A Deep Learning Radiomics Model to Identify Poor Outcome in COVID-19 Patients With Underlying Health Conditions: A Multicenter Study

Siwen Wang, Di Dong, Liang Li, Hailin Li, Yan Bai, Yahua Hu, Yuanyi Huang, Xiangrong Yu, Sibin Liu, Xiaoming Qiu, Ligong Lu, Meiyun Wang, Yunfei Zha, and Jie Tian, *Fellow, IEEE*

Abstract—Objective: **Coronavirus disease 2019 (COVID-19) has caused considerable morbidity and mortality, especially in patients with underlying health conditions. A precise prognostic tool to identify poor outcomes among such cases is desperately needed.** ***Methods:*** **Total 400 COVID-19 patients with underlying health conditions were retrospectively recruited from 4 centers, including 54 dead cases (labeled as poor outcomes) and 346 patients discharged or hospitalized for at least 7 days since initial CT scan. Patients were allocated to a training set ($n = 271$), a test set ($n = 68$), and an external test set ($n = 61$). We proposed an initial CT-derived hybrid model by combining a 3D-ResNet10 based deep learning model and a quantitative 3D radiomics model to predict the probability of COVID-19 patients reaching poor outcome. The model performance was assessed by area under the receiver operating characteristic curve (AUC), survival analysis, and subgroup analysis.** ***Results:*** **The hybrid model achieved AUCs of 0.876 (95% confidence interval: 0.752-0.999) and 0.864 (0.766-0.962) in test and external test sets, outperforming other models. The survival analysis verified the hybrid model as a significant risk factor for mortality (hazard ratio, 2.049 [1.462–2.871], $P < 0.001$) that could well stratify patients into high-risk and low-risk of reaching poor outcomes ($P < 0.001$).** ***Conclusion:*** **The hybrid model that combined deep learning and radiomics could accurately identify poor outcomes in COVID-19 patients with underlying health conditions from initial CT scans. The great risk stratification ability could help alert risk of death and allow for timely surveillance plans.**

***Index Terms*—COVID-19, deep learning, radiomics, prognosis, computed tomography.**

I. Introduction

THE PANDEMIC coronavirus disease 2019 (COVID-19) is evolving worldwide and has brought about considerable morbidity and mortality, especially in patients with underlying health conditions [1]. Many studies have reported that 25–46% of COVID-19 patients had underlying health conditions, with hypertension (15–31%), diabetes (8–20%), and cardiovascular diseases (15–54%) being the most common [2]–[4]. More importantly, COVID-19 patients with underlying diseases were closely related to severer status and poorer prognosis [4]–[6]. However, most previous studies either estimated the prevalence of underlying diseases in COVID-19 patients or analyzed the

Manuscript received August 31, 2020; revised December 13, 2020, February 25, 2021, and March 31, 2021; accepted April 24, 2021. Date of publication April 27, 2021; date of current version July 20, 2021. This work was supported in part by the National Key R&D Program of China under Grants 2017YFC1309100 and 2017YFA0205200, in part by the National Natural Science Foundation of China under Grants 82022036, 91959130, 81971776, 81771924, 81930053, and 81527805, in part by Beijing Natural Science Foundation (L182061), in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDB 38040200, in part by the Chinese Academy of Sciences under Grants GJJSTD20170004 and QYZDJ-SSW-JSC005, and in part by the Youth Innovation Promotion Association CAS 2017175. *(Siwen Wang, Di Dong, and Liang Li contributed equally to this work.) (Corresponding authors: Yunfei Zha; Jie Tian.)*

Siwen Wang and Di Dong are with the CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wangsiwen2017@ia.ac.cn; di.dong@ia.ac.cn).

Liang Li and Yunfei Zha are with the Department of Radiology, Renmin Hospital of Wuhan University, Wuhan 430060, China (e-mail: liliang_082@163.com; zhayunfei999@126.com).

Hailin Li and Jie Tian are with the CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing 100191, China (e-mail: hailin_li@buaa.edu.cn; tian@ieee.org).

Yan Bai and Meiyun Wang are with the Department of Medical Imaging, Henan Provincial People's Hospital & the People's Hospital of Zhengzhou University, Zhengzhou 450003, China (e-mail: resonance2010@126.com; mywang@ha.edu.cn).

Yahua Hu and Xiaoming Qiu are with the Department of Radiology, Huangshi Central Hospital, Affiliated Hospital of Hubei Polytechnic University, Edong Healthcare Group, Huangshi 435000, China (e-mail: huyahua.8888@126.com; 120003481@qq.com).

Yuanyi Huang and Sibin Liu are with the Department of Radiology, Jingzhou Central Hospital, Jingzhou 434020, China (e-mail: yyhuangjz@163.com; liusib9159@qq.com).

Xiangrong Yu is with the Department of Medical Imaging, Zhuhai People's Hospital, Zhuhai Hospital Affiliated with Jinan University, Zhuhai 519000, China (e-mail: yxr00125040@126.com).

Ligong Lu is with the Zhuhai Interventional Medical Center, Zhuhai Precision Medical Center, Zhuhai Hospital Affiliated with Jinan University, Zhuhai 519000, China (e-mail: luligong1969@126.com).

This article has supplementary downloadable material available at https://doi.org/JBHI.2021.3076086, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2021.3076086

association of underlying diseases with prognosis based on case-control schemes [7]–[9]. Few studies focused on developing precise prognostic tools for such population.

Radiological findings from chest computed tomography (CT) provide crucial information for the diagnosis of COVID-19 within the tense clinical settings [10], [11]. Patients who suffered from COVID-19 pneumonia tend to exhibit CT abnormalities characterized by multiple lobular ground-glass opacity (GGO), subsegmental areas of consolidation, and bilateral involvement [12]. Some attempts have also been made to capture disease progression and monitor pneumonia changes by serial CT scans [13]–[15], prompting the value of CT in prognosis management. Notwithstanding, the development of CT-based prognostic tools remains scant.

Inspirationally, artificial intelligence (AI) techniques, including deep learning and radiomics methods, have revealed great success in diagnosis and prognosis by mining knowledge from medical images [16], [17]. Particularly, researchers have suggested the image analysis for COVID-19 by advanced AI methods to develop effective diagnostic and prognostic models [18], [19]. In response to the call, Yue et al. [20] used CT radiomics features to predict COVID-19 patients' hospital stay. Wu et al. [21] developed a clinic-radiomics signature to predict COVID-19 patients with poor outcomes. Liang et al. [22] built a deep learning survival model by fitting clinical features to estimate COVID-19 patients' risk of developing critical illness. Their models achieved good performance, however, failed to take full advantages of deep learning and radiomics. Moreover, radiomics methods quantify the image features from the entire lung level, while deep learning features mostly focus on the local information of lung lesions. Thus, a combination of deep learning and radiomics may help evaluate the image features of COVID-19 patients more comprehensively.

Therefore, this study aimed to develop an initial CT-derived deep learning radiomics model to identify poor outcomes in COVID-19 patients with underlying health conditions. The risk stratification ability of the model was also explored to ensure timely prognosis surveillance as well as appropriate health resource allocations within the tense clinical settings.

II. Materials and Methods

A. Study Participants

A total of 400 COVID-19 patients with underlying health conditions were retrospectively enrolled from 4 centers between January 6, 2020 and March 6, 2020, including 339 patients from Renmin Hospital of Wuhan University, 29 patients from Huangshi Central Hospital, 22 patients from Jingzhou Central Hospital, and 10 patients from Henan Provincial People's Hospital. This study was approved by the institutional review boards at the 4 hospitals, and informed consent was waived. The inclusion criteria were: (1) laboratory-confirmed COVID-19 pneumonia by real-time reverse transcriptase-polymerase chain reaction (RT-PCR) tests; (2) one or multiple underlying health conditions based on self-reported records at admission; (3) available initial non-contrast enhanced chest CT scan; (4) a definite primary outcome (dead, discharged or hospitalized) wherein the hospitalized patient should have a follow-up duration for at least 7 days since initial CT scan. We excluded patients with poor CT image quality.

The clinical characteristics, including age, sex, underlying health conditions, survival status, and survival time, were collected from medical records (Table I). Underlying health conditions included hypertension, diabetes, coronary heart disease, chronic pulmonary disease, carcinoma, cerebrovascular disease, chronic heart disease, chronic renal disease, chronic liver disease, post operation, etc. Two outcomes were measured in this study. The primary outcome was survival status, which was defined as death (non-survivor) and discharge or hospitalization (survivor). The secondary outcome was survival time, which was recorded from initial CT scan date to the date of death or discharge, or the latest date the patient was monitored during hospitalization.

For the subsequent model development, patients from Renmin Hospital of Wuhan University (Wuhan dataset) were randomly assigned to a training set (n = 271) and a test set (n = 68) at a ratio of 4:1. Patients from the other three centers constituted an external test set (n = 61). The models were constructed based on the training set and finally evaluated on the test set and external test set.

B. Lung Volume Segmentation and Preprocessing

All patients underwent chest CT scans at admission. The initial CT scans were exported from the standard picture archiving and communication system.

In this study, we first adopted an automatic lung volume segmentation scheme via the threshold segmentation and flood fill algorithm [23]. The binarization for initial chest CT image was first done by a threshold of Hounsfield Units value of -300 to get a preliminary lung shape. Then, the flood fill algorithm searched and detected all the nodes that were connected to the given seed nodes within the lung region by a path in the three-dimensional (3D) array, generating the connected lung domains. Herein, the closing operation method was used to denoise and keep the maximal connected domain. Then, we calculated the number of connected domains in each image slice, with which the lung volume and non-lung area (i.e., background and other organs) were finally formed, and the segmented lung volume and the corresponding lung mask were both acquired.

For the deep learning model development, the segmented lung volume was resized to a 3D volume of interest (VOI) with the size of $48\times240\times360$ [24], of which 24 slices were center cropped and used as the inputs. For the radiomics model development, a 20-layer 3D VOI was chosen with the image slice that had the largest area of lung mask as the center. Then, the VOIs were normalized by the min-max normalization method to minimize the influence of voxel distribution and contrast variation. The overall workflow of this study is shown in Fig. 1.

C. Deep learning Model Construction and Training

We first developed an end-to-end deep learning model to predict the probability of reaching poor outcomes. We utilized

TABLE I
CLINICAL CHARACTERISTICS OF PATIENTS

Clinical characteristics	Training set (271)		*P*	Test set (68)		*P*	External test set (61)		*P*
	Non-survivor (40)	Survivor (231)		Non-survivor (10)	Survivor (58)		Non-survivor (4)	Survivor (57)	
Age, mean ± SD, years	73.0 ± 13.8	63.3 ± 12.7	<0.001*	66.0 ± 14.0	63.5 ± 12.3	0.768	72.0 ± 20.1	58.6 ± 14.1	0.103
Sex, No. (%)			0.182			0.597			0.883
Male	15 (37.5%)	113 (48.9%)		7 (70.0%)	32 (55.2%)		3 (75.0%)	33 (57.9%)	
Female	25 (62.5%)	118 (51.1%)		3 (30.0%)	26 (44.8%)		1 (25.0%)	24 (42.1%)	
Underlying health conditions, No. (%)			<0.001*			0.461			0.643
1	14 (35.0%)	158 (68.4%)		8 (80.0%)	36 (62.1%)		2 (50.0%)	35 (61.4%)	
⩾2	26 (65.0%)	73 (31.6%)		2 (20.0%)	22 (37.9%)		2 (50.0%)	22 (38.6%)	
Survival time, median (IQR), days	11.5 (7.0-17.75)	23.0 (14.0-33.0)	<0.001*	6.0 (3.5-9.5)	23.0 (18.0-30.5)	<0.001*	14.0 (3.0-18.0)	19.0 (14.0-29.0)	0.046*

NOTE. *P* values were calculated to assess the differences of clinical characteristics between non-survivors and survivors. For continuous variables, Mann-Whitney U test was used. For categorical variables, Chi-square test or Fisher exact test was used, as appropriate. * represents a significant difference. SD, standard deviation; IQR, interquartile range.

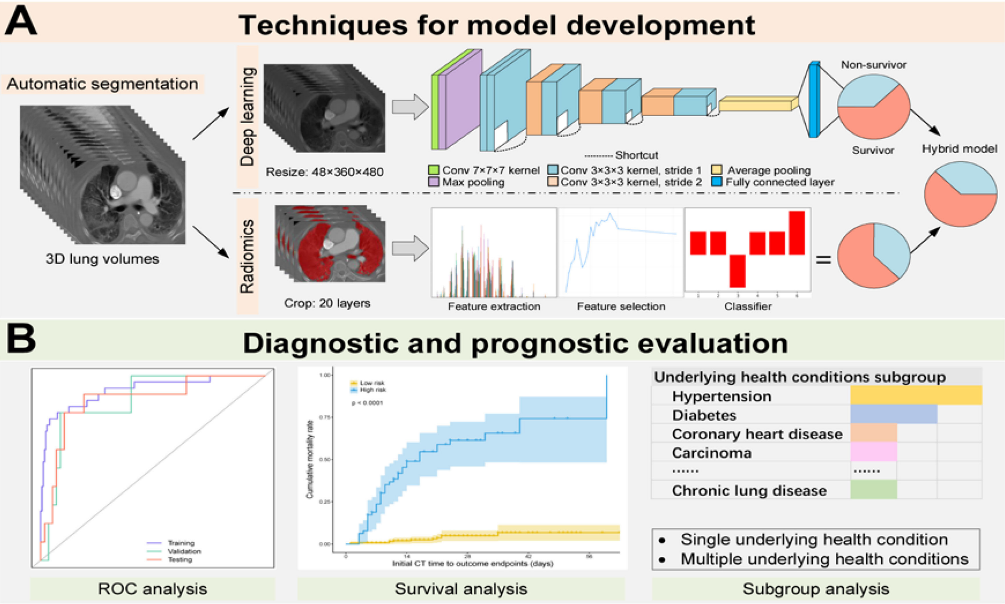

Fig. 1. The overall workflow of this study. (A) Techniques for model development, including automatic lung volume segmentation, deep learning model construction, and radiomics model construction. The deep learning model was developed based on a 3D-ResNet10 architecture. The radiomics workflow involved feature extraction, feature selection, and classifier construction. The model probabilities were combined to build a hybrid model. (B) Diagnostic and prognostic evaluation, including ROC curve analysis, survival analysis, and subgroup analysis.

the deep residual network in a 3D fashion as the backbone. Specifically, our model was designed as a 10-layer architecture (3D-ResNet10), mainly comprising 4 residual blocks ended with shortcut connections, mimicking the computational neurons and links in brains (Appendix Table S1). Under the 3D-ResNet10 settings, the model shared a similar architecture to 2D-ResNet18 [25], but had better feature representation and explored more dimensional knowledge through 3D convolution kernels. Meanwhile, the model has fewer parameters compared with the commonly used 3D-ResNet18 and 3D-ResNet34, and may be more suitable for a relatively small dataset.

In detail, each residual block was stacked with multiple convolution layers followed by batch normalization layers and ReLU activation layers to avoid gradient vanishing during the training process. The shortcut connection was used to combine the information of two distant convolution layers, enhancing the gradient flow in the network. The output of the average pooling layer was a 1D vector with the length of 512, representing the deep learning features learned from the input lung volume. The final output after the fully connected layer and the SoftMax function was the predicted probability of reaching poor outcome, which we termed as deep learning model probability. Here, the

SoftMax function restricted the deep learning model probability to a range of 0 to 1.

The model was trained on the training set with an initial learning rate of 2e-6 for 100 epochs, and then finetuned with a learning rate of 2e-7 for 50 epochs. We adopted cross-entropy loss function and stochastic gradient descent optimizer. The batch size was set to 16, and the weight decay was 0.01. The training process was realized in Python using PyTorch (version 1.3.0) and performed on an NVIDIA Titan RTX Graphics Card.

D. Radiomics Model Development

A radiomics model was also built in radiomics fashion to predict the probability of reaching poor outcomes. To quantitatively measure the intensity distribution of lung volumes, predefined first-order statistical features were extracted from each 3D VOI by using algorithms provided in PyRadiomics (version v3.0). The image types included original image, Laplacian of Gaussian (LoG) filtered image, and wavelet filtered image. Thereinto, the LoG filter was an edge enhancement filter using sigma parameters of 1.0, 3.0, and 5.0. The wavelet filter yielded 8 decompositions by applying either a High (H) or a Low (L) pass filter in each of the 3 dimensions, including LLH, LHL, LHH, HLL, HLH, HHL, HHH, and LLL [26]. The feature types included maximum, minimum, mean, median, range, interquartile range, 10 percentile, 90 percentile, entropy, skewness, kurtosis, mean absolute deviation, robust mean absolute deviation, root mean squared, energy, total energy, uniformity, and variance. All the extracted features were normalized.

Feature selection was then conducted to find the optimal feature subset. We adopted the recursive feature elimination (RFE) algorithm incorporating 10-fold cross-validation (Algorithm 1). The RFE is a backwards feature selection method [27], and the 10-fold cross-validation was used as the outer resampling method [28], [29]. In practice, the classification performance of feature subsets was measured according to a Kappa metric [30], and we finally selected the feature subset with the highest Kappa value.

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

$$p_o = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

$$p_e = \frac{(TP + FN) \times (TP+FP)+(FP+TN) \times (FN+TN)}{(TP+FP+FN+TN)^2} \tag{3}$$

where non-survivors are denoted as positive and survivors as negative in the binary classification task, TP (True Positive) represents a positive sample (non-survivor) correctly classified, TN (True Negative) represents a negative sample (survivor) correctly classified, FP (False Positive) represents that a negative sample (survivor) is wrongly classified as positive (non-survivor), and FN (False Negative) represents that a positive sample (non-survivor) is wrongly classified as negative (survivor).

Algorithm 1: Recursive Feature Elimination Incorporating 10-fold Cross-Validation.

```
for Each fold iteration do
    Partition data into training and test samples via resampling
    Train the model on training samples using all the N extracted radiomics features
    Predict the model outputs on test samples
    Calculate feature importance and rank features by importance
    for Each subset size S = 1, 2, 3 … N do
        Keep the S top ranked radiomics features
        Train the model on training samples using the S radiomics features
        Predict the model outputs on test samples
    end for
end for
Calculate the Kappa performance over the S radiomics features using test samples
Determine the appropriate number of features according to the highest Kappa value
Use a consensus ranking to determine the final features to retain in the final model
```

Then, the selected features by RFE were fed to a logistic regression, commonly used in binary classification tasks, to fit a mathematical formula for calculating the probability of reaching poor outcome, which we called the radiomics model probability. Here, the sigmoid function of the logistic regression made the radiomics model probability between 0 and 1.

E. Hybrid Model Development and Evaluation

To explore the potential compound value of the end-to-end deep learning model and the quantitative radiomics model, we proposed a hybrid model that ensembled the deep learning model probability and radiomics model probability via a multivariate logistic regression. The prediction here was called the hybrid model probability. Likewise, the hybrid model probability was restricted to a range of 0 to 1 by the sigmoid function of the logistic regression. As some clinical characteristics have been reported to be risk factors for COVID-19 patients, we also built a clinical non-imaging model based on clinical characteristics (age, sex, and single/multiple underlying health conditions) for comparison and to investigate the advantage of using imaging data.

We evaluated the 4 models with regard to: (1) Area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, specificity, and accuracy, with the optimal classification thresholds determined based on the training set according to Youden index (or Youden's J statistic) [31]. The Youden index is defined as:

$$J = Sensitivity + Specificity - 1 \tag{4}$$

And, we reported the optimal classification threshold that got the maximal Youden index in this study:

$$J_{max} = max_t \; \{Sensitivity\,(t) + Specificity\,(t) - 1\}, \tag{5}$$

where *t* denotes the optimal classification threshold for which J is maximal. (2) Gradient-weighted class activation maps (Grad-CAM) from the last convolution layer of the deep learning model that visualize the model attention on lung lesions [32]. We depicted the Grad-CAMs for two typical cases (1 non-survivor and 1 survivor) and superimposed the maps on the original CT images. (3) Descriptions of the radiomics features embedded in the radiomics model. We provided a calculation formula for the radiomics model, and tried to discuss the clinical relevance of the radiomics features.

In prognosis analysis, the survival curves illustrating the time-dependent cumulative probability of reaching poor outcomes were plotted [33]. We also adopted Cox proportional hazards regression to estimate the hazard ratio (HR) of the hybrid model to compare with age, sex, and single/multiple underlying health conditions for identifying whether the hybrid model was an independent risk factor for mortality.

Given that several underlying health conditions were involved in this study and many patients had two and more underlying health conditions, single underlying health condition subgroup and multiple underlying health conditions subgroup were analyzed to further verify the risk stratification ability of the hybrid model.

F. Statistical Analysis

To evaluate the effectiveness and robustness of the deep learning model, radiomics model, and hybrid model, we conducted 5-fold cross-validation based on the Wuhan dataset. We also did this for the clinical non-imaging model.

The statistical analysis was conducted with Python (version 3.6.9; https://www.python.org/) and R (version 3.5.3; https://www.r-project.org/). The R packages used in this study were summarized in Appendix Text S1. To measure statistical differences, Mann-Whitney U test was used for continuous variables, and Chi-square test or Fisher exact test was used for categorical variables. The log-rank test and the Gehan-Breslow-Wilcoxon test were both used to compare the survival curves. The AUC and HR were reported with 95% confidence interval (95% CI). The 5-fold cross-validation results were reported with mean ± standard deviation (SD). A 2-sided $P < 0.05$ represents a statistically significant difference.

III. Results

A. Patient Characteristics

A total of 400 COVID-19 patients with underlying health conditions were finally included from 4 centers. The mean age was 63.8 years (SD, 13.5 years), and 203 patients (50.8%) were male. Hypertension was the most common underlying health condition (250 patients, 62.5%), followed by diabetes (88 patients, 22.0%) and coronary heart disease (42 patients, 10.5%). A relatively large proportion of patients (147 patients, 36.8%) had more than one underlying health conditions. There were 54 patients dead, 308 patients discharged, and 38 patients still hospitalized. No significant difference was captured between non-survivors and survivors in sex ($P = 0.482$), whereas age ($P < 0.001$) and number of underlying health conditions ($P = 0.002$) differed significantly. Also, the survival time was significantly different (non-survivors, 10.0 days; IQR, 5.3–17.0 days vs. survivors, 22.0 days; IQR, 15.0–31.0 days; $P < 0.001$).

B. Training and Visualization of Deep Learning Model

In the test set, the deep learning model achieved an AUC of 0.759 (95% CI, 0.573–0.944) in identifying poor outcomes, which was further verified in the external test set with an AUC of 0.746 (95% CI, 0.458–1.000) (Fig. 2). When the deep learning model was well trained, we could calculate the gradient of the predicted value, which could inform us how the model responded to the lung lesion changes within the VOI. If the deep learning model learned the abstract mappings between high-order features and the primary outcomes and could discriminate non-survivors from survivors, different model response patterns would be demonstrated in Grad-CAM. Hence, visualizing these gradients may help interpret the attention of the deep learning model. Assisted by the Grad-CAM, we discovered some high responses in pneumonia area (Fig. 3), which may suggest high risk of reaching poor outcome for the patient. We have uploaded the deep learning model online for open access (please see http://www.radiomics.net.cn/post/136).

C. Development and Assessment of Radiomics Model

A total of 216 predefined first-order statistical features were extracted. The parameter file for radiomics feature extraction was uploaded online (please refer to http://www.radiomics.net.cn/post/136). After normalization, the RFE algorithm identified a potential feature subset of 11 features, which were then fitted by a multivariate logistic regression. Detailed descriptions of the selected radiomics features were given in Appendix Text S2. Finally, the reserved features weighted by corresponding regression coefficients generated a radiomics model formula to predict the probability of poor outcome. The constructed 11-feature radiomics model exhibited good classification performance, yielding AUCs of 0.872 (95% CI, 0.817–0.927) and 0.855 (95% CI, 0.744–0.966) in the training and test set.

We calculated the radiomics model probability from a sigmoid function on a logit scale. The calculation formula is: radiomics model probability = sigmoid (–2.485 – 0.714 × wavelet.HLL_firstorder_MeanAbsoluteDeviation + 0.175 × log.sigma.1.0.mm.3D_firstorder_InterquartileRange + 1.757 × log.sigma.1.0.mm.3D_firstorder_10Percentile + 3.321 × log.sigma.1.0.mm. 3D_ firstorder_RobustMeanAbsoluteDeviation + 0.018 × log.sigma.3.0.mm.3D_firstorder_10Percentile + 11.707 × wavelet.HHH_firstorder_InterquartileRange + 0.319 × wavelet.HLH_firstorder_Uniformity – 0.242 × log.sigma.5.0.mm.3D_firstorder_Entropy – 11.508 × wavelet.HHH_firstorder_RobustMeanAbsoluteDeviation – 0.950 × wavelet.LLL_firstorder_Kurtosis + 1.171 × log.sigma.5.0.mm.3D_firstorder_10Percentile).

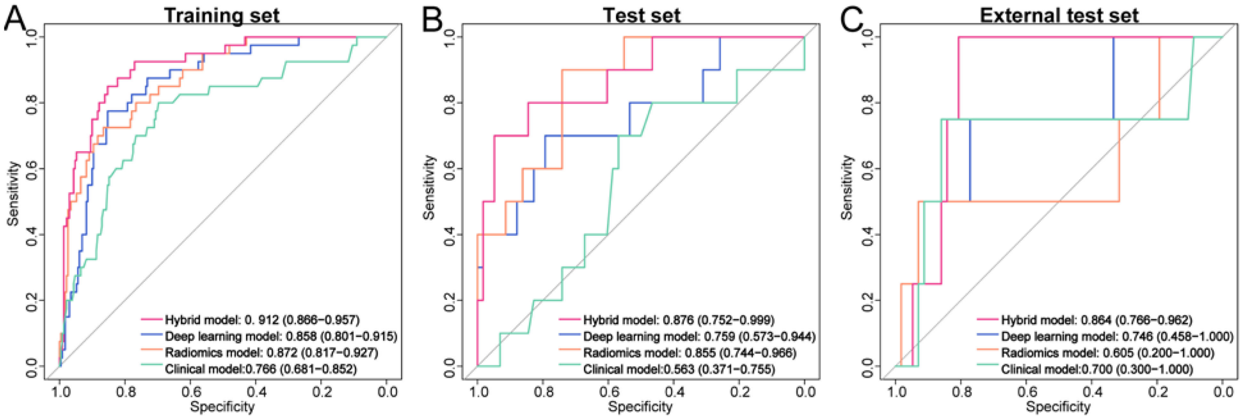

Fig. 2. Receiver operating characteristic curves of the hybrid model, deep learning model, radiomics model, and clinical non-imaging model in the three sets.

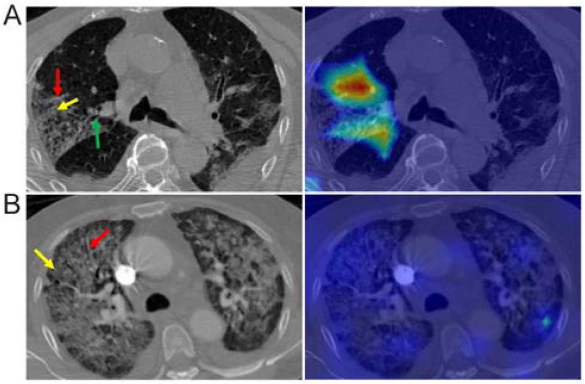

Fig. 3. Visualization of deep learning model attention. (A) A dead case: an 89-year-old male patient with single underlying health condition. (B) A discharged case: a 78-year-old female patient with multiple underlying health conditions. The yellow arrows represent GGO, the red arrows represent tractive vasodilation, and the green arrow represents tractive bronchiectasis. The highlighted areas indicated that the deep learning model was more sensitive to image features of patients with high risk of death, while less sensitive to patients with low risk of death.

D. Diagnostic Performance of Hybrid Model

The hybrid model, taking advantages of the deep learning model and radiomics model predictions, outperformed other models in identifying poor outcomes, with an AUC of 0.876 (95% CI, 0.752–0.999) in the test set. The sensitivity, specificity, and accuracy were 0.700, 0.845, and 0.824, respectively (Table II). This was further verified in external test set with an AUC of 0.864 (95% CI, 0.766–0.962). The sensitivity, specificity, and accuracy in external test set were 0.750, 0.842, and 0.836, respectively. The multivariate logistic regression results for hybrid model are demonstrated in Table III. Furthermore, the 5-fold cross-validation results proved that our hybrid model performed best and showed robustness even based on a relatively small dataset (Table IV). The ROC curves for each fold in cross-validation are plotted in Appendix Fig. S1 with the mean ROC curves and standard deviation regions highlighted.

TABLE II
DIAGNOSTIC PERFORMANCE OF MODELS

Models / Sets	AUC (95% CI)	SEN	SPE	ACC
Clinical non-imaging model				
Training set	0.766 (0.681-0.852)	0.800	0.697	0.712
Test set	0.563 (0.371-0.755)	0.800	0.345	0.412
External test set	0.700 (0.300-1.000)	0.750	0.702	0.705
Deep learning model				
Training set	0.858 (0.801-0.915)	0.775	0.853	0.841
Test set	0.759 (0.573-0.944)	0.600	0.828	0.794
External test set	0.746 (0.458-1.000)	0.500	0.807	0.787
Radiomics model				
Training set	0.872 (0.817-0.927)	0.725	0.866	0.845
Test set	0.855 (0.744-0.966)	0.600	0.845	0.809
External test set	0.605 (0.200-1.000)	0.500	0.912	0.885
Hybrid model				
Training set	0.912 (0.866-0.957)	0.850	0.853	0.852
Test set	0.876 (0.752-0.999)	0.700	0.845	0.824
External test set	0.864 (0.766-0.962)	0.750	0.842	0.836

NOTE. AUC, area under the curve; CI, confidence interval; SEN, sensitivity; SPE, specificity; ACC, accuracy.

TABLE III
MULTIVARIATE LOGISTIC REGRESSION FOR HYBRID MODEL

Index	Coefficient	Multivariate *P* value
deep learning model probability	13.708	< 0.0001*
radiomics model probability	4.823	< 0.0001*
Intercept	-4.970	< 0.0001*

NOTE. Hybrid model probability = sigmoid (13.708 × deep learning model probability + 4.823 × radiomics model probability – 4.970). * represents a statistical significance.

TABLE IV
5-FOLD CROSS-VALIDATION RESULTS OF DIFFERENT MODELS

Sets / Evaluation Metrics		Deep learning model	Radiomics model	Hybrid model	Clinical non-imaging model
Training set	AUC	0.835±0.032	0.866±0.005	**0.903±0.012**	0.704±0.035
	Accuracy	0.796±0.038	0.827±0.021	**0.839±0.049**	0.691±0.082
	Sensitivity	0.750±0.057	0.810±0.051	**0.860±0.058**	0.690±0.096
	Specificity	0.805±0.046	0.830±0.030	**0.836±0.066**	0.691±0.109
Test set	AUC	0.783±0.017	0.831±0.038	**0.869±0.029**	0.703±0.098
	Accuracy	0.790±0.048	0.814±0.022	**0.832±0.023**	0.608±0.158
	Sensitivity	0.640±0.049	0.680±0.117	0.680±0.133	**0.700±0.179**
	Specificity	0.816±0.060	0.837±0.042	**0.858±0.036**	0.592±0.213

NOTE. The model performances were presented as mean ± standard deviation. The best results were shown in bold. AUC, area under the curve.

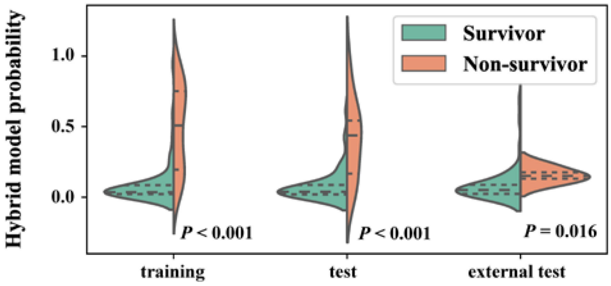

Fig. 4. There were significant differences in the hybrid model score distribution between non-survivors and survivors. The hybrid model showed good discriminative ability.

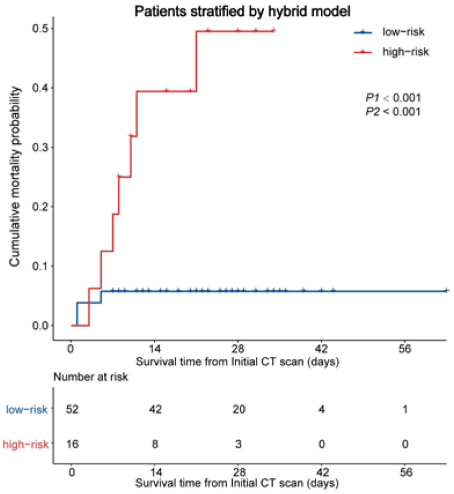

Fig. 5. Risk stratification ability of the hybrid model verified in the test set. Patients could be stratified into high-risk and low-risk of reaching poor outcomes by the classification threshold of hybrid model.

We also built a feature hybrid model for comparison by fusing features extracted from the deep learning model and the radiomics model. Specifically, we froze the weights of convolution layers in the deep learning model and concatenated the 11 features from the radiomics model with the extracted 512 deep learning features. Then, the 523 features were fed to fully connected layers and trained for 100 epochs with a learning rate of 5e-5 and a batch size of 16. The feature hybrid model achieved an AUC of 0.786 (95% CI, 0.614–0.958) in the test set, and the sensitivity, specificity, and accuracy were 0.700, 0.759, and 0.750, respectively, performing worse than the prediction hybrid model.

Also, the hybrid model probabilities showed significant differences between non-survivors and survivors in both test set (median, 0.438; IQR, [0.167, 0.544] vs. median, 0.038; IQR, [0.019, 0.087]; $p < 0.001$) and external test set (median, 0.151; IQR, [0.131, 0.176] vs. median, 0.051; IQR, [0.023, 0.087]; $p = 0.016$) (Fig. 4). The significant P values did identify the good classification ability of the hybrid model.

E. Prognostic Analysis of Hybrid Model

To further assess the prognostic value of the hybrid model, survival analysis was performed to assess the risk stratification ability. In the test set, the median survival time was 22.0 (IQR, 12.8–29.0) days. Using the classification threshold of the hybrid model, patients were stratified into high-risk group (survival time [IQR], 13.5 [8.0 to 23.0] days; hybrid model probability [IQR], 0.400 [0.224 to 0.532]) and low-risk group (survival time [IQR], 23.0 [16.0 to 29.5] days; hybrid model probability [IQR], 0.032 [0.018 to 0.060]). As illustrated in Fig. 5, patients with higher hybrid model probabilities had higher cumulative risk of reaching poor outcomes (log-rank test, $P1 < 0.001$; Gehan-Breslow-Wilcoxon test, $P2 < 0.001$).

The hybrid model was also identified as an independent risk factors (HR, 2.049; 95% CI, 1.462–2.871; $P < 0.001$) in the test set, compared with age (HR, 1.020; 95% CI, 0.966–1.077; $P = 0.468$) and sex (HR, 1.667; 95% CI, 0.431–6.452; $P = 0.454$). If the HR excesses 1, we consider the indicator as a significant risk factor.

F. Subgroup Analysis

As shown in Fig. 6, patients with single underlying health condition and multiple underlying health conditions were screened

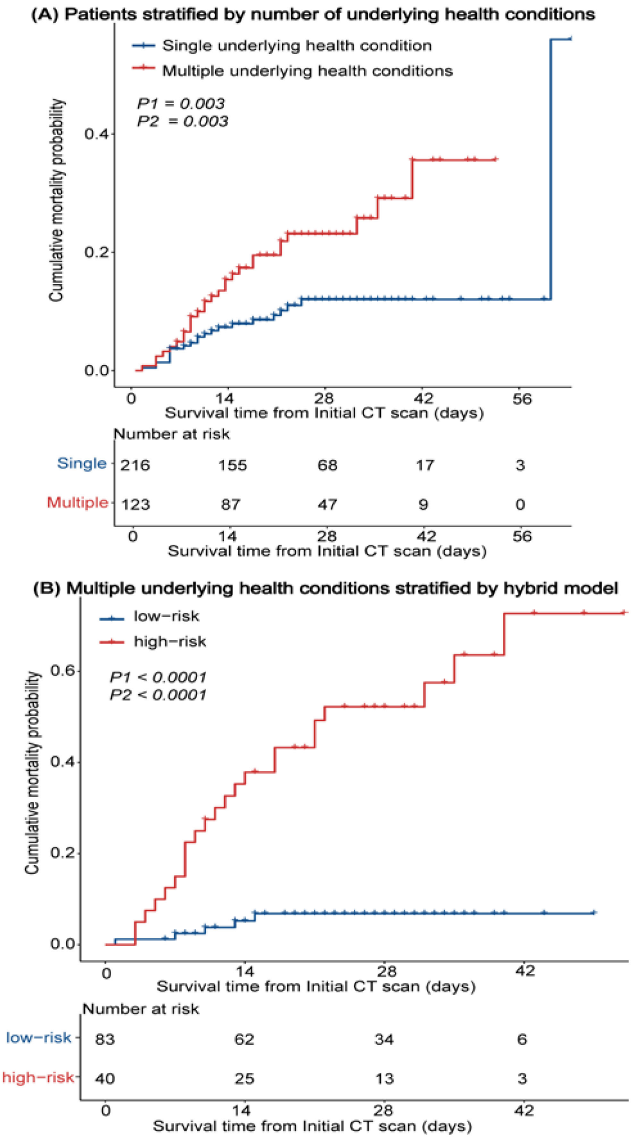

Fig. 6. (A) Patients were stratified by the number of underlying health conditions. The survival curves showed that COVID-19 patients with multiple underlying health conditions had higher risk of death than those with single. (B) Risk stratification ability of the hybrid model verified in multiple underlying health conditions subgroup. *P1* was calculated by log-rank test, and *P2* was calculated by Gehan-Breslow-Wilcoxon test.

as subgroups, respectively. The survival curves showed that COVID-19 patients with multiple underlying health conditions had higher risk of death (log-rank test, $P1 = 0.003$; Gehan-Breslow-Wilcoxon test, $P2 = 0.003$). Further, in the multiple underlying health conditions subgroup, the hybrid model still worked well in risk stratifying COVID-19 patients (log-rank test, $P1 < 0.0001$; Gehan-Breslow-Wilcoxon test, $P2 < 0.0001$).

IV. Discussion

This study enrolled a four-center dataset of 400 COVID-19 patients with underlying health conditions and attempted to develop a prognostic tool based on initial CT scan for identifying poor outcomes among such population. To our knowledge, this is the first investigation that identified a CT-derived prognostic model for COVID-19 patients with underlying health conditions using deep learning and radiomics. Our proposed hybrid model showed competitive performance in precisely stratifying patients at high risk or low risk of death, so that patients could be benefitted with more appropriate personalized protective strategies, including intensive care and timely surveillance.

Evidence from previous researches has revealed the commonness of comorbidities in COVID-19 patients, and clinical outcomes of such patients differed a lot from those without [4]–[6]. Specifically, the associations of hypertension and diabetes with mortality risk were well discussed [7], [34], echoing the common sense and clinical hypothesis that patients with underlying health conditions need more supervision due to probable poor prognosis. On the other hand, early discussions suggested that chest CT interpretation could be a preferred supplementary diagnostic and prognostic criterion for COVID-19 patients, in that radiological pneumonia lesion changes could be captured. The two main reasons provided motivation and theoretical support for this study that CT-derived models had compatible value in prognosis analysis for COVID-19 patients with underlying health conditions.

As shown in recent studies [4], [21], the primary outcome including admission to intensive care unit, invasive ventilation, or death, was met more often. With a relatively long follow-up duration, however, most patients in our study had a definite outcome. Thus, the primary outcome (death, discharge or hospitalization) and the secondary outcome (survival time) were analyzed. Automatic lung volume segmentation was done within tense clinical situations to avoid time-consuming and labor-extensive delineations. The threshold setting and flood fill method enabled the focus on lung areas. Based on segmented lung volumes, the deep learning model explored high-order image features by utilizing convolution kernels, and the radiomics model extracted quantitative low-order features by calculating statistical metrics. Our hybrid model combining the deep learning model and radiomics model predictions was capable of boosting the predictive performance. A probable reason is that deep learning and radiomics methods focus on different lung image scales (local and global), and the combination of different types of features may help analyze the image more comprehensively. Attempts were also made to compare between the feature fusion strategy and the prediction fusion strategy used in this study. The feature fusion strategy failed to outperform the prediction fusion strategy, due to that there were overmuch deep learning features, making the fully connected layers hard to learn the radiomics features effectively. Thus, the promotion the radiomics features brought to the deep learning features was limited.

We have already made the models available to improve the clinical applicability of this study. To help interpret the models, Grad-CAM was depicted to highlight the pneumonia responses to deep learning model, indicating high-risk lesions that may need more attention. The radiomics features used in the radiomics model mainly included three types, i.e., Interquartile Range, Robust Mean Absolute Deviation, and 10 Percentile. They are all first-order statistical features that quantify the gray-level intensity distribution within the lung volumes. For example, the higher the 10 Percentile value, the more the high gray-level pixels within the lung volumes, which may be correlated with the pneumonia lesions.

Given the relatively small dataset size, cross-validation was conducted to make sure the robustness of the models. The results showed that imaging-based models (deep learning model, radiomics model, and hybrid model) performed better than the clinical non-imaging model, which proved the value of images in COVID-19 analysis. Meanwhile, our proposed hybrid model showed best classification ability and was even robust. The development of initial CT-based hybrid model could act as an indispensable tool in identifying poor outcomes in COVID-19 patients with underlying health conditions.

Taking a step forward, survival analysis verified and highlighted the risk stratification ability of the hybrid model. Patients may be categorized as having escalating risk of death at a significant level by the classification threshold of hybrid model. Thus, patients at admission manifesting high risk could receive timely treatment. Moreover, despite risk factors presented by demographics, clinical characteristics, and laboratory findings [35], [36], image-derived risk factors indeed had potential prognostic value. In this study, the hybrid model even had a higher HR than age. Also, as a wide range of underlying health conditions were recorded, prognostic value of hybrid model was also evaluated in subgroups. Despite variations in the proportion of underlying health conditions in individual studies due to different sample sizes and regions, hypertension, diabetes, coronary heart diseases, and carcinoma remained the most common [2]–[5], similar to our statistics. As the number of underlying health conditions may also affect the prognosis of COVID-19 patients, the multiple underlying health conditions subgroup was also evaluated. Our findings implied the effectiveness and stability of the risk stratification ability of the hybrid model.

Regarding the COVID-19 issues, there have been several related studies that achieved good performance [20]–[22]. Compared with them, a major disadvantage of our study is the lack of clinical data, by incorporating which our hybrid model performance may be further improved. But our study also shows advantages in two aspects: 1) Previous studies more focused on all the COVID-19 population, either predicting hospital stay or estimating the disease worsening. Our study, however, could provide a precise tool to plan for care and surveillance in advance only for patients with underlying health conditions; 2) Their models failed to take full advantages of deep learning and radiomics methods, but our hybrid model proved that combining deep learning and radiomics was capable of boosting the predictive performance (Appendix Table S2).

This study still has some limitations. First, a larger dataset is desired to generalize the model performance in the future. Second, the severity of underlying diseases is an important factor in predicting the outcome of COVID-19 patients with underlying health conditions. Hard to collect and quantify, however, the severity information was not incorporated in our models. Future studies may take it into account. Thirdly, due to the lack of clinical data (serial CT scans, changes in symptoms, multiple laboratory tests, etc.), our models may fail in capturing the changes in disease progression during hospitalization. To realize the real-time monitoring and prognosis prediction, serial CT scans can be used for time-related analysis. Fourthly, research has revealed that quantifying GGO caused by COVID-19 was powerful in estimating patients' survival outcomes [37]. Whether our proposed methods could outperform other GGO classifiers was also worth studying.

V. Conclusion

This multi-center study proposed a deep learning and radiomics based hybrid model for accurately identifying poor outcomes in COVID-19 patients with underlying health conditions from initial CT images at admission. The hybrid model outperformed other models and showed great risk stratification ability, which could be a powerful tool for alerting risk of death and arranging individualized surveillance plans.

References

[1] World Health Organization, "Coronavirus disease (COVID-19) outbreak," 2020. [Online]. Available: https://www.who.int

[2] 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.

[3] D. Wang *et al.*, "Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China," *JAMA*, vol. 323, no. 11, pp. 1061–1069, 2020.

[4] W.-J. Guan *et al.*, "Comorbidity and its impact on 1590 patients with COVID-19 in China: A nationwide analysis," *Eur. Respir. J.*, vol. 55, 2020, Art. no. 2000547.

[5] J. Yang *et al.*, "Prevalence of comorbidities in the novel Wuhan coronavirus (COVID-19) infection: A systematic review and meta-analysis," *Int. J. Infect. Dis.*, vol. 94, no. C, pp. 91–95, 2020.

[6] A. Clark *et al.*, "Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: A modelling study," *Lancet Glob. Health*, vol. 8, no. 8, pp. e1003–e1017, 2020.

[7] W. Guo *et al.*, "Diabetes is a risk factor for the progression and prognosis of COVID-19," *Diabetes/Metab. Res. Rev.*, vol. 36, no. 7, p. e3319, 2020.

[8] P. Zhang *et al.*, "Association of inpatient use of angiotensin converting enzyme inhibitors and angiotensin II receptor blockers with mortality among patients with hypertension hospitalized with COVID-19," *Circ. Res.*, vol. 126, no. 12, pp. 1671–1681, 2020.

[9] S. Shi *et al.*, "Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China," *JAMA Cardiol.*, vol. 5, no. 7, pp. 802–810, 2020.

[10] Y. Li and L. Xia, "Coronavirus disease 2019 (COVID-19): Role of chest CT in diagnosis and management," *Amer. J. Roentgenol.*, vol. 214, no. 6, pp. 1280–1286, 2020.

[11] D. Dong *et al.*, "The role of imaging in the detection and management of COVID-19: A review," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 16–29, 2020.

[12] M. Chung *et al.*, "CT imaging features of 2019 novel coronavirus (2019-nCoV)," *Radiology*, vol. 295, no. 1, pp. 202–207, 2020.

[13] F. Pan *et al.*, "Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia," *Radiology*, vol. 295, 2020, Art. no. 200370.

[14] Y. Wang *et al.*, "Temporal changes of CT findings in 90 patients with COVID-19 pneumonia: A longitudinal study," *Radiology*, vol. 296, 2020, Art. no. 200843.

[15] L. Huang *et al.*, "Serial quantitative chest CT assessment of COVID-19: Deep-learning approach," *Radiol.: Cardiothoracic Imag.*, vol. 2, no. 2, 2020, Art. no. e200075.

[16] W. L. Bi *et al.*, "Artificial intelligence in cancer imaging: Clinical challenges and applications," *CA: Cancer J. Clin.*, vol. 69, no. 2, pp. 127–157, 2019.

[17] D. Dong *et al.*, "Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: An international multi-center study," *Ann. Oncol.*, vol. 31, no. 7, pp. 912–920, 2020.

[18] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, “Artificial intelligence and machine learning to fight COVID-19,” *Physiol. Genomics*, vol. 52, no. 4, pp. 200–202, 2020.

[19] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, “Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: A review,” *Chaos, Solitons Fractals*, vol. 139, 2020, Art. no. 110059.

[20] H. Yue *et al.*, “Machine learning-based CT radiomics method for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study,” *Ann. Transl. Med.*, vol. 8, no. 14, p. 859, 2020.

[21] Q. Wu *et al.*, “Radiomics analysis of computed tomography helps predict poor prognostic outcome in COVID-19,” *Theranostics*, vol. 10, no. 16, 2020, Art. no. 7231.

[22] W. Liang *et al.*, “Early triage of critically ill COVID-19 patients using deep learning,” *Nature Commun.*, vol. 11, no. 1, pp. 1–7, 2020.

[23] M. Januszewski *et al.*, “High-precision automated reconstruction of neurons with flood-filling networks,” *Nature Methods*, vol. 15, no. 8, pp. 605–610, 2018.

[24] S. Wang *et al.*, “A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis,” *Eur. Respir. J.*, vol. 56, 2020, Art. no. 2000775.

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[26] J. J. Van Griethuysen *et al.*, “Computational radiomics system to decode the radiographic phenotype,” *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, 2017.

[27] N. Louw and S. Steel, “Variable selection in kernel fisher discriminant analysis by means of recursive feature elimination,” *Comput. Statist. Data Anal.*, vol. 51, no. 3, pp. 2043–2055, 2006.

[28] 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, 2002.

[29] V. Svetnik, A. Liaw, C. Tong, and T. Wang, “Application of Breiman’s random forest to modeling structure-activity relationships of pharmaceutical molecules,” in *Proc. Int. Workshop Mult. Classifier Syst.*, Springer, 2004, pp. 334–343.

[30] M. L. McHugh, “Interrater reliability: The kappa statistic,” *Biochemia medica: Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[31] N. J. Perkins and E. F. Schisterman, “The inconsistency of ‘optimal’ cutpoints obtained using two criteria based on the receiver operating characteristic curve,” *Amer. J. Epidemiol.*, vol. 163, no. 7, pp. 670–675, 2006.

[32] 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.*, 2017, pp. 618–626.

[33] H. T. Kim, “Cumulative incidence in competing risks data and competing risks regression analysis,” *Clin. Cancer Res.*, vol. 13, no. 2, pp. 559–565, 2007.

[34] L. Fang, G. Karakiulakis, and M. Roth, “Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection?,” *Lancet. Respir. Med.*, vol. 8, no. 4, p. e21, 2020.

[35] X. Li *et al.*, “Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan,” *J. Allergy Clin. Immunol.*, vol. 146, no. 1, pp. 110–118, 2020.

[36] J. Zhang *et al.*, “Risk factors for disease severity, unimprovement, and mortality of COVID-19 patients in Wuhan, China,” *Clin. Microbiol. Infect.*, vol. 26, no. 6, pp. 767–772, 2020.

[37] W. Ye *et al.*, “Detection of pulmonary ground-glass opacity based on deep learning computer artificial intelligence,” *Biomed. Eng. Online*, vol. 18, no. 1, pp. 1–12, 2019.