# Integrated Clinical and CT Based Artificial Intelligence Nomogram for Predicting Severity and Need for Ventilator Support in COVID-19 Patients: A Multi-Site Study

Amogh Hiremath, Kaustav Bera, Lei Yuan, Pranjal Vaidya, Mehdi Alilou, Jennifer Furin, Keith Armitage, Robert Gilkeson, Mengyao Ji, Pingfu Fu, Amit Gupta, Cheng Lu, and Anant Madabhushi

***Abstract*—Almost 25% of COVID-19 patients end up in ICU needing critical mechanical ventilation support. There is currently no validated objective way to predict which patients will end up needing ventilator support, when the disease is mild and not progressed. N = 869 patients from two sites ($D_1$: N = 822, $D_2$: N = 47) with baseline clinical characteristics and chest CT scans were considered for this study. The entire dataset was randomly divided into 70% training, $D_1{}^{train}$ (N = 606) and 30% test-set ($D^{test}$: $D_1{}^{test}$ (N = 216) + $D_2$ (N = 47)). An expert radiologist delineated ground-glass-opacities (GGOs) and consolidation regions on a subset of $D_1{}^{train}$, ($D_1{}^{train\_sub}$, N = 88). These regions were automatically segmented and used along with their corresponding CT volumes to train an imaging AI predictor (AIP) on $D_1{}^{train}$ to predict the need of mechanical ventilators for COVID-19 patients. Finally, top five prognostic clinical factors selected using univariate analysis were integrated with AIP to construct an integrated clinical and AI imaging nomogram (CIAIN). Univariate analysis identified lactate dehydrogenase, prothrombin time, aspartate aminotransferase, %lymphocytes, albumin as top five prognostic clinical features. AIP yielded an AUC of 0.81 on $D^{test}$ and was independently prognostic irrespective of other clinical parameters on multivariable analysis ($p<0.001$). CIAIN improved the performance over AIP yielding an AUC of 0.84 ($p = 0.04$) on $D^{test}$. CIAIN outperformed AIP in predicting which COVID-19 patients ended up needing a ventilator. Our results across multiple sites suggest that CIAIN could help identify COVID-19 with severe disease more precisely and likely to end up on a life-saving mechanical ventilation.**

***Index Terms*—Convolutional neural networks, COVID-19, deep learning, nomograms, prognosis, ventilator.**

Manuscript received November 17, 2020; revised January 15, 2021, March 10, 2021, and June 1, 2021; accepted July 29, 2021. Date of publication August 13, 2021; date of current version November 5, 2021. This work was supported in part by the National Cancer Institute under Awards nos. 1U24CA199374-01, R01CA249992-01A1, R01CA202752-01A1, R01CA208236-01A1, R01CA216579-01A1, R01CA220581-01A1, R01CA257612-01A1, 1U01CA239055-01, 1U01CA248226-01, and 1U54CA254566-01, in part by National Heart, Lung, and Blood Institute under Grants 1R01HL15127701A1 and R01HL15807101A1, in part by the National Institute of Biomedical Imaging and Bioengineering under Grant 1R43EB028736-01, in part by the National Center for Research Resources under Award no. 1 C06 RR12463-01, in part by the U.S. Department of Veterans Affairs Biomedical Laboratory Research and Development Service under VA Merit Review Award no. IBX004121A, in part by the Office of the Assistant Secretary of Defense for Health Affairs through the Breast Cancer Research Program under Grant W81XWH-19-1-0668, in part by the Prostate Cancer Research Program under Grants W81XWH-15-1-0558 and W81XWH-20-1-0851, in part by the Lung Cancer Research Program under Grants W81XWH-18-1-0440 and W81XWH-20-1-0595, in part by the Peer Reviewed Cancer Research Program under Grant W81XWH-18-1-0404, in part by the Kidney Precision Medicine Project (KPMP) Glue Grant, in part by the Ohio Third Frontier Technology Validation Fund, in part by the Clinical and Translational Science Collaborative of Cleveland under Grant UL1TR0002548 from the National Center for Advancing Translational Sciences (NCATS) component of the National Institutes of Health and NIH Roadmap for Medical Research, in part by the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering at Case Western Reserve University, in part by Bristol Myers-Squibb, Boehringer-Ingelheim, and Astrazeneca, and in part by the National Natural Science Foundation of China under Grant 81901817. *(Corresponding author: Mengyao Ji.)*

Please see the Acknowledgment section of this article for the author affiliations.

Digital Object Identifier 10.1109/JBHI.2021.3103389

## I. Introduction

THE CORONAVIRUS Disease of 2019 (COVID-19) caused by novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-Cov-2) is an ongoing pandemic which has led to widespread deaths in patients who are older (usually over 65 years [1]) with significant comorbidities especially susceptible [2], [3].

Multiple studies have shown the presence of lung changes in CT scans in some COVID-19+ patients who initially tested negative for the virus on Reverse transcription polymerase chain reaction (RT-PCR) [4]. CT is currently recommended in patients with COVID-19 exhibiting moderate to severe clinical features or worsening respiratory status features as well as for treatment monitoring [5].

Artificial intelligence (AI) approaches on CT scans, including convolutional neural network (CNN) based deep-learning (DL) approaches have recently been proposed for detecting COVID-19 even in the asymptomatic stage [6], [7]. The majority of these studies were limited to being single-site, typically developed on imaging alone, and focused on automated diagnosis or differentiation as opposed to severity assessment or prognosis

© IEEE 2021. This article is free to access and download, along with rights for full text and data mining, re-use and analysis.

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2024 at 03:38:49 UTC from IEEE Xplore. Restrictions apply.

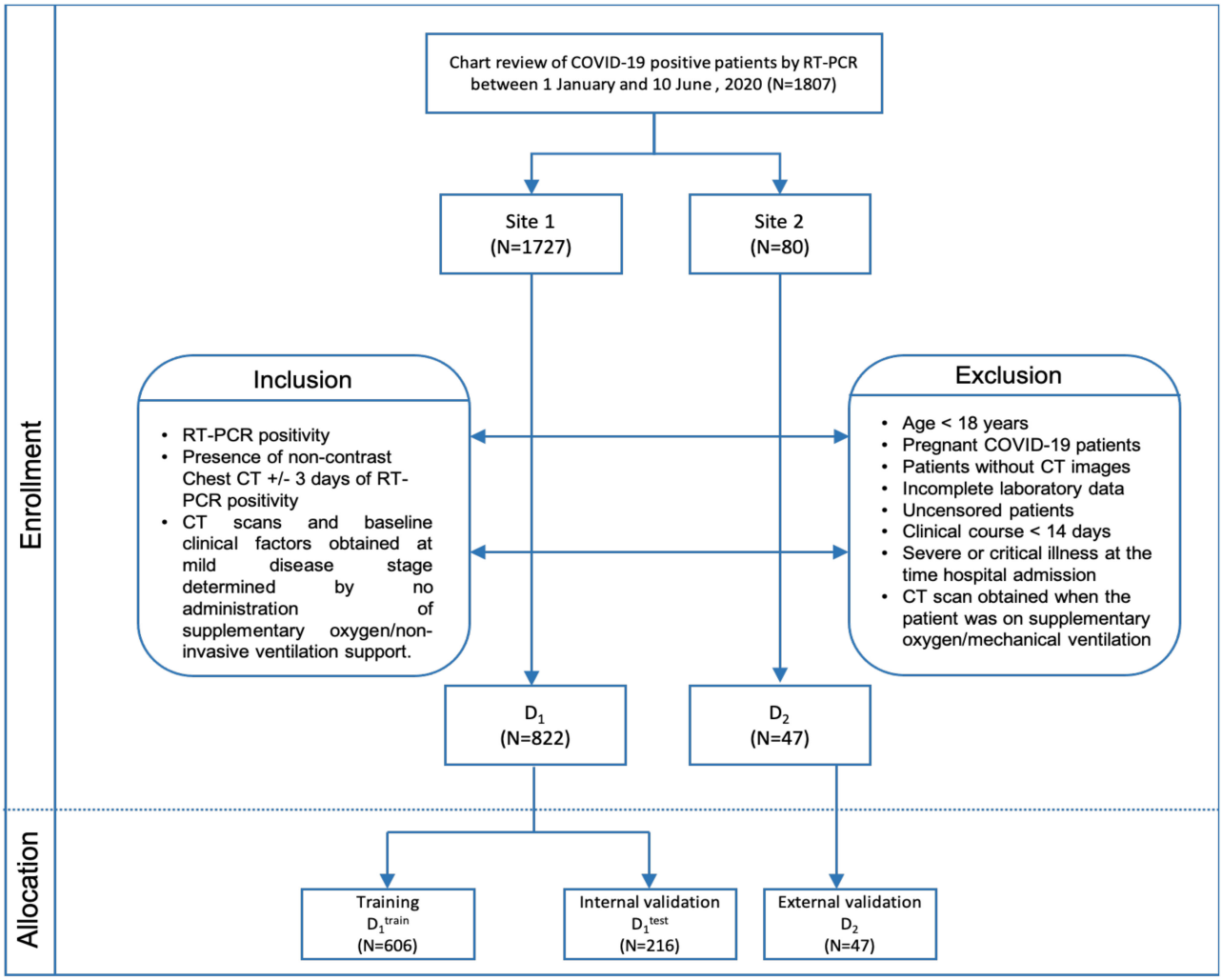


Fig. 1. Patient flow chart including patient enrollment, eligibility and exclusion criteria and allocation of the data into different splits (training, internal validation and external validation).

of COVID-19. Identifying those patients at an earlier disease presentation who would progress on to more aggressive disease and hence would need mechanical ventilation would decrease disease progression and mortality.

In this work, we present a novel first of its kind hybrid AI approach combining DL on CT scans at diagnosis (RT-PCR positivity) as well as clinical factors to construct an integrated imaging and clinical nomogram (ClAIN) to predict which COVID-19 patients would have a severe disease phenotype and end up needing invasive mechanical ventilation. The presented approach is fully automated, where a U-Net based CNN is first used to segment ground glass opacities (GGOs) and consolidation regions. Subsequently a second network, a 3D CNN, is used to capture image representations from the areas identified by the first network, followed by the integration of the imaging features with the clinical parameters. The models were trained on $D_1{}^{train}$ ($N = 606$ patients) and evaluated on data from two different sites ($D^{test}$: $D_1{}^{test}$ ($N = 216$) + $D_2$ ($N = 47$)). The integrated nomogram ClAIN was compared with the imaging AI predictor (AIP) and clinical AI predictor (ACP) models built using only CT imaging and clinical features respectively.

## II. Materials and Methods

### A. Patients

The retrospective chart review study was approved by the Institutional Review Board committee of record at University Hospitals, Cleveland and Ethics committee of the Renmin Hospital of Wuhan University (ethics number: V1.0), and the need for the written consent was waived. Following the inclusion and exclusion criteria (Fig. 1), the study included $D_1$ ($N = 822$) patients, from Renmin hospital of Wuhan University, Hubei General Hospital and $D_2$ ($N = 47$) patients from University Hospitals, Cleveland. A stratified random sampling was performed to split the data into 70% training, $D_1{}^{train}$ ($N = 606$) and 30% test set $D^{test}$ ($D_1{}^{test}$ ($N = 216$) + $D_2$ ($N = 47$)). The patients were acquired by consecutive chart review for patients who were seen between 1 January and 10 June 2020. The baseline clinical characteristics and chest CT scans was acquired for all the patients. The CT examinations included GE medical systems, United Imaging Healthcare, Philips and SIEMENS manufacturers with standard chest imaging protocol (the patient placed in supine position, and helical scanning performed during

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2024 at 03:38:49 UTC from IEEE Xplore. Restrictions apply.

breath hold at end inspiration). The scan parameters include 120 kVp, automatic tube current modulation and slice thicknesses ranging from 2 mm to 5 mm. Additional details of the chest CT acquisition parameters are provided in SUPPLEMENTARY TABLE I.

### B. Detection and Segmentation of Lung Lesions on the Baseline Chest CT Scans of COVID-19 Patients

An expert radiologist with 14 years of experience delineated GGOs and consolidation regions on a subset of $D_1^{train}$ ($D_1^{train\_sub}$, N = 88) and a subset of $D_1^{test}$ ($D_1^{test\_sub}$, N = 96 patients). A CNN with U-Net architecture [8] (Supplementary Fig. 1) was employed to segment out these regions in the lung on the baseline chest CT scans.

A previously used automatic lung segmentation method utilizing watershed transform [9] was used to segment out and crop the CT volume around the regions of the lung. Each 2D slice of the cropped volume was resized to a shape of 256x320 and parts of the lung region (right, left) were given as separate inputs (input size: 256x160) to the network to segment GGOs and consolidation regions (see supplementary section, Appendix ***A***, for further data pre-processing and data augmentation details).

### C. A Deep Learning Network on Chest CT Scans for Predicting the Need for Mechanical Ventilation in COVID-19 Patients

A 3D CNN [10] with 6 convolutional layers and 3 dense layers (Supplementary Fig. 2) was used to construct an imaging AI predictor (AIP). The network consisted of two distinct input channels with CT volume cropped around the lung region being the first input and the corresponding volume of automatically segmented binary segmentations of GGOs and consolidations being the second. The details of the network initialization parameters, hyper-parameters and the programming software / platform used to build the DL framework is provided in the supplementary section (Appendix ***B, C***).

### D. Evaluation Metrics and Statistical Analysis

*Performance of detection and segmentation of GGOs and consolidation regions*: Dice Similarity Coefficient (DSC) was used to evaluate the voxel wise segmentation performance as compared to an expert radiologist reader.

*Performance of the classifiers in predicting which COVID-19 patients would end up needing invasive mechanical ventilation*: The outcome of interest was disease severity as invasive mechanical ventilation/ECMO/death vs. no invasive ventilator support (no respiratory distress, oxygen supplementation, non-invasive ventilation).

The receiver operating characteristic (ROC) analysis with sensitivity, specificity, Area under ROC curve (AUC) and positive predictive value (PPV) as performance metrics was used to evaluate the accuracy of detection of COVID-19 regions on CT and the performance of the models that predict the need of invasive mechanical ventilation for COVID-19 patients. DeLong test [11] was used to compare the statistical significance of difference in AUCs between two models.

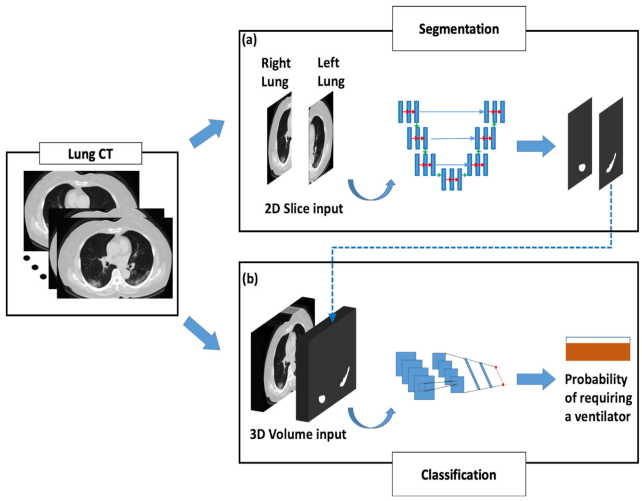


Fig. 2. An end-end deep learning framework consisting of (a) 2D U-Net based segmentation network for segmenting the GGOs and consolidation regions in the lung and (b) an imaging AI predictor (AIP) constructed using 3D CNN to predict which COVID-19 patients would end up getting a ventilator. The segmentations of ground glass opacities and consolidation regions in the lung are performed using a 2D U-Net based network. At first, the lung region is segmented and part of the lung slice (right, left) is given as input to the network. A 3D CNN is used to construct AIP. The CT volume cropped around the lung and its corresponding segmented regions of ground glass opacities and consolidations are given as input to the AIP.

95% confidence intervals (CI) were calculated to determine statistical significance, and cross validation results on $D_1^{train}$ were reported as mean ± standard deviation.

## III. Experimental Design

### A. Experiment 1: Deep Learning Classifier Using Baseline Chest CT for Predicting the Need for Mechanical Ventilation in COVID-19 Patients

An end-end deep learning framework consisted of a) 2D U-Net based segmentation network for segmenting the GGOs and consolidation regions in the lung and b) AIP constructed using 3D CNN to predict the need of a ventilator in COVID-19 patients (Fig. 2).

The U-Net network was trained on the training set, $D_1^{train\_sub}$, with a 3-fold cross-validation setting and further evaluated on $D_1^{test\_sub}$ (N = 96). We used the segmentation maps outputted by the networks to evaluate detection performance of the network in detecting the GGOs and consolidations. We defined the region as being detected if $\geq$ 0.2 DSC overlap existed between the network segmentation map and the ground-truth delineation of that corresponding region. We report the segmentation accuracy of the detected regions in terms of DSC.

Subsequently, the AIP was trained on $D_1^{train}$ with a 3-fold cross validation setting for predicting which COVID-19 patients would end up getting on a ventilator. Further, the ensemble of the AIP's predictions trained on 3-fold cross validation (average predictions of the predictors) was used to evaluate the performance on $D^{test}$. Additionally, the performance of the presented

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2024 at 03:38:49 UTC from IEEE Xplore. Restrictions apply.

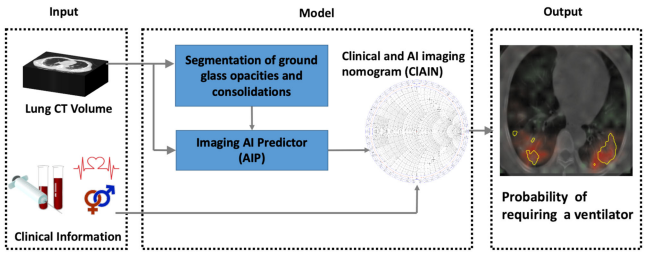


Fig. 3. Overall modelling framework to create an integrated clinical and imaging-based nomogram (ClAIN) to predict which COVID-19 patients would end up getting a ventilator. The decision obtained from an imaging AI predictor (AIP) and the most discriminating clinical features are integrated in a multivariate logistic regression model to construct ClAIN.

architecture of AIP was compared with other widely used architectures such as 3D ResNet [12], [13] and DenseNet [14].

### *B. Experiment 2: An Integrated Clinical and Imaging Nomogram to Predict the Need of Mechanical Ventilation in COVID-19 Patients*

Univariate analysis was performed to identify the top 5 most prognostic clinical (age, gender) and laboratory parameters (CBC, Inflammation markers, coagulation markers, general chemistry). We excluded those clinical factors which were found in less than 50% of the cases in the training set (such as IL6 which is not routinely captured). For a full list of clinical and laboratory parameters used see SUPPLEMENTARY TABLE II. The results of univariate analysis performed of each variable is shown in the SUPPLEMENTARY TABLE III. A median value based missing value imputation was used to impute the missing values in the dataset. Subsequently, min-max normalization was applied to each of the clinical features, normalizing the values between 0 and 1. A clinical AI predictor (ACP) was built by training a logistic regression classifier on the top 5 selected features. Finally, the top 5 features were integrated with the probability scores obtained by AIP and a multivariable logistic regression was trained to generate an integrated clinical and AI imaging nomogram (ClAIN). The overall workflow diagram to construct ClAIN is depicted in Fig. 3. Both the models, ACP and ClAIN were validated on $D^{test}$. DeLong [11] test was used to compare the performance between the models.

## IV. Results

### *A. Study Population Characteristics*

TABLE I lists the study population characteristics for the two sites $D_1$ and $D_2$. The median age of the patients was 59 in $D_1$ and 60 in $D_2$, and, 46.8%, 53.2% and 51%, 49% were male, female in $D_1$ and $D_2$ respectively. (For detailed patient characteristics, refer to TABLE I). 41.9%, 55.3% had a mild disease which resolved while 58.1%, 44.7% had severe ARDS needing invasive mechanical ventilation in $D_1$ and $D_2$ respectively. The clinical characteristics were not significantly different between the training and validation datasets with age ($p = 0.2$), sex ($p = 0.09$) being largely similar.

TABLE I
DEMOGRAPHIC INFORMATION AND CLINICAL LAB TEST INFORMATION

| Variables | $D_1$ | $D_2$ |
|---|---|---|
| Age Median (IQR) | 59 (49 - 69) | 60 (51.25 – 68.75) |
| Sex | | |
| Male (%) | 385 (46.8%) | 24 (51%) |
| Female (%) | 437 (53.2%) | 23 (49%) |
| Ventilator | | |
| Yes (%) | 478 (58.1%) | 21 (44.7%) |
| No (%) | 344 (41.9%) | 26 (55.3%) |
| Laboratory findings Median (IQR) | | |
| Lactate Dehydrogenase (units/liter) | 230 (187 - 310.75) | 250 (230 – 368.5) |
| Prothrombin Time (seconds) | 11.6 (11 – 12.2) | 12.7 (11.3 – 14.1) |
| Lymphocytes % | 23.45 (14.6 – 31.8) | 16.6 (12.35 – 23.55) |
| Albumin (grams/liter) | 38.1 (34.6 – 41.37) | 37 (33 - 39) |
| Aspartate Aminotransferase (units/liter) | 25 (19-37) | 27 (22-46) |

IQR: Interquartile range.

TABLE II
U-NET SEGMENTATION RESULTS IN SEGMENTING GROUND GLASS OPACITIES AND CONSOLIDATION REGIONS

| | Detected | False positives | Sensitivity (%) | PPV (%) | DSC |
|---|---|---|---|---|---|
| $D_1^{train}$ | 1017/1260 | 449 | 80.71% | 69.3% | 0.60 ±0.02 |
| $D_1^{sub}$ | 1071/1353 | 470 | 79.15% | 69.5% | 0.59 |

PPV: Positive predictive value.DSC: Dice similarity coefficient.

### *B. Experiment 1: Deep Learning Classifier Using Baseline Chest CT for Predicting the Need for Mechanical Ventilation in COVID-19 Patients*

The performance of U-Net in detection (number of 3D connected regions detected, false positives, sensitivity and positive predictive value (PPV)), and segmentation (DSC) of GGOs and consolidation regions is presented in TABLE II. The U-Net network resulted in detection sensitivity and PPV of 80.71% and 69.3% respectively, and segmentation (Fig. 4) DSC of $0.60 \pm 0.02$ on $D_1^{train\_sub}$. On $D_1^{test\_sub}$ ($N = 96$), the sensitivity and PPV was found to be 79.15% and 69.5% respectively, and the corresponding DSC of the segmentations was 0.59.

The 3-fold cross-validation results and performance on $D^{test}$ of AIP and other classical architectures in predicting which

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2024 at 03:38:49 UTC from IEEE Xplore. Restrictions apply.

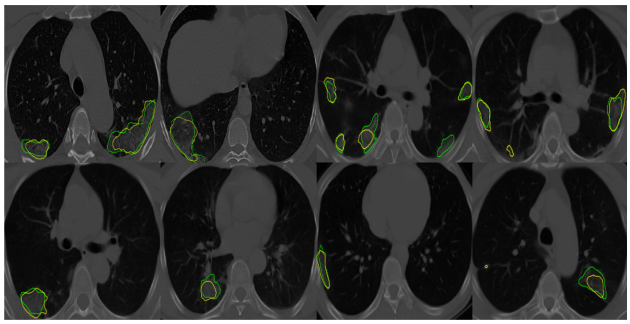

Fig. 4. Segmentations results of a 2D U-Net in segmentation of ground glass opacities and consolidation regions in the lung. Green contours represent ground-truth delineations of the GGOs and consolidations and their corresponding network segmentation contours are represented in yellow. The 2D U-Net network was trained on the training set of N=88 patients and was evaluated on a subset of the testing set $D_1^{sub}$ (N=96 patients).

TABLE III
COMPARISON OF PERFORMANCE (AUC) OF THE PRESENTED ARCHITECTURE AIP WITH OTHER ARCHITECTURES (RESNET AND DENSENET) ON $D_1^{TRAIN}$ AND $D^{TEST}$

| Architecture | AUC (mean ± std.) ($D_1^{train}$) | ACC (%) ($D_1^{train}$) | AUC (95% CI) ($D^{test}$) | ACC (%) ($D^{test}$) |
|---|---|---|---|---|
| ResNet10 | 0.72±0.05 | 71.6% | 0.777 (0.72-0.83) | 70.0% |
| ResNet50 | 0.75±0.04 | 72.9% | 0.785 (0.73-0.83) | 71.0% |
| ResNet101 | 0.78±0.06 | 70.0% | 0.802 (0.75-0.85) | 70.5% |
| DenseNet121 | 0.75±0.03 | 76.3% | 0.806 (0.75-0.86) | 71.9% |
| DenseNet169 | 0.79±0.04 | 76.5% | 0.820 (0.77-0.87) | 76.0% |
| DenseNet201 | 0.81±0.02 | 75.6% | 0.807 (0.76-0.85) | 74.8% |
| AIP | 0.86±0.01 | 79.5% | 0.809 (0.75-0.87) | 75.3% |

AUC: area under the receiver operating characteristic curve. ACC: Accuracy. std: standard deviation.

COVID-19 patients needed mechanical ventilation support are presented in TABLE III. AIP yielded a 3-fold cross-validation AUC of 0.86±0.01 on $D_1^{train}$, outperforming other architectures such as ResNet [12], [13] and DenseNet [14] whose performance was found to be in the range of 0.72-0.81. The ensemble of the AIP's predictions (average predictions of AIP trained on three different folds of cross-validation) resulted in an AUC of 0.809; 95% CI [0.75-0.87], on $D^{test}$. Although, DenseNet169 yielded an AUC = 0.820; 95% CI [0.77-0.87], we chose AIP as the base architecture since the difference in the performance was not found to be statistically significant ($p>0.05$), and AIP (1323768 parameters) has 18x fewer parameters compared to DenseNet169 (18568002 parameters). The unsupervised clustered heat map of deep features extracted from the pre-final layer of AIP shows separate clusters for patients who ended up getting on a ventilator and for those who did not (Supplementary Fig. 3).

TABLE IV
COMPARISON OF PERFORMANCE (AUC) WITH AND WITHOUT THE USE OF BINARY SEGMENTATION MAPS AS SECONDARY CHANNEL TO AIP ON $D_1^{TRAIN}$ AND $D^{TEST}$

| Dataset | ($D_1^{train}$) | | ($D^{test}$) | |
|---|---|---|---|---|
| Model | AUC (mean ± std) | ACC (%) | AUC (95% CI) | ACC |
| Without segmentations | 0.76±0.07 | 75.1% | 0.753 (0.69-0.81) | 71.6% |
| With Segmentations | 0.86±0.01 | 79.5% | 0.809 (0.75-0.87) | 75.3% |

AUC: area under the receiver operating characteristic curve.ACC: Accuracy. std: standard deviation.

Gradient-weighted Class Activation Maps (Grad-CAM) [15] were used to showcase some examples of interpretation of AIP. Grad-CAM helps in highlighting the regions in an image that are associated with a particular class. From Fig. 5, we observe that the network primarily focuses on the segmented GGOs and consolidation regions. Additionally, when the network was trained without the use of binary segmentation masks as an auxiliary input channel to the network, a performance drop of 10% AUC was observed on $D_1^{train}$ (TABLE IV). The corresponding AUCs on $D^{test}$ was found to be 0.753; 95% CI [0.69-0.81] leading to drop in AUC of 5.6% on $D^{test}$. Therefore, the use of binary segmentations as the second channel input to the network aids the network in setting an attention region [10] helping the network to focus on these regions, while at the same time, providing the context of the whole lung region.

### C. Experiment 2: An Integrated Clinical and Imaging Nomogram to Predict Need for Mechanical Ventilation in COVID-19 Patients

The most prognostic clinical factors (see SUPPLEMENTARY TABLE II for the list of all the clinical factors) identified were prothrombin time, albumin, lactate dehydrogenase, aspartate aminotransferase, and % lymphocyte. The ACP model trained using the most discriminating clinical factors yielded an AUC of 0.74; 95% CI [0.67-0.80] on $D^{test}$.

The integrated model, ClAIN outperformed AIP resulting in an AUC of 0.84; 95% CI [0.79-0.89] ($p = 0.04$) on $D^{test}$. TABLE V shows other performance metrics such as sensitivity and specificity for all the models AIP, ACP and ClAIN on $D^{test}$ at two optimal cut-off points on the ROC (maximizing sensitivity on $D_1^{train}$, and maximizing F1- score on $D_1^{train}$). At an optimal operating point on the ROC determined by maximizing F1-score [16] on the $D_1^{train}$, ClAIN resulted in the accuracy, sensitivity and specificity of 77.9%, 97.3% and 52.6% on $D^{test}$ respectively.

The multivariate logistic regression analysis of the ClAIN model (TABLE VI) revealed that AIP, Lactate Dehydrogenase (LDH) and Prothrombin Time (PT), added independent prognostic value irrespective of other clinical parameters on multivariable analysis ($p < 0.001$). Fig. 6 depicts the integrated clinical nomogram, ClAIN. We can observe that LDH, PT, and

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2024 at 03:38:49 UTC from IEEE Xplore. Restrictions apply.

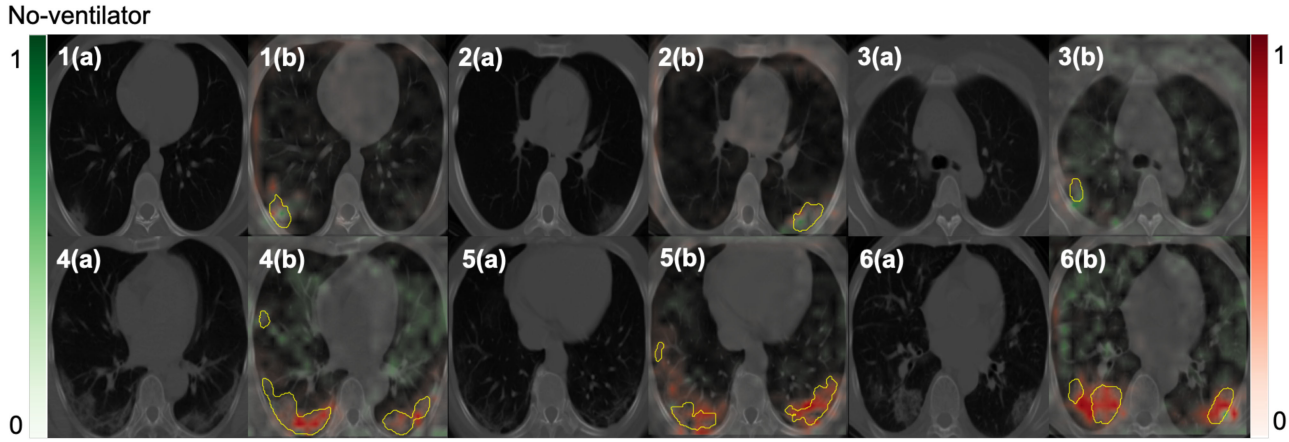


Fig. 5. Model interpretability results with guided gradient weighted class activation maps (Grad-CAM) based on Chest CT scan to differentiate COVID-19 patients not requiring a ventilator (top row patients 1, 2 and 3) from the ones needing a ventilator (bottom row patients 4, 5 and 6). The color "red" indicates the pixel contributing towards the need of a ventilator while the color "green" corresponds to the pixel contributing towards not requiring a ventilator. The color bar gradient corresponds to the strength of the contribution. We can observe that the use of binary segmentation maps as an input channel aids the network to focus on areas of ground glass opacities and consolidations. We can also observe that for patients who need a ventilator, the majority of the GGOs and consolidations are illustrated in red.

TABLE V
PERFORMANCE METRICS (AUC, SENSITIVITY, SPECIFICITY) OF THREE MODELS; AIP, ACP, CIAIN ON $D^{TEST}$ (N=263)

| $D^{test}$ | AUC (95% CI) | Accuracy (%) | Sensitivity (%) | Specificity (%) | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|---|---|---|
| | | Cutoff: Maximizing accuracy on $D_1^{train}$ (N=606) | | | Cutoff: Maximizing F1- score on $D_1^{train}$ (N=606) | | |
| AIP | 0.81 (0.75-0.87) (p= 0.04) | 77.2% | 93.3% | 56.1% | 75.3% | 96.0% | 48.2% |
| ACP | 0.74 (0.67-0.80) (p<0.001) | 75.3% | 89.3% | 57.0% | 73.0% | 95.3% | 43.9% |
| ClAIN | 0.84* (0.79-0.89) | 77.9% | 90.6% | 61.4% | 77.9% | 97.3% | 52.6% |

(AIP: A 3D convolutional neural network (CNN) trained on baseline non-contrast chest CT images), clinical AI predictor (ACP**:** A logistic regression-based model trained on the most discriminable clinical factors) and ClAIN (a logistic regression based integrated clinical and AI imaging nomogram).

TABLE VI
MULTIVARIABLE LOGISTIC REGRESSION ANALYSIS OF CIAIN

| Variable | Log (Odds Ratio) | p-value |
|---|---|---|
| AIP (fraction) | 2.6366 | <0.001 |
| ALB (grams/liter) | -0.0182 | 0.487 |
| LDH (units/liter) | 0.0066 | <0.001 |
| PT (seconds) | 0.2497 | <0.001 |
| AST (units/liter) | -0.0009 | 0.563 |
| % LYM | -0.0015 | 0.708 |

ClAIN: integrated clinical and imaging AI nomogram.

risk score from AIP are the major contributing factors in the integrated nomogram, ClAIN. The decision curve analysis on $D^{test}$ indicated an added net-benefit in using ClAIN over AIP and ACP (Fig. 6).

## V. DISCUSSION

In this study we constructed an integrated clinical and AI based nomogram (ClAIN) to predict at baseline which patients have a severe phenotype of COVID-19 and would end up developing severe ARDS needing intubation and mechanical ventilation. In order to decrease bias, we explicitly used those patients along with baseline CT scans and laboratory parameters in the mild stage of disease without any respiratory assistance. ClAIN comprised a state-of-the-art DL model (AIP) based off baseline non contrast CT scans. AIP incorporated a CNN where automatically segmented COVID-19 regions on lung CT was taken as an auxiliary input, beside the whole segmented lung as the region of interest, to predict the need for invasive ventilation

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2024 at 03:38:49 UTC from IEEE Xplore. Restrictions apply.

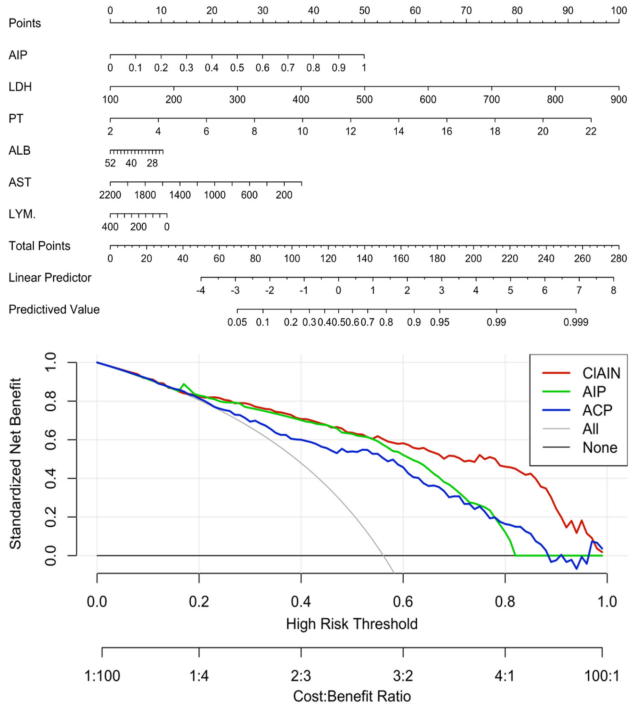


Fig. 6. A ClAIN (Clinical and imaging integrated nomogram) (top) constructed using output probability score of the imaging AI predictor (AIP: a 3D convolutional neural network trained for predicting which COVID-19 patients would end up getting on a ventilator) and five most discriminable clinical features with decision curve analysis (bottom) to evaluate the net benefit of using ClAIN over the models AIP (only imaging) and ACP (only clinical). The models were trained on $D_1{}^{train}$ (N=606 patients) and was evaluated on $D_1{}^{test}$ (N=216 patients) and $D_2$ (N = 47 patients).

as the outcome of interest. Meanwhile, the ACP incorporated prothrombin time, albumin, lactate dehydrogenase, aspartate aminotransferase, % lymphocyte out of routine clinical parameters. Additionally, the three models were trained and evaluated on a large multi-institutional dataset making this the largest study we are aware of involving the use of AI for prognosis of COVID-19 patients.

The integrated ClAIN model outperformed AIP and ACP models in predicting which COVID-19 patients would ultimately need invasive mechanical ventilation on both internal ($D_1{}^{test}$) and external validation sets ($D_2$). ClAIN improved performance by over 10% ($p<0.001$) over ACP and by 3% ($p = 0.04$) over AIP in terms of AUC with the performance increase found to be statistically significant by DeLong's test. AIP model performance was found to be independently predictive irrespective of other clinical parameters on multivariable analysis. The ClAIN model was also used to individualize risk assessments by constructing a nomogram which showed benefit over using only the DL approach or the clinical factors. Nomogram predicted score of 0.28 and greater (optimal cutoff point on the ROC curve) suggested the need for mechanical ventilation while scores less than or equal to 0.28 could be managed conservatively.

Given current expert recommendations for early intubation and initiation of invasive mechanical ventilation to significantly reduce disease progression and decrease COVID-19 related mortality [17], there is an urgent need to build a validated prognostic approach using routine clinical tools to risk assess patients who have tested positive for COVID-19 and are at a relatively early course of disease. This would not only allow for early initiation of medications or supportive interventions to decelerate disease progression for these patients [17], [18], but in the face of worldwide ventilator shortage [19] allow for early identification of ideal ventilator candidates - those at increased risk of developing severe ARDS or death.

At present, increase in laboratory parameters including but not limited to lymphocytopenia, acute phase reactants like procalcitonin, LDH, IL-6, coagulation factors like Prothrombin time, D-dimer, Fibrinogen have been correlated with severe disease or respiratory deterioration, but there is no currently validated way to risk assess these patients rapidly on the bedside immediately when they test positive by RT-PCR for COVID-19.

Ellinghaus et al[20]. meanwhile took a GWAS approach on 1980 patients and found ABO blood group and 3p21.31 gene cluster as a genetic susceptibility locus for disease severity defined by the need for invasive mechanical ventilation. Our study differs from the above in being easy to deploy for routine clinical use unlike expensive genomic assays as it makes use of routinely acquired non-contrast CT scans and routine laboratory parameters. Additionally, to the best of our knowledge, ours is one of the first studies to i) construct and validate a hybrid (clinical and CT-derived DL) approach to predict the need for intubation and mechanical ventilation in RT-PCR determined COVID-19 patients, ii) validate our approach in over 800 patients from two different institutions; iii) completely automated – from disease detection on CT to assessment of severity; iv) predicts disease outcome rather than using AI for diagnostic or differentiation applications.

While our approach makes use of a DL network with its inherent limitations including the relative lack of robust biological underpinning, we rectified that to an extent by passing the binary segmentation of GGOs and consolidation regions as an input to the CNN. The automatically segmented region of COVID-19 lesions as an auxiliary input to the network directs the network to clinical regions of interest for its decision making. In fact, when compared to the same network but without the segmentations as an input to the network, performance decreased by 10% and 5.6% on $D_1{}^{train}$ and $D^{test}$ respectively. This seemed to suggest that it is subtle changes that the network picks up from COVID-19 regions which allows it to predict outcome. Additionally, on post-hoc analysis of network activation maps (Fig. 5) by Grad-CAM, we could observe the regions with GGOs, and consolidations got activated based on the class that they belong to (patients who would end up getting on a ventilator and the ones who did not).

Our study did have its limitations. Our study was retrospective in nature and the two cohorts were not homogeneously defined, and hence to ensure the clinical usefulness of ClAIN, we need to validate the tool in a prospective setting by following up patients till discharge. Additionally, the retrospective nature of the study also precluded us from standardizing the time between RT-PCR positivity and CT scans across the cohort. Furthermore, while a single experienced radiologist delineated the COVID regions

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2024 at 03:38:49 UTC from IEEE Xplore. Restrictions apply.

(GGOs and consolidations), employing multiple readers could have yielded a consensus annotation which might have been less prone to sensitivity of the annotations made by the individual reader. Additionally, since the dataset in the study was retrospectively obtained from multiple institutions, we did not have access to the raw pre reconstructed data. Therefore, standardizing and constructing all the scans with only one reconstruction kernel to perform the analysis remains part of future research directions. Finally, we did not explicitly compare segmentation and prediction performances between the AI model and expert radiologist interpretations, because our final goal was to build a prognostic model, and not focus solely on region detection accuracy.

## VI. Conclusion

In conclusion, we constructed an integrated DL and clinical parameter prognostic model using routinely available blood parameters and standard-of-care CT scans at baseline in SARS-CoV-2 positive patients at the milder stage of disease. We showed in a multi-institutional cohort that our integrated model was able to accurately identify as to which of these patients would decline to severe respiratory distress and would need intubation and mechanical ventilation assistance. Further multi-site prospective validation would allow for clinical deployment of ClAIN specially to triage patients for ventilator usage, in the face of worldwide shortages in availability of mechanical ventilators. The developed tool once prospectively validated could provide an objective way to risk stratify patients immediately following diagnosis with COVID-19.

## Appendix

### A. Data Preprocessing and Augmentation

All CT scans were first pre-processed by converting them from Hounsfield units to image intensities by considering the air in the lungs as having zero intensity value. Augmentations were performed on 3D volume. A random combination with certain probabilities of rotation (3°, 5°, 8°) along the axial plane, and shearing was performed to increase the size of the training dataset. For each of the volumes 5 different combinations of randomly chosen augmentations were performed.

### B. Initialization and Hyper-Parameter Settings of the Deep Learning Networks

U-Net (segmentation of GGOs and consolidation regions) and the imaging AI predictor (AIP) were both implemented in pytorch (0.4.1).

*Initialization*: Both the CNNs were initialized with a manual seed.

*Loss function*: Dice Loss = 1 – dice similarity co-efficient (DSC) function was used train U-Net for segmentation of GGOs and consolidation regions. Binary cross entropy loss function was used for training AIP for predicting which COVID-19 patients would end up getting on a ventilator.

Stopping criteria: An early stopping criterion (patience = 10) was used to stop the network training with respect to the leave one out cross validation loss.

Optimizer: The training of both the networks was performed using an Adam optimizer.

Learning rate and optimizer weight decay: A grid search was performed to choose the learning rate ($10^{-5}$–$10^{-3}$) and weight decay parameter ($10^{-5}$–$10^{-3}$) for the optimizer. A learning-rate of $10^{-4}$ and weight decay of $10^{-5}$ was chosen based on highest cross-validation AUC (SUPPLEMENTARY TABLE IV).

Size of the training set: An ablation study was conducted to choose the size of the training set (10%, 30%, 50%, 70%). Using 70% of the training set, lead to the highest cross-validation AUC = 0.86 (SUPPLEMENTARY TABLE IV). Therefore, 70% training set and 30% test set was chosen to build and evaluate all the models (AIP, ACP and ClAIN).

Batch size: A batch-size of 24 was used to train the networks.

### C. Programming Language

Most of the analysis in this study was done using python as a programming language as well as "R" for some of the statistical analysis.

Specific packages such as matplotlib (2.2.2), numpy (1.17.4), scipy (1.1.0), scikit-learn (0.20.2), pytorch (0.4.1) was used.

Area under the receiver characteristic operating curve (AUC) was calculated using R package 'pROC'. A trapezoidal rule was used along with 95% confidence interval (CI) obtained by performing 2000 stratified bootstrap replicates to calculate AUCs. A DeLong test was used to compare the difference between two AUCs.

## Acknowledgment

The content is solely the responsibility of the authors and does not necessarily represent the official views of the , the U.S. Department of Veterans Affairs, the Department of Defense, or the United States Government.

### Authors Affiliation

Amogh Hiremath, Kaustav Bera, Pranjal Vaidya, Mehdi Alilou, and Cheng Lu are with the Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH 44106 USA (e-mail: axh672@case.edu; kxb413@case.edu; pvv3@case.edu; mxa574@case.edu; cxl884@case.edu).

Lei Yuan is with the Department of Information Center, Renmin Hospital of Wuhan University, Wuhan, Hubei 430060, China (e-mail: yuanlei009@whu.edu.cn).

Jennifer Furin and Keith Armitage are with the Department of Infectious Diseases, University Hospitals Cleveland Medical Center, Cleveland, OH 44106 USA (e-mail: jennifer.furin@uhhospitals.org; keith.armitage@uhhospitals.org).

Robert Gilkeson and Amit Gupta are with the Department of Radiology, University Hospitals Cleveland Medical Center, Cleveland, OH 44106 USA (e-mail: robert.gilkeson@uhhospitals.org; amit.gupta@uhhospitals.org).

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2024 at 03:38:49 UTC from IEEE Xplore. Restrictions apply.

Mengyao Ji is with the Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan, Hubei 430060, China (e-mail: amy_5840@whu.edu.cn).

Pingfu Fu is with the Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106 USA (e-mail: pxf16@case.edu).

Anant Madabhushi is with the Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH 44106 USA, and also with the Louis Stokes Cleveland Veterans Administration Medical Center, Cleveland, OH 44106 USA (e-mail: axm788@case.edu).

## REFERENCES

[1] CDC, "Coronavirus disease 2019 (COVID-19)," *Centers Dis. Control Prevention*, Feb. 2020. Accessed: Jun. 19, 2020, [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html

[2] D. Wang *et al.*, "Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China," *JAMA*, no. 11, pp. 1061–1069, Mar 2020, doi: 10.1001/jama.2020.1585.

[3] F. Zhou *et al.*, "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study," *Lancet*, vol. 395, no. 10229, pp. 1054–1062, Mar. 2020, doi: 10.1016/S0140-6736(20)30566-3.

[4] P. Huang *et al.*, "Use of chest CT in combination with negative RT-PCR assay for the 2019 novel coronavirus but high clinical suspicion," *Radiology*, vol. 295, no. 1, pp. 22–23, Feb. 2020, doi: 10.1148/radiol.2020200330.

[5] "ACR recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection," Accessed: Mar. 2020, [Online]. Available: https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection

[6] N. Lessmann, et al., "Automated assessment of COVID-19 reporting and data system and chest CT severity scores in patients suspected of having COVID-19 using artificial intelligence," *Radiology* 298, no. 1, pp. E18–E28, Jan. 2021.

[7] X. Mei *et al.*, "Artificial intelligence–enabled rapid diagnosis of patients with COVID-19," *Nature Med.*, vol. 26, no. 8, pp. 1224–1228, Aug. 2020, doi: 10.1038/s41591-020-0931-3.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, Cham, Switzerland, Oct. 2020, vol. 9351, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[9] R. Shojaii, J. Alirezaie, and P. Babyn, "Automatic lung segmentation in CT images using watershed transform," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2005, vol. 2, pp. 2–1270.

[10] S. Eppel, "Setting an attention region for convolutional neural networks using region selective features, for recognition of materials within glass vessels," Aug, 2017, *arXiv:1708.08711*.

[11] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, Sep. 1988, doi: 10.2307/2531595.

[12] H. Kataoka, T. Wakamiya, K. Hara, and Y. Satoh, "Would mega-scale datasets further enhance spatiotemporal 3D CNNs?," Apr. 2020, Accessed: Jan. 2021. [Online]. Available: http://arxiv.org/abs/2004.04968

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

[14] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," Jan. 2018, Accessed: Dec. 2019. [Online]. Available: http://arxiv.org/abs/1608.06993

[15] 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," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[16] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, "Thresholding classifiers to maximize F1 score," May 2014, Accessed: Jul. 2020. [Online]. Available: http://arxiv.org/abs/1402.1892

[17] J. J. Marini and L. Gattinoni, "Management of COVID-19 respiratory distress," *JAMA*, vol. 323, pp. 2329–2330, Apr. 2020, doi: 10.1001/jama.2020.6825.

[18] S. Iwanami *et al.*, "Rethinking antiviral effects for COVID-19 in clinical studies: Early initiation is key to successful treatment," *medRxiv*, Jun. 2020, doi: 10.1101/2020.05.30.20118067.

[19] G. M. Piscitello, E. M. Kapania, W. D. Miller, J. C. Rojas, M. Siegler, and W. F. Parker, "Variation in ventilator allocation guidelines by US state during the coronavirus disease 2019 pandemic: A systematic review," *JAMA Netw. Open*, vol. 3, no. 6, p. e2012606, Jun. 2020, doi: 10.1001/jamanetworkopen.2020.12606.

[20] D. Ellinghaus *et al.*, "Genomewide association study of severe COVID-19 with respiratory failure," *New England J. Med.*, vol. 383, no. 16, pp. 1522–1534, Jun. 2020, doi: 10.1056/NEJMoa2020283.

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2024 at 03:38:49 UTC from IEEE Xplore. Restrictions apply.