixel maps
Soda *et al.* [57]	mild *vs.* severe	820 / CXR / 6	clinical, radiomic, deep	SVM, LR, RF, MLP, CNNs	acc=0.769±0.054	Grad-CAM
Barbano *et al.* [5]	COVID-positive *vs.* COVID-negative	451 (129+322) / CXR / 1	clinical, deep	fully-connected ANN	AUC=0.84	Grad-CAM
Guarrasi *et al.* [17]	mild *vs.* severe	1103 (820+283) / CXR / 6	clinical, deep	CNNs	acc=73.36±1.95 (only CXR), acc=77.61±1.10 (CXR+clinical)	feature importance, Grad-CAM
Proposed approach	mild *vs.* severe	1589 (820+283+486) / CXR / 6	clinical, radiomic	RF, SVM	AUC=0.819, acc=0.733	multi-level (SHAP analysis)

and it is the architecture itself that extracts features *via* convolution (i.e., learned features). Learned features are not directly intelligible and the meaning is unknown. Instead, by providing the radiomic features directly, we would know *a priori* the meanings of the input, improving the explainability aspect and allowing clinical introspection;

- *deep learning models explainability techniques are not robust and unreliable:* there are several methods to explain the features extracted *via* deep neural networks, focusing mainly on saliency maps computation. These methods aim to highlight the areas that most influence the model decision [34], [35]. Unfortunately, some experimental findings demonstrated unsatisfactory results in the clinical domain. For example, poorly localized, and spatially blurred visualization was found in some cases [40], [56]. In addition, it has been shown that the saliency map does not change even when adversarial attacks lead to incorrect model predictions [16]. Moreover, it has been shown that different saliency map computation techniques can produce conflicting results, as reported in [46] and [69]. Finally, these methods enable just a local explanation (i.e., for each specific patient), not allowing a global validation of the system decisions. For this reason, saliency maps have still to demonstrate to be an effective tool for validating clinical findings;
- *comparable performance:* results obtained from our model are comparable with deep approaches available in [5] and [57]. To evaluate this aspect, additional tests were performed using the architectures proposed in [5] and [57]. In particular, ResNet-11, ResNet-18, and ResNet-50 were trained - by exploiting Adam optimizer - using as input images resized to 224 × 224. The use of these deep architectures didn't allow relevant improvements over the proposed radiomic approach. The best accuracy obtained with ResNet-50 resulted lower than our SVM-based model. Moreover, with deep learning approaches, the clinical interpretation of the results - performed by means of saliency maps - would be reduced to visual/qualitative evaluation, which is a subjective and operator-dependent tool.



(a) SVM classifier with radiomic features

(b) SVM classifier with clinical features

(c) RF classifier with radiomic features

(d) RF classifier with clinical features

FIGURE 7. Accuracy trend obtained in the *preliminary selection* by SVM (in (a) and (b)) and RF (in (c) and (d)) classifiers, during the features selection, performed using SFFS algorithm: in (a) and (c) results on radiomic features; in (b) and (d) results on clinical features.

VI. CONCLUSION AND FUTURE DIRECTIONS

This work defined predictive models using clinical and radiomic features for COVID-19 prognosis prediction. Different ML classifiers and feature selection strategies were implemented, to optimally combine clinical, laboratory and radiomic features. In a clinical scenario, the findings have to be correct, effective, and also interpretable and clinically justifiable. Most of the literature works merely provide a set of features, focusing poorly on explaining the results. In this work, a multi-level clinical explanation was proposed through Explainable AI algorithms, considering the point of view of the stakeholders involved in the care process (i.e., developers, physicians, and patients). To make predictive models intrinsically explainable, they must be trained with explainable inputs. For this reason, the use of deep features (i.e., extracted through neural networks) was avoided. A global explanation was used for distributional drift detection and to clinically justify the behaviour of features that most influence classification. On the other hand, a local explanation was essential to make implementable the concept of a CDSS, which transparently gives the prediction and the explanation to the physician and patient.

Our work represents an attempt to implement an X-CDSS. The development of explainable radiomics-powered predictive models could accelerate their incorporation into personalized medicine [15], [27]. For this reason, it is essential in the future to realize large-scale studies using explainable ML models to define the associations between data and clinical outcomes. Providing reliable and explainable diagnostic and prognostic biomarkers for precision medicine is the ultimate goal of this research field.

APPENDIX A
PROVIDED CLINICAL AND LABORATORY DATA

This appendix provides the complete list of clinical and laboratory features associated with CXR images.

In particular, clinical data are: *Hospital*, *Age*, *Sex*, *Positivity at Admission*, *Temperature*, *Days of Fever*, *Cough*, *Difficulty in Breathing*, *Cardiovascular Disease*, *Ischemic Heart Disease*, *Atrial Fibrillation*, *Heart Failure*, *Ictus*, *High Blood Pressure*, *Diabetes*, *Dementia*, *Chronic Obstructive Bronchopneumopathy (BPCO)*, *Cancer*, *Chronic Kidney Disease*, *Respiratory Failure*, *Obesity*, *Position*, *Prognosis*, *Death*.

Instead, laboratory data are: *White Blood Cell (WBC)*, *Red Blood Cell (RBC)*, *C-Reactive Protein (CRP)*, *Fibrinogen*, *Glucose*, *Procalcitonin (PCT)*, *Lactate Dehydrogenase (LDH)*, *International Normalized Ratio (INR)*, *D-Dimer*, *Oxigen Percentage*, *Partial Pressure of Oxygen (PaO2)*, *Arterial Oxygen Saturation (SaO2)*, *Partial Pressure of Carbon Dioxide (PaCO2)*, *pH*.

APPENDIX B
PRELIMINARY FEATURE SELECTION: ACCURACY TRENDS

SFFS was used to select the best features subset. More details are provided in the following Figure 7, showing the accuracy trends obtained in the preliminary features selection results to evaluate the unimodal models (considering only clinical/laboratory features, and only radiomic features). The number of features maximizing the accuracy was considered.

REFERENCES

[1] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, "Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases," *Radiology*, vol. 296, no. 2, pp. 32–40, Aug. 2020.

[2] N. Ali, "Elevated level of C-reactive protein may be an early marker to predict risk for severity of COVID-19," *J. Med. Virol.*, vol. 92, no. 11, pp. 2409–2411, Nov. 2020.

[3] M. Amadasun and R. King, "Textural features corresponding to textural properties," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 5, pp. 1264–1274, Oct. 1989.

[4] E. Angeli, S. Dalto, S. Marchese, L. Setti, M. Bonacina, F. Galli, E. Rulli, V. Torri, C. Monti, R. Meroni, G. D. Beretta, M. Castoldi, and E. Bombardieri, "Prognostic value of CT integrated with clinical and laboratory data during the first peak of the COVID-19 pandemic in northern Italy: A nomogram to predict unfavorable outcome," *Eur. J. Radiol.*, vol. 137, Apr. 2021, Art. no. 109612.

[5] C. Alberto Barbano, E. Tartaglione, C. Berzovini, M. Calandri, and M. Grangetto, "A two-step explainable approach for COVID-19 computer-aided diagnosis from chest X-ray images," 2021, *arXiv:2101.10223*.

[6] A. Benfante, S. Principe, M. N. Cicero, M. Incandela, G. Seminara, C. Durante, and N. Scichilone, "Management of severe asthma during the first lockdown phase of SARS-CoV-2 pandemic: Tips for facing the second wave," *Pulmonary Pharmacol. Therapeutics*, vols. 73–74, Jun. 2022, Art. no. 102083.

[7] S. Bignardi, R. Sandhu, and A. Yezzi, "Radar-based shape and reflectivity reconstruction using active surfaces and the level set method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3617–3631, Mar. 2023.

[8] A. Borghesi and R. Maroldi, "COVID-19 outbreak in Italy: Experimental chest X-ray scoring system for quantifying and monitoring disease progression," *La Radiologia Medica*, vol. 125, no. 5, pp. 509–513, May 2020.

[9] C. Combi and G. Pozzi, "Health informatics: Clinical information systems and artificial intelligence to support medicine in the CoViD-19 pandemic," in *Proc. IEEE 9th Int. Conf. Healthcare Informat. (ICHI)*, Aug. 2021, pp. 480–488.

[10] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J. H. Moore, M. Zitnik, and J. H. Holmes, "A manifesto on explainability for artificial intelligence in medicine," *Artif. Intell. Med.*, vol. 133, Nov. 2022, Art. no. 102423.

[11] T. D. Do, S. Skornitzke, U. Merle, M. Kittel, S. Hofbaur, C. Melzig, H.-U. Kauczor, M. O. Wielpütz, and O. Weinheimer, "COVID-19 pneumonia: Prediction of patient outcome by CT-based quantitative lung parenchyma analysis combined with laboratory parameters," *PLoS ONE*, vol. 17, no. 7, Jul. 2022, Art. no. e0271787.

[12] ELI5 Website, *Eli5 Documentation*, 2022.

[13] M. M. Galloway, "Texture analysis using gray level run lengths," *Comput. Graph. Image Process.*, vol. 4, no. 2, pp. 172–179, Jun. 1975.

[14] M. M. Ghiasi and S. Zendehboudi, "Application of decision tree-based ensemble learning in the classification of breast cancer," *Comput. Biol. Med.*, vol. 128, Jan. 2021, Art. no. 104089.

[15] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016.

[16] J. Gu and V. Tresp, "Saliency methods for explaining adversarial attacks," 2019, *arXiv:1908.08413*.

[17] V. Guarrasi and P. Soda, "Multi-objective optimization determines when, which and how to fuse deep networks: An application to predict COVID-19 outcomes," *Comput. Biol. Med.*, vol. 154, Mar. 2023, Art. no. 106625.

[18] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Sep. 2019.

[19] Hackathon Website, *COVID CXR Hackathon Competition*, 2022.

[20] R. M. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE*, vol. 67, no. 5, pp. 786–804, May 1979.

[21] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[22] B. M. Henry, G. Aggarwal, J. Wong, S. Benoit, J. Vikse, M. Plebani, and G. Lippi, "Lactate dehydrogenase levels predict coronavirus disease 2019 (COVID-19) severity and mortality: A pooled analysis," *Amer. J. Emergency Med.*, vol. 38, no. 9, pp. 1722–1726, Sep. 2020.

[23] F. Innocenti, A. De Paris, A. Lagomarsini, L. Pelagatti, L. Casalini, A. Gianno, M. Montuori, P. Bernardini, F. Caldi, I. Tassinari, and R. Pini, "Stratification of patients admitted for SARS-CoV2 infection: Prognostic scores in the first and second wave of the pandemic," *Internal Emergency Med.*, vol. 17, no. 7, pp. 2093–2101, 2022.

[24] A. Jacobi, M. Chung, A. Bernheim, and C. Eber, "Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review," *Clin. Imag.*, vol. 64, pp. 35–42, Aug. 2020.

[25] S. Kabiraj, M. Raihan, N. Alvi, M. Afrin, L. Akter, S. A. Sohagi, and E. Podder, "Breast cancer risk prediction using XGBoost and random forest algorithm," in *Proc. 11th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2020, pp. 1–4.

[26] M. Khan, M. T. Mehran, Z. U. Haq, Z. Ullah, S. R. Naqvi, M. Ihsan, and H. Abbass, "Applications of artificial intelligence in COVID-19 pandemic: A comprehensive review," *Exp. Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115695.

[27] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong, J. van Timmeren, S. Sanduleanu, R. T. H. M. Larue, A. J. G. Even, A. Jochems, and Y. van Wijk, "Radiomics: The bridge between medical imaging and personalized medicine," *Nature Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 749–762, Dec. 2017.

[28] S. Leger, A. Zwanenburg, K. Pilz, F. Lohaus, A. Linge, K. Zöphel, J. Kotzerke, A. Schreiber, I. Tinhofer, V. Budach, and A. Sak, "A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling," *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, Oct. 2017.

[29] Z. Li, S. Zhao, Y. Chen, F. Luo, Z. Kang, S. Cai, W. Zhao, J. Liu, D. Zhao, and Y. Li, "A deep-learning-based framework for severity assessment of COVID-19 with CT images," *Exp. Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115616.

[30] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[31] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020.

[32] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.

[33] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 4765–4774.

[34] R. Margolin, L. Zelnik-Manor, and A. Tal, "Saliency for image manipulation," *Vis. Comput.*, vol. 29, no. 5, pp. 381–392, May 2013.

[35] R. Mechrez, E. Shechtman, and L. Zelnik-Manor, "Saliency driven image manipulation," *Mach. Vis. Appl.*, vol. 30, no. 2, pp. 189–202, Mar. 2019.

[36] C. Militello, F. Prinzi, G. Sollami, L. Rundo, L. La Grutta, and S. Vitabile, "CT radiomic features and clinical biomarkers for predicting coronary artery disease," *Cognit. Comput.*, vol. 15, no. 1, pp. 238–253, Jan. 2023.

[37] C. Militello, L. Rundo, M. Dimarco, A. Orlando, I. D'Angelo, V. Conti, and T. V. Bartolotta, "Robustness analysis of DCE-MRI-derived radiomic features in breast masses: Assessing quantization levels and segmentation agreement," *Appl. Sci.*, vol. 12, no. 11, p. 5512, May 2022.

[38] N. T. Nguyen, J. Chinn, M. De Ferrante, K. A. Kirby, S. F. Hohmann, and A. Amin, "Male gender is a predictor of higher mortality in hospitalized adults with COVID-19," *PLoS ONE*, vol. 16, no. 7, Jul. 2021, Art. no. e0254066.

[39] Q. Niu, X. Jiang, Q. Li, Z. Zheng, H. Du, S. Wu, and X. Zhang, "Texture features and pharmacokinetic parameters in differentiating benign and malignant breast lesions by dynamic contrast enhanced magnetic resonance imaging," *Oncol. Lett.*, pp. 4607–4613, Jul. 2018.

[40] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.

[41] E. K. Oikonomou, M. C. Williams, C. P. Kotanidis, M. Y. Desai, M. Marwan, A. S. Antonopoulos, K. E. Thomas, S. Thomas, I. Akoumianakis, L. M. Fan, and S. Kesavan, "A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CT angiography," *Eur. Heart J.*, vol. 40, no. 43, pp. 3529–3543, Nov. 2019.

[42] N. Papanikolaou, C. Matos, and D. M. Koh, "How to develop a meaningful radiomic signature for clinical use in oncologic patients," *Cancer Imag.*, vol. 20, no. 1, p. 33, Dec. 2020.

[43] *The Impact of the General Data Protection Regulation on Artificial Intelligence*, Eur. Parliament Directorate-General Parliamentary Res. Services, Eur. Parliament, 2021, doi: 10.2861/293.

[44] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Amsterdam, The Netherlands: Elsevier, 2011.

[45] S. Principe, A. Grosso, A. Benfante, F. Albicini, S. Battaglia, E. Gini, M. Amata, I. Piccionello, A. G. Corsico, and N. Scichilone, "Comparison between suspected and confirmed COVID-19 respiratory patients: What is beyond the PCR test," *J. Clin. Med.*, vol. 11, no. 11, p. 2993, May 2022.

[46] F. Prinzi, M. Insalaco, A. Orlando, S. Gaglio, and S. Vitabile, "A YOLO-based model for breast cancer detection in mammograms," *Cognit. Comput.*, pp. 1–14, Aug. 2023.

[47] F. Prinzi, C. Militello, V. Conti, and S. Vitabile, "Impact of wavelet kernels on predictive capability of radiomic features: A case study on COVID-19 chest X-ray images," *J. Imag.*, vol. 9, no. 2, p. 32, Jan. 2023.

[48] F. Prinzi, A. Orlando, S. Gaglio, M. Midiri, and S. Vitabile, "ML-based radiomics analysis for breast cancer classification in DCE-MRI," in *Proc. Int. Conf. Appl. Intell. Inform.* Cham, Switzerland: Springer, 2022, pp. 144–158.

[49] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[50] S. Ruano, G. Gallego, A. Yezzi, C. Cuevas, and N. García, "Robust image registration with global intensity transformation," in *Proc. Int. Symp. Consum. Electron. (ISCE)*, 2015, pp. 1–2.

[51] S. Raschka, "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack," *J. Open Source Softw.*, vol. 3, no. 24, p. 638, Apr. 2018.

[52] H. Shi, Z. Xu, G. Cheng, H. Ji, L. He, J. Zhu, H. Hu, Z. Xie, W. Ao, and J. Wang, "CT-based radiomic nomogram for predicting the severity of patients with COVID-19," *Eur. J. Med. Res.*, vol. 27, no. 1, p. 13, Dec. 2022.

[53] I. Shiri, H. Arabi, Y. Salimi, A. Sanaat, A. Akhavanallaf, G. Hajianfar, D. Askari, S. Moradi, Z. Mansouri, M. Pakbin, and S. Sandoughdaran, "COLI-Net: Deep learning-assisted fully automated COVID-19 lung and infection pneumonia lesion detection and segmentation from chest computed tomography images," *Int. J. Imag. Syst. Technol.*, vol. 32, no. 1, pp. 12–25, Jan. 2022.

[54] I. Shiri, Y. Salimi, M. Pakbin, G. Hajianfar, A. H. Avval, A. Sanaat, S. Mostafaei, A. Akhavanallaf, A. Saberi, Z. Mansouri, and D. Askari, "COVID-19 prognostic modeling using CT radiomic features and machine learning algorithms: Analysis of a multi-institutional dataset of 14,339 patients," *Comput. Biol. Med.*, vol. 145, Jun. 2022, Art. no. 105467.

[55] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, 1979.

[56] A. Signoroni, M. Savardi, S. Benini, N. Adami, R. Leonardi, P. Gibellini, F. Vaccher, M. Ravanelli, A. Borghesi, R. Maroldi, and D. Farina, "BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102046.

[57] P. Soda, N. C. D'Amico, J. Tessadori, G. Valbusa, V. Guarrasi, C. Bortolotto, M. U. Akbar, R. Sicilia, E. Cordelli, D. Fazzini, and M. Cellina, "AIforCOVID: Predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study," *Med. Image Anal.*, vol. 74, Dec. 2021, Art. no. 102216.

[58] J. Q. Su and J. S. Liu, "Linear combinations of multiple diagnostic markers," *J. Amer. Stat. Assoc.*, vol. 88, no. 424, pp. 1350–1355, Dec. 1993.

[59] C. Sun and W. G. Wee, "Neighboring gray level dependence matrix for texture classification," *Comput. Vis., Graph., Image Process.*, vol. 23, no. 3, pp. 341–352, Sep. 1983.

[60] G. Thibault, J. Angulo, and F. Meyer, "Advanced statistical matrices for texture characterization: Application to cell classification," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 3, pp. 630–637, Mar. 2014.

[61] M. J. Tobin, A. Jubran, and F. Laghi, "PaO_2/FIO_2 ratio: The mismeasure of oxygenation in COVID-19," *Eur. Respiratory J.*, vol. 57, Mar. 2021, Art. no. 2100274.

[62] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.*, vol. 77, no. 21, pp. 104–107, Nov. 2017.

[63] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2008, pp. 705–718.

[64] D. Wang, C. Huang, S. Bao, T. Fan, Z. Sun, Y. Wang, H. Jiang, and S. Wang, "Study on the prognosis predictive model of COVID-19 patients based on CT radiomics," *Sci. Rep.*, vol. 11, no. 1, p. 11591, Jun. 2021.

[65] D. S. Weld and G. Bansal, "The challenge of crafting intelligible intelligence," *Commun. ACM*, vol. 62, no. 6, pp. 70–79, May 2019.

[66] P. H. Westfall, J. F. Troendle, and G. Pennello, "Multiple McNemar tests," *Biometrics*, vol. 66, no. 4, pp. 1185–1191, Dec. 2010.

[67] Q. Wu, S. Wang, L. Li, Q. Wu, W. Qian, Y. Hu, L. Li, X. Zhou, H. Ma, H. Li, M. Wang, X. Qiu, Y. Zha, and J. Tian, "Radiomics analysis of computed tomography helps predict poor prognostic outcome in COVID-19," *Theranostics*, vol. 10, no. 16, pp. 7231–7244, 2020.

[68] Z. Xu, L. Zhao, G. Yang, Y. Ren, J. Wu, Y. Xia, X. Yang, M. Cao, G. Zhang, T. Peng, J. Zhao, H. Yang, J. Hu, and J. Du, "Severity assessment of COVID-19 using a CT-based radiomics model," *Stem Cells Int.*, vol. 2021, pp. 1–10, Sep. 2021.

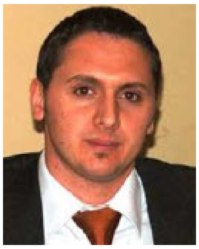
[69] J. Zhang, H. Chao, M. K. Kalra, G. Wang, and P. Yan, "Overlooked trustworthiness of explainability in medical AI," *medRxiv*, 2021.

[70] Q. Zhang, Y. Peng, W. Liu, J. Bai, J. Zheng, X. Yang, and L. Zhou, "Radiomics based on multimodal MRI for the differential diagnosis of benign and malignant breast lesions," *J. Magn. Reson. Imag.*, vol. 52, no. 2, pp. 596–607, Aug. 2020.

[71] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Berlin, Germany: Springer, 2018, pp. 3–11.

[72] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, and M. Bogowicz, "The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, 2020.

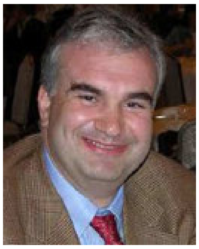
FRANCESCO PRINZI received the master's degree in computer science engineering from the University of Palermo, in 2020, where he is currently pursuing the Ph.D. degree in biomedical engineering with the BiND Department. He is also a Visiting Ph.D. Student with the University of Cambridge, U.K. His research interests include the development of diagnostic and predictive models through machine learning, explainable artificial intelligence, and medical imaging analysis methods. The main works aim at the development of data-driven systems to support the physician diagnostic process.



CARMELO MILITELLO is currently a Research Scientist with the Italian National Research Council (CNR). His research interests include medical imaging, radiomics, applied machine learning, clinical DSS, digital architectures, and hardware programmable devices. His research is focused on the analysis and processing of biomedical images, with a particular interest in the development of semi-automatic/automatic tools able to support the decision-making activities of clinicians. He has been the coordinator and scientific responsible for activities in different national and international research projects. He is the coauthor of more than 65 publications in international journals, book chapters, and conference proceedings.

NICOLA SCICHILONE is currently a Full Professor in respiratory diseases with the University of Palermo, Italy. He is also the Head of the Division of Respiratory Diseases, Department of Health Promotion Sciences, Maternal and Infant Care, Internal Medicine and Medical Specialties (PROMISE) and the Director of the COVID-19 Unit, University of Palermo.

SALVATORE GAGLIO (Life Member, IEEE) received the Graduate degree in electronic engineering from the University of Genoa, Italy, in 1977, and the M.S.E.E. degree from the Georgia Institute of Technology, Atlanta, USA, in 1978. From 1986 he is professor of computer science and artificial intelligence at the University of Palermo, Italy. From 2007 he is the scientific responsible for the Ph.D. program in Computer Science and Technological Innovation at the University of Palermo. From 2015 to 2018 he has been the President of the Italian National Academy of Sciences, Humanities, and Arts of Palermo. He has been member of various committees for projects of national interest in Italy and he is referee of various scientific congresses and journals. His present research activities are in the area of artificial intelligence and robotics.



SALVATORE VITABILE is a Full Professor with the Department of Biomedicine, Neuroscience and Advanced Diagnostics at the University of Palermo, Italy. He is co-author of more than 200 scientific papers in referred journals and conferences. He has chaired, organized, and served as member of the organizing committee of several international conferences and workshops. He is an Associate Editor of the Human-centric Computing and Information Sciences journal and an Editorial Board Member of Electronics. His research interests include medical data processing and analysis, clinical decision support systems, specialized architecture design and prototyping, and machine and deep learning applications.

Open Access funding provided by 'Consiglio Nazionale delle Ricerche-CARI-CARE-ITALY' within the CRUI CARE Agreement