BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Development and multicentric external validation of a prognostic COVID-19 severity model based on thoracic CT

Ine Dirks[1,2*], Matías Nicolás Bossa[1], Abel Díaz Berenguer[1], Tanmoy Mukherjee[1], Hichem Sahli[1,2], Nikos Deligiannis[1,2], Emma Verelst[3], Bart Ilsen[3], Simon Van Eyndhoven[4], Lucie Seyler[5], Arne Witdouck[5], Gilles Darcis[6], Julien Guiot[7], Athanasios Giannakis[8,9,10] and Jef Vandemeulebroucke[1,2,3]

Abstract

Background Risk stratification of COVID-19 patients can support therapeutic decisions, planning and resource allocation in the hospital. In times of high incidence, a prognostic model based on data efficiently retrieved from one source can enable fast decision support.

Methods A model was developed to identify patients at risk of developing severe COVID-19 within one month based on their age, sex and imaging features extracted from the thoracic computed tomography (CT). The model was trained on publicly available data from the Study of Thoracic CT in COVID-19 (STOIC) challenge and validated on unseen data from the same study and an external, multicentric dataset. The model, trained on data acquired before any variant of concern dominated, was assessed separately on data collected at later stages of the pandemic when the delta and omicron variants were most prevalent.

Results A logistic regression based on handcrafted features was found to perform on par with a direct deep learning approach, and the former was selected for simplicity. Volumetric and intensity-based features of lesions and healthy lung parenchyma proved most predictive, in addition to patient age and sex. The model reached an area under the curve of 0.78 on the challenge test set and 0.74 on the external test set. The performance did not drop for the subset acquired at a later stage of the pandemic.

Conclusions A logistic regression utilizing features from thoracic CT and its metadata can provide rapid decision support by estimating short-term COVID-19 severity. Its stable performance underscores its potential for real-world clinical integration. By enabling rapid risk stratification using readily available imaging data, this approach can support early clinical decision-making, optimize resource allocation, and improve patient management, particularly during surges in COVID-19 cases. Furthermore, this study provides a foundation for future research on prognostic modelling in respiratory infections.

Keywords COVID-19, Computed tomography, Prognosis, Disease severity, Logistic regression

*Correspondence:
Ine Dirks
ine.dirks@vub.be

Full list of author information is available at the end of the article

Background

The coronavirus disease 2019 (COVID-19) has had a great impact worldwide. As of 19 April 2023, there have been over 700 million confirmed cases and over six million deaths [1]. The global pandemic has posed huge challenges for the medical sector and put enormous pressure on intensive care units (ICUs). Early identification of patients that are at risk of developing severe disease is imperative for optimal patient care and proper resource allocation.

Computed tomography (CT) has been indicated for severity scoring and triage support for high-prevalent times when the availability of reverse transcription polymerase chain reaction (RT-PCR) tests is limited. Patterns suggestive of COVID-19 can be perceived on CT at an early stage [2–6].

A fully automated system that can select patients at risk of a severe outcome based on that same baseline CT scan without requiring additional data or a delay awaiting laboratory results would be of great benefit. Early aggressive therapy in patients at higher risk can be considered to improve prognosis. Moreover, an indication of prospective resource requirements could notably enhance hospital workflow and alleviate the extreme pressure put on the staff.

Related work

A considerable amount of literature has been published on predictive models for COVID-19 prognosis. The following section provides an overview of recent studies leveraging CT-derived parameters alone or combined with demographic, laboratory and/or other clinical features. The cited papers are limited to those evaluating the model on, at least, an internal test set.

Cai et al. [7] assessed the use of random forests for the prediction of the need and duration of admission to the ICU, duration of oxygen inhalation, hospitalisation, positive sputum nucleic acid test and the prediction of prognosis, in the context of partial recovery versus prolonged recovery. For a set of 99 patients, they collected 40 clinical parameters and lungs and lesions were segmented automatically to extract quantitative and textural image features. In repeated ten-fold cross-validation with 100 repetitions, the model for prognosis prediction achieved an area under the curve (AUC) of the receiver operating characteristic (ROC) of 0.96.

Meng et al. [8] proposed De-COVID19-Net, a three-dimensional (3D) densely connected convolutional neural network (CNN) that merges CT-derived features with clinical parameters, including sex, age, severity grade, and a binary indicator for chronic disease, to categorise COVID-19 patients as high or low mortality risk. They used 366 patients from four institutes and about a third from each centre was taken to make up the test set. A total of 70 patients died within 14 days and were labelled as high-risk. The remaining 296 subjects were considered low-risk. The model achieved an AUC of 0.95 on the training set and 0.94 on the test set.

Ning et al. [9] developed a framework named HUST-19 to predict morbidity and mortality outcomes. The former comprised negative control subjects, patients with mild or regular COVID-19 (type I) and severe or critically ill patients (type II) as defined by the Guidance for COVID-19 (sixth edition) released by the National Health Commission of China. The framework consists of different steps, including a CNN-based classification of axial CT slices as non-informative, negative or positive and their integration into a patient-level prediction. A separate deep neural network (DNN) makes a prediction based on clinical features. Then, a penalised logistic regression combines all the information to make a final prediction. They collected data from a total of 1521 patients from two hospitals, including a CT scan and 130 clinical features. The morbidity outcome, trained on the first cohort, achieved AUCs of 0.86, 0.88 and 0.94 in the second cohort for predicting type I, type II and negative patients, respectively. Due to a relatively low number of deceased patients and an unknown mortality status for a portion of the patients, both cohorts were merged to train the mortality prediction. This reached an AUC of 0.86 in a ten-fold cross-validation.

Fu et al. [10] assessed the use of CT-derived radiomic features to classify COVID-19 patients as stable or progressive. A support vector machine (SVM) considering seven radiomic features achieved an AUC of 0.83 in a leave-one-out validation on 64 patients.

Li et al. [11] developed a model to distinguish between severe and critical COVID-19 patients. They collected a total of 217 patients from three institutes and randomly divided them into 80% for training and 20% for testing. Lung masks were automatically segmented and used to extract 102 radiomic features from the CT. Six features remained after a number of selection steps and were combined with features from a 3D-Resnet-10 in a logistic regression, SVM, decision tree, and random forest. The logistic regression demonstrated good performance and generalisability with an AUC of 0.90 in the training set and 0.86 in the testing set.

Yue et al. [12] developed a model to predict a patient's duration of hospital admission as short or long based on a 10-day cutoff. They used a dataset of 52 patients, collected at five institutes of which 26 patients from four centres constituted the training set and the remaining five patients from a different hospital were used for testing. Lungs and lesions were segmented semi-automatically and 1218 radiomic features were calculated per lesion of which six were retained. In a head-to-head comparison, a logistic regression generalised slightly better

than random forest classifiers, achieving an AUC of 0.92 in five-fold cross-validation and 0.97 on the independent test set.

Wu et al. [13] compared four models to predict severe COVID-19, defined as respiratory failure requiring mechanical ventilation, shock, ICU admission, organ failure and/or death during hospitalisation. A dataset of 299 patients was acquired at one centre in China of which 80% was used for training and 20% for internal validation. An additional five sets totalling 426 patients were collected at eight centres in China, Italy and Belgium which were used for testing. Four logistic regression models were trained, each with a different set of features. The first model achieved a validation AUC of 0.83 through the use of baseline clinical features without symptoms. The second model included additional clinical features and reached an AUC of 0.74 in the validation set. The third model combined CT-derived features with age and sex and obtained a validation AUC of 0.83. The final model considered all selected clinical, laboratory and CT-derived features and achieved the highest validation AUC of 0.90. This last one was assessed on the five test datasets and reached scores of 0.84 to 0.93.

Fang et al. [14] investigated the prediction of an unfavourable progression for patients admitted with a mild type of COVID-19. The CT scan, represented by a 128-dimensional feature vector extracted through a 3D ResNet, was combined in a neural network with 61 clinical parameters and processed by a Long Short-Term Memory (LSTM) network. The system reached a mean AUC of 0.92 in a five-fold cross-validation. Additionally, they proposed a domain adaptation method where the model trained on data from one hospital achieved an AUC of 0.86 on the data from a different centre using 10 labelled samples from that institute.

Wang et al. [15] proposed a system for COVID-19 progression prediction using CT-derived and clinical features. A dataset of 1051 patients acquired at nine hospitals was randomly divided into training (70%), validation (10%), and testing sets (20%). Lungs and lesions were automatically segmented and the 10 axial slices with the largest area of affected tissue served as input to a deep learning severity prediction model. This could predict whether or not a COVID-19 patient would develop a critical illness with an AUC of 0.86. The features extracted by this model were combined with clinical data and given to a random survival forest for progression prediction. This achieved AUC scores of 0.82, 0.81, and 0.83 for prediction at three, five and seven days, respectively.

Wang et al. [16] discussed the prediction of COVID-19 progression. They used a training set of 124 patients and a test set of 64 subjects collected at a different hospital. Lungs and lesions were segmented semi-automatically from which radiomic features were extracted. The best-performing model proved to be a logistic regression combining radiomic, clinical and laboratory results and achieved AUC values of 0.92 and 0.87 on the training and test sets, respectively.

In December 2021, the Study of Thoracic CT in COVID-19 (STOIC) [17] launched a COVID-19 artificial intelligence (AI) challenge [18] with the aim of developing models that can predict if a patient will develop severe COVID-19 within one month based on the initial CT scan. The STOIC project collected CT scans, clinical data, RT-PCR test results and outcome at one month of over 10,000 patients treated at 20 different French hospitals. CT scans were classified as positive or negative for COVID-19 by seven junior and 13 senior radiologists. Of the 4238 positive patients, determined by both the CT reading and the RT-PCR test, 1000 developed severe disease within one month, defined as death or the need for intubation. Two logistic regression models were tested for one-month severity prediction. Clinical risk factors included age, sex, oxygen supplementation at presentation, hypertension, and coronary artery disease. A second model included the same features in addition to a coronary artery calcium score and quantified disease extent from the CT analysis. These extra parameters proved predictive with an AUC of 0.69 compared to 0.64 for the first method.

Lassau et al. [19] proposed an AI-severity score constructed of a deep learning model to extract features from the CT scan and five clinical and biological parameters, namely age, sex, oxygenation, urea, and platelet count. The model was trained for the severity outcome defined as an oxygen flow rate of 15L/min or higher, the need for mechanical ventilation or death. The AI-severity method outperformed 11 existing severity scores on an internal holdout set of 150 patients with an AUC of 0.77 and on an external validation set of 135 patients with an AUC of 0.79.

Shiri et al. [20] analysed the use of CT-derived radiomic features to predict the overall survival of COVID-19 patients. In a dataset of 14,339 patients from 19 different institutes, they extracted 107 radiomic features from CT using automatically-segmented lung masks. Optimal models were searched through four feature selection algorithms, seven classifiers and 10 different splitting and cross-validation strategies. The best result reached an AUC of 0.83. This included a relief-based feature selection with random forest classifier, trained on 70% of each institute's dataset and tested on the remaining data.

Kienzle et al. [21] adapted the ConvNeXt architecture to process 3D CT images and pretrained it in different ways. First, a 2D version was trained on grayscale images of the ImageNet dataset and the weights were converted to 3D via inflation. Next, the network was trained for segmentation, once using a lung lesion segmentation dataset

and once on the COV19-CT-DB [22] after generating pseudo-labels. Pseudo-labels were also generated for the STOIC data to train for severity classification using real labels and segmentation using pseudo-labels for the STOIC dataset. This resulted in a framework for severity prediction and infection detection based on CT analysis. The method achieved an F1 score of 0.68 in five-fold cross-validation and 0.49 on the official COV19-CT-DB test set. The performance was validated by ranking second and third in two challenges with respective F1 scores of 0.51 and 0.86.

Duan et al. [23] developed two CT-based radiomics models to predict COVID-19 progression. The total set of 44 patients was divided into aggravating and relief groups. For each patient, the slice with the largest lesion was selected and the lesion was delineated manually. From this, 782 radiomic features were extracted, of which 10 were used to build the prediction model. A second model was developed by extracting the same features from a scan taken one to two weeks after the first CT and subtracting the corresponding feature values. The first and second models achieved AUCs of 0.99 and 1.00, respectively, in a 10-fold cross-validation.

In summary, a considerable amount of work has been done on predictive model development for COVID-19 prognosis using CT-derived parameters. The related work described above is summarized in Table 1. Reported AUC values (0.77–0.95) and dataset characteristics vary largely across studies. Several studies included small datasets with less than 100 patients and only a limited number of studies performed an external validation on data acquired at a different centre. In addition, many methods require inputs on top of the CT image with its metadata, leading to longer processing times (e.g., awaiting lab results) which is a disadvantage for emergency situations.

Goal and contributions of this study

The primary aim of this work, classified as type 3 according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [24], was the development, internal and external validation of a predictive model for COVID-19 prognosis at one month. A key differentiator of this work is its focus on leveraging thoracic CT scan data in Digital Imaging and Communications in Medicine (DICOM) format. Age and sex, read from the metadata, were combined with a number of features extracted from the imaging data to create a model that requires all input features to originate from a single source. This facilitates seamless integration in clinical routine, a challenge often overlooked in similar studies but of utmost importance in emergency settings.

To achieve this, we explored and compared two fundamentally different and widely used modelling approaches: logistic regression, a traditional statistical method, and deep learning. By directly comparing these methods, we provide insights into their relative performance, robustness, and applicability in real-world scenarios.

Table 1 Overview of related literature

Publication	n	Methods	Input	Performance
Cai et al. [7]	99	random forest	automated lung and lesion segmentation, 40 clinical parameters	cv10 AUC = 0.96
Meng et al. [8]	366	CNN	CT, sex, age, severity grade, chronic disease status	test set AUC = 0.94
Ning et al. [9]	1521	DNN, CNN, logistic regression	CT, 130 clinical features	cv10 AUC = 0.86
Fu et al. [10]	64	SVM	CT-derived radiomics	leave-one-out AUC = 0.83
Li et al. [11]	217	CNN, logistic regression, SVM, decision tree, random forest	CT-derived radiomics	test set AUC = 0.86
Yue et al. [12]	52	logistic regression, random forest	CT-derived radiomics	independent test set AUC = 0.97
Wu et al. [13]	725	logistic regression	semi-automatically derived CT findings, 7 clinical features	independent test set AUC = 0.84 to 0.93
Fang et al. [14]	1040	CNN, RNN	CT, 61 clinical parameters	cv5 AUC = 0.92, domain adaptation AUC = 0.86
Wang et al. [15]	1051	CNN, DNN, random forest	CT, 15 clinical parameters	test set AUC = 0.81 to 0.83
Wang et al. [16]	188	logistic regression	CT-derived radiomics, 24 clinical parameters	test set AUC = 0.87
Revel et al. [17]	10735	logistic regression	manually-derived CT findings, 7 clinical parameters	AUC = 0.64
Lassau et al. [19]	1003	DNN	CT, 5 clinical parameters	test set AUC = 0.79
Shiri et al. [20]	14339	random forest	CT-derived radiomics	test set AUC = 0.83
Kienzle et al. [21]	2476	CNN	CT	test set F1 = 0.49
Duan et al. [23]	44	random forest	CT-derived radiomics	cv10 AUC = 0.99 to 1.00

n: total number of patients, cvx: x-fold cross-validation, DNN: deep neural network, CNN: convolutional neural network, RNN: recurrent neural network, SVM: support vector machine

A significant strength of this study lies in the used datasets. The selected model was developed on a public dataset collected in March and April of 2020 and internally validated on a large, private dataset from the same initiative. In addition, a dataset of 1318 patients was collected from different institutions and countries than the training set, and used for external validation. This approach contrasts with many existing studies, which often rely on much smaller datasets [7, 8, 10–13, 16, 23] and/or lack validation on external test sets from different institutions [7–11, 15–17, 19–21, 23], making it impossible to determine their generalizability. Furthermore, to assess the robustness and adaptability of our model, the performance was assessed separately on data collected at later stages of the pandemic when the delta and omicron variants were most prevalent.

As a secondary aim, we investigated the feasibility of a long-term prediction model to identify patients who still experience COVID-19-related symptoms three months after the initial positive RT-PCR test. A smaller dataset was employed, enriched with more detailed patient information obtained from lab results. We report our initial findings on model development using this single dataset.

Methods

This section describes the steps taken for the development and validation of the proposed prognostic models as illustrated in Fig. 1. The study was approved by the Institutional Review Board/Ethics Committee of Universitair Ziekenhuis Brussel (Commissie Medische Ethiek O.G. 016, EC-2023-014), of Centre Hospitalier Universitaire de Liège (committee reference 707, study references 2020/139 and 2022/21) and of Universitätsklinikum Heidelberg (S-293/2020).

Training dataset

A prediction model for short-term prognosis was developed based on the public data of the STOIC challenge. Access was provided to part of the associated database, consisting of a total of 10,735 patients of which 6448 had a positive RT-PCR test. Of the latter, 1602 developed severe COVID-19, defined as death or the need for intubation within one month. Besides the CT scan, the patient sex and age were available, where the latter was assigned according to ranges of 10 years. The dataset was randomly divided into a public training set, three private test sets and a private training set. The former was used for model development and consisted of 2000 patients (1148 male, 852 female) with a median age of 65 years old [35–85]. A total of 1205 people tested positive and 301 were classified as severe [25].

To allow for feature and classifier selection, the available dataset was divided into five equally-sized sets. Four sets were used for development, while one set was kept aside for further model comparison and ensemble testing. The subsets were stratified for COVID status, severity outcome, age and sex.

The long-term prognosis prediction model was developed using part of the dataset described by Darcis et al. [26]. The aim was to create a model for predicting a patient's likelihood to still experience symptoms at three months, plus or minus three weeks after the first positive RT-PCR test. Moreover, it was evaluated if the addition of other clinical data could improve the performance of this task.

For 149 patients, the CT and three-month follow-up data were available. In this group, 43 were admitted to the ICU at some point and 102 still experienced one or more symptoms three months after recording the first symptom. The median number of days until the follow-up closest to three months was 96.5 [3.0–293.0]. Recorded comorbidities included: chronic renal failure (13), diabetes (52), arterial hypertension (82), dyspnea (53), severe pathology (41), chronic pneumoptology (24), immunosuppression (9), smoking (80), asthma (16), chronic obstructive bronchopneumonia (9) and active cancer (14). Collected biological data at discharge, summarised in Table 2, included: levels of C-reactive protein (CRP), creatinine, white blood cell, haemoglobin, lymphocytes, aspartate aminotransferase, alanine aminotransferase, creatinine kinase, D-dimer, ferritin and glomerular filtration rate. Missing data were replaced by the median taken over the remaining patients.

Dataset for external validation

A large, multicentric dataset was collected for external validation from three institutions: Centre Hospitalier Universitaire de Liège (CHUL, Liège, Belgium), Universitätsklinikum Heidelberg (UKHD, Heidelberg, Germany) and Universitair Ziekenhuis Brussel (UZB, Brussels, Belgium). For all patients, we collected the chest CT scan and corresponding acquisition date, the patient's age, sex and the dates of intubation and/or death if applicable. At UZB, additional data were collected through a query of the registry of severe acute respiratory infections (SARI) [27]. The subset for which all required data were available comprised 982 cases from CHUL, 28 cases from UKHD and 308 subjects from UZB.

In contrast to the training data, which were gathered when there were no variants of concern (VOCs), the external validation set contained data collected at a later stage when certain variants dominated and more people were vaccinated. For these, CT predictors might be less pronounced and the performance of the initial model needs to be evaluated. Since the COVID-19 variant was not recorded, this was assigned based on the scan date and the information on the Belgian epidemiological situation [28], which was assumed to be similar for Germany.

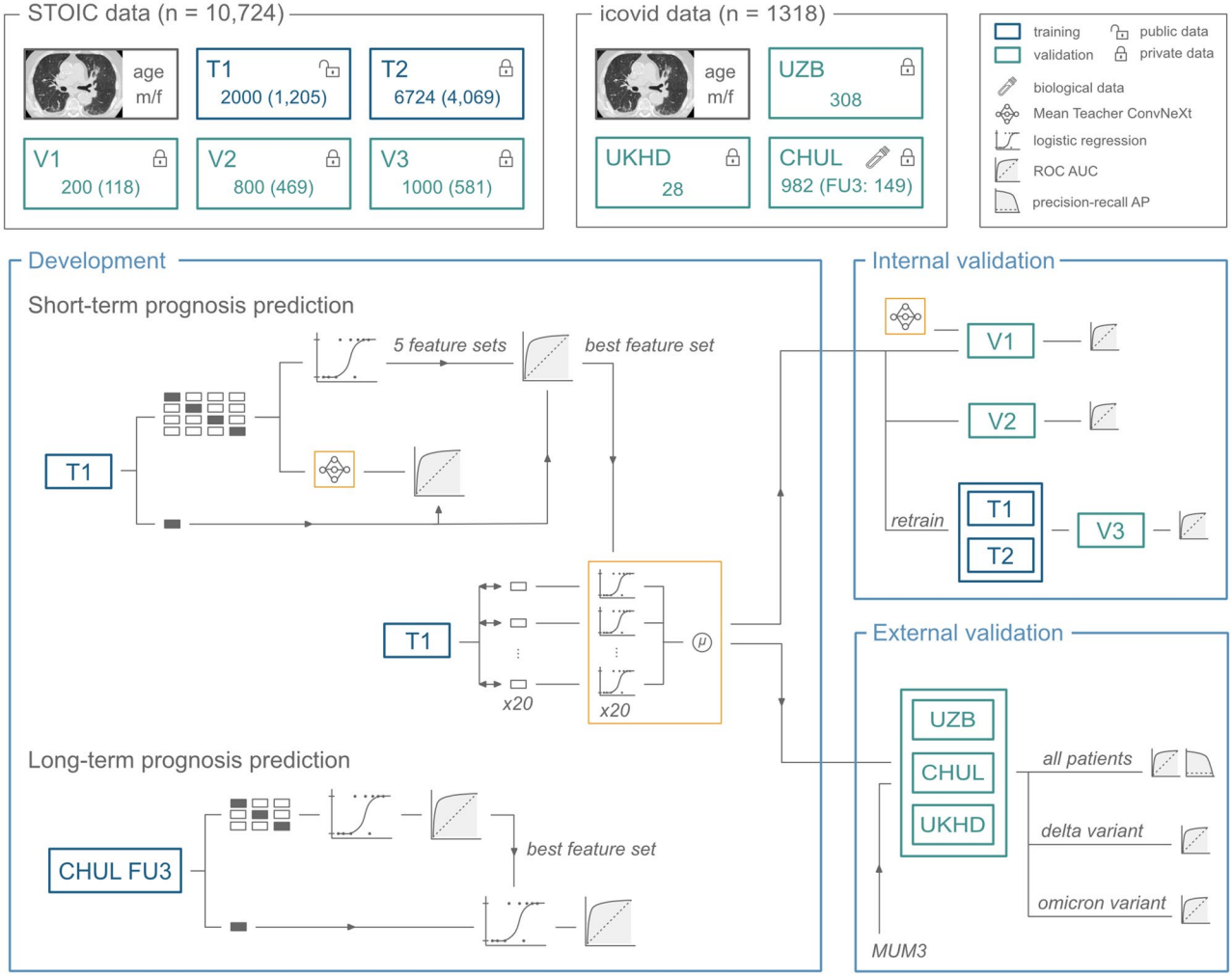

Fig. 1 Overview of the steps taken for development and validation of the proposed prognostic models and their respective datasets. The Study of Thoracic CT in COVID-19 (STOIC) challenge dataset consisted of one public set and four private sets used for several training (T) and validation (V) steps. Per set, the amount of samples is given with the number of covid-positive patients in parentheses. The icovid dataset consisted of covid-positive patients acquired at Universitair Ziekenhuis Brussel (UZB), Universitätsklinikum Heidelberg (UKHD) and Centre Hospitalier Universitaire de Liège (CHUL). In the latter, three-month follow-up (FU3) data were available for a subset of 149 patients. For all images, the lungs and lung lesions were segmented and handcrafted features were extracted. This is omitted from the scheme to avoid needlessly convoluting the figure. **Development** focused on one-month severity and three-month symptomatology. For the latter, preliminary tests were performed in a three-fold cross-validation and on a holdout set created from the subset with three-month follow-up at CHUL. For the one-month severity prediction, both a deep model and a logistic regression exploiting handcrafted features were developed in a four-fold cross-validation and tested on a holdout set created from the public STOIC data. For the logistic regression, five feature sets achieved the same area under the curve (AUC) of the receiver operating characteristic (ROC) in the four-fold cross-validation. The combination of features with the highest AUC on the holdout set was selected to continue. Next, an ensemble model was created, averaging the probabilities predicted by 20 logistic regression models, each trained on 2000 samples selected through sampling with replacement from the public STOIC data. **Internal validation** was performed through the STOIC challenge and its private datasets. After validating both the deep model and ensembled logistic regression approach on a first private validation set, preference was given to the latter method, which was then tested on a second private set. In the final stage of the challenge, the algrithm was retrained on both training sets and validated on the remaining data. **External validation** was performed for the ensemble model on the multicentre icovid dataset and through a comparison to the Maastricht University Model 3 (MUM3). In addition to the AUC, the precision-recall curve with average precision (AP) was evaluated. Besides the full set of patients, two subsets acquired in the respective timespans where the delta and omicron variants were most prevalent were considered

The alpha variant was designated to any scan acquired between 1 March 2021 and 28 June 2021. Only six scans belonged to this category, so no separate test was performed on this group. The period to the next cutoff, on 1 January 2022, was linked to the delta variant, comprising 40 patients. After that, omicron was the most prevalent VOC. No distinction was made between the different omicron sublineages due to a lack of data representing each subvariant. This resulted in a total of 35 patients from the omicron stage.

An overview of the external validation set is provided in Table 3.

Table 2 Biological data at discharge

Parameter	n	m
C-reactive protein (mg/L)	133	21
Creatinine	133	0.84
White blood cells (10^3/mm^3)	133	5.9
Haemoglobin (g/dL)	131	12
Lymphocytes (10^3/mm^3)	131	1.5
TGO (ASAT) aspartate aminotransferase (U/L)	125	35
TGP (ALAT) alanine aminotransferase (U/L)	125	45
Creatinine kinase (UI/L)	64	60
D-dimer (µg/L)	50	922
Ferritin (µg/L)	23	817
Glomerular filtration rate estimation (CKD-EPI)	133	90

n: number of patients for which the parameter was available, m: median value

Table 3 Patient characteristics of the validation set for the short-term and long-term predictions

Patient Characteristics	Short-term				Long-term
	All (n = 1318)	Alpha (n = 6)	Delta (n = 40)	Omicron (n = 35)	(n = 149)
Age (median[range])	69 [1–104]	58.5 [50–87]	60.5 [1–93]	77 [26–95]	63 [26–87]
Age category (n (%))					
<40 years	84 (6.37)	0 (0)	5 (12.5)	3 (8.57)	11 (7.38)
40–50 years	107 (8.11)	0 (0)	10 (25.0)	1 (2.86)	16 (10.7)
50–60 years	202 (15.3)	3 (50.0)	4 (10.0)	1 (2.86)	33 (22.1)
60–70 years	298 (22.6)	1 (16.7)	5 (12.5)	5 (14.3)	46 (30.9)
70–80 years	297 (22.5)	0 (0)	8 (20.0)	10 (28.6)	33 (22.1)
>80 years	330 (25.1)	2 (33.3)	8 (20.0)	15 (42.9)	10 (6.71)
Sex (n (%))					
male	823 (62.4)	1 (16.7)	17 (42.5)	25 (71.4)	98 (65.8)
female	495 (37.6)	5 (83.3)	23 (57.5)	10 (28.6)	51 (34.2)
Short-term severity (n (%))					
severe	432 (32.8)	0 (0)	3 (7.50)	6 (17.1)	119 (79.9)
not severe	886 (67.2)	6 (100)	37 (92.5)	29 (82.9)	30 (20.1)
Long-term symptoms (n (%))					
yes	–	–	–	–	102 (31.5)
no	–	–	–	–	47 (68.5)

n: number of patients

Image segmentation

On the CT scans, the lungs were segmented using an open-source model [29]. A postprocessing step was added, retaining only the two largest components with a minimum size of 10mL to exclude any regions outside the lungs that may have been segmented. Lung lobes were segmented with the LTRCLobes model [29]. Every voxel that was classified as background in the lobe segmentation but as foreground in the lung mask received the label of its nearest neighbour.

Lung lesions composed of consolidation and ground glass opacity (GGO) were segmented by a previously developed model. In brief, the nnU-Net [30] implementation of Monai [31] was employed with deep supervision and the sum of the mean Dice loss and the cross-entropy as loss function. The network was trained on 199 patients from the COVID-19 Lung CT Lesion Segmentation Challenge [32], 69 CT scans and manual lung lesion segmentations acquired within the icovid project [33], 70 samples from the public dataset of the COVID-19 Pneumonia Lesion segmentation network (COPLE-Net) [34] and 10 scans from the publicly available COVID-19 CT Lung and Infection Segmentation Dataset [35].

Short-term prognosis prediction

In this study, parameters were limited to those that can be extracted from the DICOM data (i.e., the CT image and associated metadata). Therefore, available features for model development included combinations of patient age and sex with image-derived features. Two main approaches were investigated: an end-to-end deep model and a logistic regression model using handcrafted radiomic features. The rationale for comparing these two approaches is twofold. First, logistic regression is a simple, well-established, and interpretable method that is unlikely to overfit and is often used in clinical settings. It allows for direct analysis of feature importance and easy implementation in resource-constrained environments. Second, deep learning methods, while computationally more demanding and less interpretable, have demonstrated substantial promise in medical imaging tasks, particularly when working directly with raw image data. Deep learning models are capable of capturing more complex relationships but carry higher risk of overfitting. By comparing these two fundamentally different approaches, we aim to evaluate whether the increased complexity of deep learning offers significant performance advantages over simpler, interpretable models in the context of short-term prognosis prediction.

For the deep model, the Mean Teacher method [36] was adopted using the ConvNeXt architecture [37], modified to process 3D images. This semi-supervised learning technique effectively leverages both labelled and unlabelled data by minimizing the consistency loss between a student and teacher model. By utilizing unlabelled data, which is more readily available in clinical datasets, the method helps improve model performance, particularly when labelled data is limited. The consistency regularization inherent to the Mean Teacher method further enhances robustness by encouraging the model to produce stable predictions under slight perturbations of the

input data (e.g., noise or augmentations), leading to better generalization on unseen datasets.

The CT images were masked with the lung segmentation to remove irrelevant information. A logistic regression model was trained on age and sex to predict severity. The logit of the predicted probabilities was included as an extra bias neuron to the last layer before the output layer to incorporate age and sex information to the model in a controlled way. The network was initialised with pretrained weights for COVID-19 diagnosis, obtained using COV19-CT-DB [22].

Given the considerably lower training times for logistic regression, several hand-crafted features and their combinations could be evaluated. From the lung and lesion segmentations, the number of lesions were extracted as well as volume fractions of diseased tissue and radiomic features. For the latter, features derived by SimpleITK and pyradiomics were tested.

In SimpleITK 2.0.2, the mean and median intensity, standard deviation, flatness, kurtosis, roundness and skewness were collected for the healthy lung tissue, consolidation and GGO separately and rescaled to [−1, 1]. For lesion-free patients, the values for consolidation and GGO were set to the healthy tissue value of that patient.

With pyradiomics 3.0.1, radiomic features were extracted from the 3D volumes of interest at an intensity bin width of 25 Hounsfield units (HU). The following two publications were used as a starting point for the radiomic features. Chen et al. [38] assessed the performance of SVM models for the diagnosis of COVID-19. The training and testing sets were made up of 112 and 22 patients, respectively, with COVID-19 or another type of pneumonia. Four groups of features were defined: radiomics, radiologic, quantifying, and clinical features. The method comprising all of these outperformed the ones including only one group and reached a ROC AUC of 0.92 on the test set. Huang et al. [39] evaluated models for discriminating COVID-19 and influenza pneumonia through logistic regression. Features from CT annotations, performed by three radiologists, were tested as well as 1316 radiomic features. After some feature selection steps, lesion distribution, GGO, intralobular interstitial thickening and halo sign were retained from the CT reading together with seven radiomic features. The model built on both groups of features achieved a superior AUC of 0.96 in a set of 153 COVID-19 and influenza pneumonia patients. Based on the most predictive features described by Chen et al. [38] and Huang et al. [39], a total of 168 features were derived, including 24 first-order features, 48 gray level co-occurrence matrix (GLCM) features, 36 gray level size zone matrix (GLSZM) features, 36 gray level run length matrix (GLRLM) features and 24 gray level dependence matrix (GLDM) features. Missing values for these features were replaced with the median taken over the healthy subjects in the training set. The volume-weighted average of the three largest lesions for each feature was selected and normalised through z-scoring.

A baseline logistic regression model was initially trained using only the patient's age and sex. To identify the optimal set of image features, the increase in ROC AUC during cross-validation was evaluated as features were added. The extensive set of features extracted through pyradiomics was reduced to three using both univariate and multivariate feature selection methods and results were compared. For univariate feature selection, the *SelectKBest* module of scikit-learn 0.24.2, based on the F-statistic between label and target, was employed. However, this approach returned a lot of correlated features, prompting the use of a multivariate approach. The *SelectFromModel* module was subsequently applied to identify the best combination of three features. Various feature combinations were considered, including the ones identified through this uni- and multivariate features selection, features identified by Chen et al. [38] and Huang et al. [39] and those extracted using SimpleITK. The AUC values for these feature combinations were assessed through cross-validation. The best-performing models from these experiments were retrained on data from four folds and evaluated on the holdout set. The model achieving the highest AUC was selected to perform all further validation. The statistical significance of the difference in AUC was evaluated with the DeLong test [40, 41].

Comparison to the Maastricht University Model 3

The Maastricht University Model 3 (MUM3), corresponding to the third model described by Wu et al. [13], is a publicly available predictive model that had been integrated into MyPatientCheck [42] at an earlier stage and was used to benchmark the proposed model. The choice to compare against MUM3 was motivated by the following reasons. At the time of writing, MUM3 was one of the few publicly available models for COVID-19 prognosis, ensuring access for benchmarking. MUM3 was trained on a relatively large dataset and achieved good performance, making it a relevant and credible baseline for comparison. In addition, the model relies solely on age and the total CT-derived lung involvement score. This aligns closely with one of the key goals of our study to develop models that efficiently leverage data derived from a single source (DICOM-based thoracic CT scans). The total CT-derived lung involvement score was determined by the sum of the involvement scores per lobe based on the lesion fraction, for which the automated segmentations were used.

Long-term prognosis prediction

The short-term prognosis prediction model estimates the probability of developing severe COVID-19 within one month of acquiring the first CT scan. However, COVID-related symptoms can persist for months after the initial infection.

For the subset of patients for whom three-month follow-up data were available, the features derived for short-term prediction were combined with the extra parameters described in section "Training dataset". The data were split randomly into four sets, stratified on the outcome. Three subsets were used for cross-validation experiments and recursive feature elimination, while one set was kept aside to test the final model on. The number of features was reduced through an iterative process, each time selecting the feature pair with the highest Pearson correlation coefficient and removing one of them. This was repeated until no sets with a coefficient of 0.5 or higher remained. A parameter search was performed on the folds to select the optimal settings for a logistic regression model.

Table 4 Mean receiver operating characteristic (ROC) area under the curve (AUC) ± standard deviation (SD) from the four-fold cross-validation for the most relevant models that were tested

Model	Mean AUC ± SD
Mean Teacher ConvNeXt	**0.79** ± 0.033
Logistic regression with features added to the baseline model:	
None	0.65 ± 0.031
Number of lesions	0.66 ± 0.037
Volume fractions	**0.74** ± 0.045
Volume fraction per lung lobe	0.71 ± 0.039
Mean intensity, kurtosis and skewness	0.72 ± 0.033
Volume fractions, mean intensity, kurtosis and skewness	**0.74** ± 0.033
Radiomic features from Chen et al. [38]	0.69 ± 0.025
Radiomic features from Huang et al. [39]	0.72 ± 0.037
Volume fractions, radiomic features from Huang et al. [39]	0.73 ± 0.040
Best three radiomic features from univariate selection	**0.74** ± 0.037
Volume fractions, best three radiomic features from univariate selection	**0.74** ± 0.038
Best three radiomic features from multivariate selection	0.71 ± 0.032
Volume fractions, best three radiomic features from multivariate selection	**0.74** ± 0.029

For the logistic regression, the baseline model considers the patient age and sex. The other regression models build on the baseline through the addition of different features. The number of lesions and the volume fractions include the respective values for ground glass opacity (GGO) and consolidation separately. Mean intensity, kurtosis and skewness include the values for healthy lung parenchyma, GGO and consolidation separately. Best three radiomic features from univariate feature selection: lbp-3D-k_glszm_ZoneVariance, original_shape_Maximum2DDiameterColumn, lbp-3D-m1_glrlm_LongRunLowGrayLevelEmphasis. Best three radiomic features from multivariate feature selection: wavelet-HLL_glcm_MaximumProbability, wavelet-LLL_glrlm_HighGrayLevelRunEmphasis, wavelet-LHL_glcm_Correlation

Results

Short-term prognosis prediction

The baseline logistic regression, including only patient age and sex, obtained a mean AUC of 0.65 ± 0.031. The addition of image-derived features was found to always improve this performance. The most relevant results from the cross-validation experiments are summarised in Table 4.

The deep model achieved the highest AUC of 0.79 ± 0.033 on the cross-validation. The logistic model based on handcrafted features reached an average AUC over the four folds of 0.74 for five different combinations of features.

Volume fractions per lobe per lesion type proved less helpful than overall volume fractions per lesion type, with a mean AUC of 0.71 versus 0.74. Though the fractions per lobe add spatial and physiological information to the model, the more challenging nature of their segmentation may have led to inferior results. Adding mean intensity, kurtosis and skewness for each lesion type and healthy lung parenchyma to the volume fractions did not yield an increase in ROC AUC. Still, they did slightly reduce the variation between the folds. Models employing more advanced radiomic features performed comparably or worse.

To select the preferred regression model, the five best-performing ones were retrained on all patients from the folds and evaluated on the holdout set. ROC curves are visualised in Fig. 2 and AUC scores are summarised in Table 5. Results are similar except for the logistic regression considering patient age, sex, the volume fractions of GGO and consolidation, the mean intensity, kurtosis and skewness of healthy tissue, GGO and consolidation. The ROC AUC of 0.82 is significantly higher than the scores of the other models, equalling 0.74 for the volume fractions ($p = 1.62 \times 10^{-3}$), 0.71 for pyradiomics features from univariate feature selection ($p = 5.94 \times 10^{-5}$), 0.72 for the volume fraction and pyradiomics features from univariate feature selection ($p = 1.04 \times 10^{-4}$) and 0.73 or the volume fraction and pyradiomics features from multivariate feature selection ($p = 1.11 \times 10^{-3}$). With a significance level of 0.05 and a Bonferroni correction for four comparisons, the difference is significant in all cases ($p < 0.0125$). Moreover, the result is comparable to the AUC of 0.81 achieved by the Mean Teacher ConvNeXt model.

Internal validation was performed through participation in the STOIC challenge. For the deep model, the version trained on the 1600 patients from the four folds was submitted. The selected regression model was retrained on all 2000 patients. To reduce the variance that was observed between the folds, an ensemble model was created through bagging, averaging the probabilities predicted by 20 different models. Each of these was

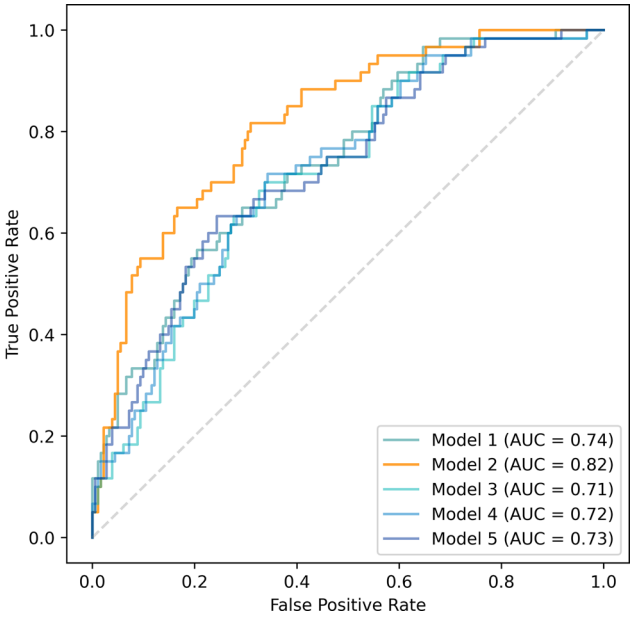

Fig. 2 Receiver operating characteristic (ROC) plots with their area under the curve (AUC) on the holdout set. Each model includes different features on top of patient age and sex. Lesion types include ground glass opacity (GGO) and consolidation. Tissue types refer to GGO, consolidation and healthy lung parenchyma. Model 1: volume fractions per lesion type, Model 2: volume fractions per lesion type, mean intensity, kurtosis and skewness per tissue type, Model 3: best 3 radiomic features from univariate feature selection, Model 4: volume fractions per lesion type, best 3 radiomic features from univariate feature selection, Model 5: volume fractions per lesion type, best 3 radiomic features from multivariate feature selection

Table 5 Receiver operating characteristic (ROC) area under the curve (AUC) on the holdout set of the best models from the four-fold cross-validation experiments

Model	AUC
Mean Teacher ConvNeXt	0.81
Logistic regression with features added to the baseline model:	
Volume fractions	0.74
Volume fractions, mean intensity, kurtosis and skewness	**0.82**
Best three radiomic features from univariate selection	0.71
Volume fractions, best three radiomic features from univariate selection	0.72
Volume fractions, best three radiomic features from multivariate selection	0.73

For the logistic regression, the baseline model considers the patient age and sex. Volume fractions include the respective values for ground glass opacity (GGO) and consolidation separately. Mean intensity, kurtosis and skewness include the values for healthy lung parenchyma, GGO and consolidation separately. Best three radiomic features from univariate feature selection: lbp-3D-k_glszm_ZoneVariance, original_shape_Maximum2DDiameterColumn, lbp-3D-m1_glrlm_LongRunLowGrayLevelEmphasis. Best three radiomic features from multivariate feature selection: wavelet-HLL_glcm_MaximumProbability, wavelet-LLL_glrlm_HighGrayLevelRunEmphasis, wavelet-LHL_glcm_Correlation

trained on 2000 samples, selected through sampling with replacement.

Both the deep model and logistic regression were evaluated on a private qualification test set of 200 samples, achieving AUC scores of 0.74 and 0.75, respectively. Preference was given to proceed with logistic regression for a number of reasons. Though the deep model outperformed the logistic regression in the four-fold cross-validation, a similar performance was observed on the holdout and qualification test sets. These results were deemed to not outweigh the considerably higher model complexity. Moreover, deep models do not necessarily perform as well on a new dataset and may be challenging to adapt to new data [43]. In contrast, the logistic regression is expected to generalise well and is easier to interpret.

The selected regression model achieved an AUC of 0.78 on the final, private qualification set of 800 patients, corresponding to fourth place. In the final round of the challenge, reserved for the ten best-performing submissions, the model was retrained on 8724 patients and evaluated on the test set of 1000 patients, reaching a comparable AUC score of 0.77 and ranked sixth [44].

External validation

The external dataset was used for the external validation of the model and to conduct a performance comparison with MUM3. We should note a discrepancy in the definition of severe COVID-19, defined as the need for mechanical ventilation, shock, ICU admission, organ failure or death during hospitalisation for MUM3. The number of patients that would be classified differently is assumed negligible. Generally, any patient going into shock as a result of COVID-19 or experiencing organ failure will be admitted to the ICU and often require intubation. In the external validation set, the fraction of patients admitted to the ICU without intubation, and thus considered not severe according to the proposed model, was 4%.

The final performance of the proposed model and MUM3 was quantified through the ROC AUC value and the average precision (AP). The statistical significance of the difference in AUC was evaluated with the DeLong test [40, 41]. The 95% confidence interval (CI) around the ROC and precision-recall curves was calculated by bootstrapping the predictions with replacement in 1000 iterations. Results for the proposed model and MUM3 on the external validation set are summarised in Table 6 and visualised in Fig. 3. The proposed model obtained an AUC of 0.74 (95% CI: 0.72–0.77) versus 0.68 (95% CI: 0.65–0.71) (DeLong $p = 9.59 \times 10^{-6}$) achieved by MUM3.

In this set of 1318 patients, 40 date from when the delta variant was most prevalent while 35 are from the period of the omicron waves. AUC scores on the subset of

Table 6 Overview of the validation data and the receiver operating characteristic (ROC) area under the curve (AUC) for the proposed model and Maastricht University Model (MUM3)

Evaluated on	n	n_{severe}	AUC proposed	AUC MUM3	DeLong *p*-value
Challenge qualification	469	118	0.78	–	–
Validation set	1318	432	0.74	0.68	9.59 e-06
Validation set delta	40	3	0.86	0.84	0.907
Validation set omicron	35	6	0.82	0.76	0.609

n: number of patients, n_{severe}: number of severe patients

patients from the delta period are 0.86 for the proposed model and 0.84 for the MUM3 model. For the omicron period, these become 0.82 and 0.76, respectively. On the data from these later stages, there is no longer a significant difference in AUC between the two models (DeLong $p > 0.05$).

To illustrate the effect of the different features on the estimated prognosis, an interactive visualisation of the short-term prognostic model is available in the supplemental material online.

Long-term prognosis prediction

When retraining the short-term severity model on the data for long-term prognosis, the mean AUC through three-fold cross-validation was 0.52 ± 0.028. Age and sex were found to not be predictive of the persistence of symptomatology. A model considering only these two parameters achieved a mean AUC of 0.42 ± 0.098. The best performance was obtained when taking into account kurtosis for healthy tissue, GGO and consolidation, ICU admission, the level of white blood cells, TGO (ASAT) aspartate aminotransferase and TGP (ALAT) alanine aminotransferase. The model achieved an AUC of 0.63 ± 0.073 in the 3-fold cross-validation and 0.62 applied to the holdout set.

Discussion

We developed a predictive model for COVID-19 severity at one month and considered two main approaches during model development: a deep learning model operating directly on the lung field and a logistic model using handcrafted features extracted from segmented lesions and healthy lung tissue. Both approaches led to similar performance during model development, and the logistic model was selected for simplicity.

Alternate methods

Methods based on gradient boosting, like LightGBM [45] and XGBoost [46], have been described as superior for tabular data [43, 47]. Preliminary tests were performed applying these methods to the hand-crafted features alone and in combination with logits derived by the Mean Teacher ConvNeXt. However, no performance improvement was achieved. A possible explanation lies in the fact the training dataset in this work was considerably smaller than those used in the original papers and benchmarking publications, where sample sizes ranged from 7000 to 1.7 billion.

Alternatively, a deep model has several orders of magnitude more degrees of freedom and is therefore expected

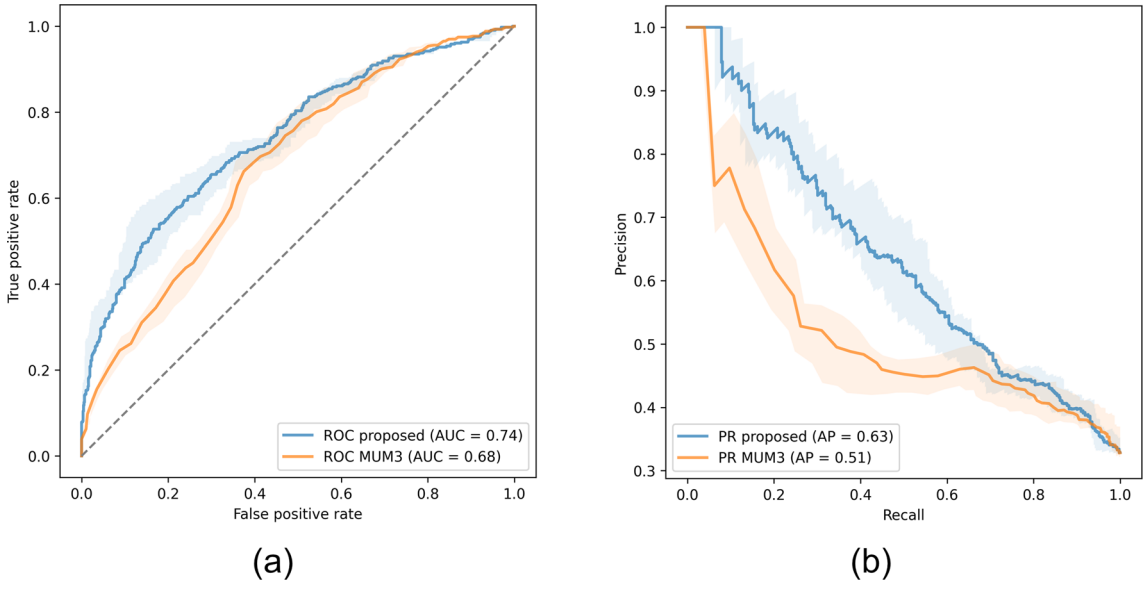

Fig. 3 Receiver operating characteristic (ROC) with the area under the curve (AUC) (**a**) and precision-recall curve (PR) with average precision (AP) (**b**) with 95% confidence interval determined through bootstrapping for the proposed model and Maastricht University Model 3 (MUM3) applied to the validation set

to benefit more from a larger training set. Indeed, the top performing models in the final phase of the STOIC challenge, in which models were retrained on more than 8000 patients, adopted a deep learning backbone and benefitted from a minor performance boost, reaching 0.81 in terms of AUC.

Radiomic analysis

The final logistic regression model exploits parameters extracted from a thoracic CT scan and its DICOM metadata. The addition of image-derived features on top of patient age and sex improved the performance of the baseline model in all cases. These features—volume fractions, mean intensity, kurtosis, and skewness for healthy lung, GGO, and consolidation—capture the distribution, composition, and extent of pathological changes in the lungs. The volume fraction is a critical indicator of the extent of COVID-19. The mean intensity represents the average radiodensity within a specific region which provides insights into the severity of the affected region, the tissue composition and aeration correlating with inflammation or fibrosis. Kurtosis measures the peakedness of intensity distributions. A higher kurtosis in GGO or consolidation regions suggests a non-homogeneous tissue response, potentially indicative of localized severe inflammation or fibrosis, which are markers of advanced disease progression. Skewness quantifies the asymmetry of intensity distributions. A positive skewness in GGO or consolidation regions may reflect a predominance of lower-intensity voxels, often associated with early or mild inflammation, whereas a negative skewness might indicate higher-intensity voxels related to denser fibrotic tissue. An additional file shows an interactive plot to illustrate the contribution of each feature [see Additional file 1]. Advanced radiomic features were not found to improve performance in our experiments. The initial set of radiomic features considered was inspired by literature. However, the features described by Chen et al. [38] and Huang et al. [39] did not offer optimal results, which can be explained by the fact that these features were initially proposed for COVID-19 diagnosis, not severity prediction.

More extensive radiomic analysis was performed using the set of features extracted by pyradiomics, and conducting aggressive feature reduction to a size of three through uni- and multivariate feature selection, but this did not lead to improved performance. Nonetheless, we do not exclude that more careful examination of such features may prove to be beneficial. In particular, specific features describing COVID-19 acute lung injury, such as extensive parenchymal disease and/or lung vasculitis [48–51] seem promising.

Comparison to the state of the art

Literature reports AUC values ranging from 0.77 to 0.95 but it is important to note that these values cannot be directly compared to the performance of the proposed model due to significant differences in patient data distribution. Furthermore the prediction tasks, severity definitions, and evaluation protocols differ between studies. Nevertheless, where possible, we conducted fair comparisons which demonstrated that our model achieves state-of-the-art performance. To support this claim, we provided three levels of comparison.

First, the selected feature combination was validated with respect to performance and generalisability. Achieving similar results when training and testing on different data indicated the robustness of the model. The method proved comparable to other state-of-the-art approaches by ranking fourth out of a total of 120 submissions made by 30 international teams to the first STOIC leaderboard and 20 teams that contended for final qualification. In this final qualification, the difference in AUC with respect to the first place was only 0.029.

Second, besides participation in the challenge, the model was validated on an external dataset consisting of 1318 patients. These were collected at three different hospitals in Belgium and Germany, whereas the training set was acquired in France. The model achieved an AUC in line with previous validations and performed considerably better than MUM3 inferred on the same data.

Third, the short-term severity prediction method developed in this study outperformed both risk models described by Revel et al. [17], the authors who contributed the STOIC dataset. Important to note, however, is that these were trained and evaluated on a different dataset split. Their best-performing method, achieving an AUC of 0.69, was a logistic regression including age, sex, oxygen supplementation at presentation, hypertension, and coronary artery disease as clinical risk factors and coronary artery calcium score and disease extent read from the CT.

Clinical applications and implications

For application in the clinic, the identification of patients that will develop severe COVID-19 is important to support therapeutic decisions. Recent guidelines base the administration of an antiviral drug, like remdesivir, and/or corticosteroids, like dexamethasone, possibly with adjunctive immunomodulators (baricitinib or tocilizumab,) on the required type of oxygen supplementation and risk factors for progressing to severe disease [52–54].

Another use of the predictive model is ICU planning, ensuring sufficient resources can be made available. At the same time, however, resources should not be retained needlessly as this may lead to postponing less urgent yet important care. Moreover, the risk of cross-infection

during patient transfer is high which leads to an increased risk for the health care practice. The system can help to select patients that can be discharged early from the hospital.

Prioritising high accuracy for detecting severe COVID-19, our model achieves a sensitivity of 0.94 with a specificity of 0.22 (95% CI: 0.19–0.24). For example, for the external validation set of 1318 patients, this corresponds to a correct classification of 405 patients that will develop severe disease and 191 that will not. However, 27 patients that will progress to severe COVID-19 are not identified and 695 that will not progress are wrongly classified. Latter is a relatively high number, indicating the limitations of the performance, but the provided error estimates could be taken into account during planning so that allocations can be adapted accordingly.

Performance across variants

Though the model was trained on data that was acquired at the early stages of the pandemic, when there were no globally dominant VOCs, the performance did not drop within the subset of data collected at later stages. In fact, more patients were correctly classified by both the proposed model and MUM3. However, for a fair comparison, the fraction of severe patients should be similar to that of the entire validation set. In the complete dataset, 32.8% of the patients developed severe COVID-19 while this was 7.50% for the delta subset and 17.1% the omicron group.

Within the group of patients admitted when the delta variant was most common, the three severe patients were correctly identified at baseline with a specificity of 0.81 (95% CI: 0.68–0.92) by the proposed model or 0.57 (95% CI: 0.42–0.73) by MUM3. For the patients admitted when the omicron variant was most prevalent, the six severe patients can be detected with a specificity of 0.31 (95% CI: 0.16–0.50) by the proposed model or 0.48 (95% CI: 0.31–0.66) by the MUM3 model. A limitation is that the variants were not confirmed through viral sequencing or genotype testing, and temporal data were used instead. Moreover, we should note the lower sizes of the test set for these evaluations, reflected in the larger confidence intervals. Nonetheless, these preliminary results indicate that predictive models based on measures closely related to lung severity remain valid for other variants.

Long-term prognosis prediction

Long-term prognosis prediction was assessed in the sense of the persistence of symptoms at three months after discharge. The selected predictors included image features, demographics and results from blood tests. The obtained performance was found to be poor, not surpassing 0.63 in terms of AUC, indicating these attributes hold little predictive value for long-term prognosis. As limitations, we should note the available dataset was relatively small and a high amount of missing data had to be imputed. Moreover, the time to the follow-up closest to three months varied considerably between patients. Further research with a larger and more complete dataset is required.

Conclusions

We developed and compared COVID-19 severity prediction models using features derived only from the CT scan and its metadata. The final selected model was based on a logistic regression using the patient age and sex and several imaging features. The method was developed on the publicly available data from the STOIC challenge [18] and independently validated on an external, multicenter dataset and demonstrated good generalisability, reaching 0.74 in terms of AUC. Though the model was developed on data that was acquired at the early stages of the pandemic, preliminary results indicate the performance remains comparable on data collected at later stages, in which other variants of concern were dominant. The short-term severity estimate relies on data from a single source, and could offer fast decision support for COVID-19 and allow for better planning and resource allocation in times of high prevalence.

Abbreviations

AUC	Area under the curve
AI	Artificial intelligence
AP	Average precision
CRP	C-reactive protein
CHUL	Centre Hospitalier Universitaire de Liège
CT	Computed tomography
CI	Confidence interval
CNN	Convolutional neural network
COPLE-Net	COVID-19 Pneumonia Lesion segmentation network
COVID-19	Coronavirus disease 2019
DNN	Deep neural network
DICOM	Digital Imaging and Communications in Medicine
GLCM	Gray level co-occurrence matrix
GLDM	Gray level dependence matrix
GLRLM	Gray level run length matrix
GLSZM	Gray level size zone matrix
GGO	Ground glass opacity
HU	Hounsfield units
ICU	Intensive care unit
LSTM	Long short-term memory
MDPI	Multidisciplinary Digital Publishing Institute
MUM3	Maastricht University Model 3
PR	Precision-recall
ROC	Receiver operating characteristic
RT-PCR	Reverse transcription polymerase chain reaction
SARI	Severe acute respiratory infections
STOIC	Study of Thoracic CT in COVID-19
SVM	Support vector machine
3D	Three-dimensional
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
UZB	Universitair Ziekenhuis Brussel
UKHD	Universitätsklinikum Heidelberg
VOC	Variant of concern

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02983-z.

Supplementary Material 1

Acknowledgements
The Authors thank the icovid consortium for the prior work that has made this study possible. The Authors also thank Els Van Nedervelde for the work on the SARI registry.

Author contributions
I.D., M.N.B, A.D.B, T.M., H.S., N.D. and J.V. conceptualised the study. E.V., B.I., S.V.E., L.S., A.W., G.D., J.G., A.G. and J.V. provided the resources to conduct the study and E.V., B.I., S.V.E., L.S., A.W., G.D., J.G. and A.G. performed the data curation. I.D. and J.V. performed the investigation, formal analysis and validation and prepared the original draft and visualizations. All authors reviewed the manuscript and approved the final version of the manuscript.

Funding
The Authors acknowledge financial support by "NUM 2.0" (FKZ:01KX2121) and from the following European Union's research and innovation programs. The DRAGON project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No. 101005122. The JU receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. The iCOVID project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101016131.

Data availability
The datasets generated and/or analysed during the current study are not publicly available due to ethical/privacy reasons. The short-term prognostic model is publicly available at: https://github.com/IneDirks/Covid_severity_prediction.git. The icolung software is offered, free of charge, at: https://icovid.ai/.

Declarations

Ethics approval and consent to participate
The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Universitair Ziekenhuis Brussel (Commissie Medische Ethiek O.G. 016, EC-2023-014), of Centre Hospitalier Universitaire de Liège (committee reference 707, study references 2020/139 and 2022/21) and of Universitätsklinikum Heidelberg (S-293/2020), who waived the need for informed consent for this retrospective study.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Pleinlaan, Brussels 1050, Belgium
[2]imec, Kapeldreef, Leuven 3001, Belgium
[3]Department of Radiology, Universitair Ziekenhuis Brussel, Vrije Universiteit Brussel (VUB), Laarbeeklaan, Jette 1090, Belgium
[4]Icometrix, Kolonel Begaultlaan, Leuven 3012, Belgium
[5]Department of Internal Medicine and Infectiology, Vitality Research Group (MIPI), Universitair Ziekenhuis Brussel, Vrije Universiteit Brussel (VUB), Laarbeeklaan, Jette 1090, Belgium
[6]Department of Infectious Diseases, University Hospital of Liège, Avenue de l'Hôpital, Liège 4000, Belgium
[7]Department of Pneumology, University Hospital of Liège, Avenue de l'Hôpital, Liège 4000, Belgium
[8]Department of Diagnostic and Interventional Radiology, University Hospital of Heidelberg, Im Neuenheimer Feld, 69120 Heidelberg, Germany
[9]Translational Lung Research Center (TLRC), German Center for Lung Research (DZL), University of Heidelberg, Im Neuenheimer Feld, 69120 Heidelberg, Germany
[10]Second Department of Radiology, University General Hospital Attikon, National and Kapodistrian University of Athens, Panepistimiou, Athens 157 72, Greece

Received: 13 February 2024 / Accepted: 20 March 2025
Published online: 01 April 2025

References

1. World Health Organization. WHO coronavirus dashboard. https://covid19.who.int/. 14 Aug 2023.
2. Hermans JJR, Groen J, Zwets E, Boxma-De Klerk BM, Van Werkhoven JM, Ong DSY, et al. Chest CT for triage during COVID-19 on the emergency department: myth or truth? Emerg Radiol. 2020;27(6):641–51. https://doi.org/10.1007/s10140-020-01821-1.
3. Desmet J, Biebaû C, De Wever W, Cockmartin L, Viktor V, Coolen J, et al. Performance of low-dose chest CT as a triage tool for suspected COVID-19 patients. J Belg Soc Radiol. 2021;105(1):1–8. https://doi.org/10.5334/jbsr.2319.
4. Esposito G, Ernst B, Henket M, Winandy M, Chatterjee A, Eyndhoven SV, et al. AI-based chest CT analysis for rapid COVID-19 diagnosis and prognosis: a practical tool to flag high-risk patients and lower healthcare costs. Diagnostics. 2022;12(7). https://doi.org/10.3390/diagnostics12071608.
5. Esposito G, Guiot J, Ernst B, Louis R, Meunier P, Kolh P. (Early) Economic Evaluation of the AI-based software 'icolung' for the detection and prognosis of COVID cases from CT scans. Eur Respir J. 2022;60(suppl 66).
6. Guiot J, Ernst B, Henket M, Louis R, Meunier P, Smeets D, et al. COVID-19 diagnosis and disease severity prediction assessment through an innovative AI-based model. Eur Respir J. 2022;60(suppl 66).
7. Cai W, Liu T, Xue X, Luo G, Wang X, Shen Y, et al. CT quantification and machine-learning models for assessment of disease severity and prognosis of COVID-19 patients. Acad Radiol. 2020;27(12):1665–78. https://doi.org/10.1016/j.acra.2020.09.004.
8. Meng L, Dong D, Li L, Niu M, Bai Y, Wang M, et al. A deep learning prognosis model help alert for COVID-19 patients at high-risk of death: a multi-center study. IEEE J Biomed Health Inform. 2020;24(12):3576–84. https://doi.org/10.1109/JBHI.2020.3034296.
9. Ning W, Lei S, Yang J, Cao Y, Jiang P, Yang Q, et al. Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. Nat Biomed Eng. 2020;4(12):1197–207. https://doi.org/10.1038/s41551-020-00633-5.
10. Fu L, Li Y, Cheng A, Pang PP, Shu Z, et al. A novel machine learning-derived radiomic signature of the whole lung differentiates stable from Progressive COVID-19 infection: a retrospective Cohort study. J Thoracic Imaging. 2020;35(6):361–68. https://doi.org/10.1097/RTI.0000000000000544.
11. Li C, Dong D, Li L, Gong W, Li X, Bai Y, et al. Classification of severe and critical covid-19 using deep learning and radiomics. IEEE J Biomed Health Inform. 2020;24:3585–94.
12. Yue H, Yu Q, Liu C, Huang Y, Jiang Z, Shao C, 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. 2020;8(14):859–859. https://doi.org/10.21037/atm-20-3026.
13. Wu G, Yang P, Xie Y, Woodruff HC, Rao X, Guiot J, et al. Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. Eur Respir J. 2020;56(2). https://doi.org/10.1183/13993003.01104-2020.
14. Fang C, Bai S, Chen Q, Zhou Y, Xia L, Qin L, et al. Deep learning for predicting COVID-19 malignant progression. Med Image Anal. 2021;72:102096.
15. Wang R, Jiao Z, Yang L, Choi JW, Xiong Z, Halsey K, et al. Artificial intelligence for prediction of COVID-19 progression using CT imaging and clinical data. Eur Radiol. 2022;32(1):205–12. https://doi.org/10.1007/s00330-021-08049-8.
16. Wang D, Huang C, Bao S, Fan T, Sun Z, Wang Y, et al. Study on the prognosis predictive model of COVID-19 patients based on CT radiomics. Sci Rep. 2021;11(1):1–9. https://doi.org/10.1038/s41598-021-90991-0.
17. Revel MP, Boussouar S, de Margerie-mellon C, Saab I, Lapotre T, Mompoint D, et al. Study of thoracic CT in COVID-19: the STOIC project. Radiology. 2021;301(1).
18. Grand Challenge.: STOIC2021. https://stoic2021.grand-challenge.org/stoic2021/. 14 Aug 2023.

19. Lassau N, Ammari S, Chouzenoux E, Gortais H, Herent P, Devilder M, et al. Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. Nat Commun. 2021;12(1):1–11. https://doi.org/10.1038/s41467-020-20657-4.
20. Shiri I, Salimi Y, Pakbin M, Hajianfar G, Avval AH, Sanaat A, et al. COVID-19 prognostic modeling using CT radiomic features and machine learning algorithms: analysis of a multi-institutional dataset of 14,339 patients: COVID-19 prognostic modeling using CT radiomics and machine learning. Comput Biol Med. 2022;145(February):105467. https://doi.org/10.1016/j.compbiomed.2022.105467.
21. Kienzle D, Lorenz J, Schön R, Ludwig K, Lienhart R. COVID detection and severity prediction with 3D-ConvNeXt and custom pretrainings. In: Computer Vision – ECCV 2022 Workshops; 2022.
22. Shakouri S, Bakhshali MA, Layegh P, Kiani B, Masoumi F, Ataei Nakhaei S, et al. COVID19-CT-dataset: an open-access chest CT image repository of 1000+ patients with confirmed COVID-19 diagnosis. BMC Res Notes. 2021;14(1):178. https://doi.org/10.1186/s13104-021-05592-x.
23. Duan L, Zhang L, Lu G, Guo L, Duan S, Zhou C. A CT-based radiomics model for prediction of prognosis in patients with novel coronavirus disease (COVID-19) pneumonia: a preliminary study. MDPI Diagn. 2023;13(8).
24. Collins GS, Reitsma JB, Altman DG, Moons KGM. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 162(1):55–63. https://doi.org/10.7326/M14-0697.
25. Boulogne LH, Lorenz J, Kienzle D, Schon R, Ludwig K, Lienhart R, et al. The STOIC2021 COVID-19 AI challenge: applying reusable training methodologies to private data. ArXiv: 2306.10484.
26. Darcis G, Bouquegneau A, Maes N, Thys M, Henket M, Labye F, et al. Long-term clinical follow-up of patients suffering from moderate-to-severe COVID-19 infection: a monocentric prospective observational cohort study. Int J Infect Dis. 2021;109:209–16.
27. Seyler L, Van Nedervelde E, De Cock D, Mann C, Pien K, Allard SD, et al. Surfing the waves: differences in hospitalised COVID-19 patients across 4 variant waves in a Belgian University Hospital. Viruses. 2023;15(3):618. https://doi.org/10.3390/v15030618.
28. Sciensano. Belgium COVID-19 epidemiological situation - variants. https://lookerstudio.google.com/embed/u/0/reporting/c14a5cfc-cab7-4812-848c-0369173148ab/page/urrUC. 14 Aug 2023.
29. JoHof. Automated lung segmentation in CT under presence of severe pathologies. https://github.com/JoHof/lungmask. 14 Aug 2023.
30. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18:203–11. https://doi.org/10.1038/s41592-020-01008-z.
31. Project MONAI. Monai. https://monai.io. 24 May 2023.
32. Grand Challenge. COVID-19 lung CT lesion segmentation challenge. https://covid-segmentation.grand-challenge.org/
33. icometrix. icovid. https://icovid.ai/. 14 Aug 2023.
34. Wang G, Liu X, Li C, Xu Z, Ruan J, Zhu H, Meng T, Li K, Huang N, Zhang S. COPLE-Net: COVID-19 pneumonia lesion segmentation network. https://github.com/HiLab-git/COPLE-Net. 14 Aug 2023
35. Jun M, Cheng G, Yixin W, Xingle A, Jiantao G, Ziqi Y, et al. COVID-19 CT lung and infection segmentation dataset. Zenodo. 2020. https://doi.org/10.5281/zenodo.3757476.
36. Tarvainen A, Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: 31st Conference on Neural Information Processing Systems. 2017. p. 1195–204.
37. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. ArXiv: 2201.03545.
38. Chen HJ, Mao L, Chen Y, Yuan L, Wang F, Li X, et al. Machine learning-based CT radiomics model distinguishes COVID-19 from non-COVID-19 pneumonia. BMC Infect Dis. 2021;21(1):1–13. https://doi.org/10.1186/s12879-021-06614-6.
39. Huang Y, Zhang Z, Liu S, Li X, Yang Y, Ma J, et al. CT-based radiomics combined with signs: a valuable tool to help radiologist discriminate COVID-19 and influenza pneumonia. BMC Med Imaging. 2021;21(1):1–12. https://doi.org/10.1186/s12880-021-00564-w.
40. DeLong E, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–45.
41. Sun X, Xu W, et al. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett. 2014;21(11):1389–93. https://doi.org/10.1109/LSP.2014.2337313.
42. Maastricht University and Comunicare. MyPatientCheck. https://mypatientcheck.web.app/home. 14 Aug 2023.
43. Shwartz-Ziv R, Armon A, et al. Tabular data: deep learning is not all you need. Inf Fusion. 2022;81(June 2021):84–90. https://doi.org/10.1016/j.inffus.2021.11.011.
44. Boulogne LH, Lorenz J, Kienzle D, Schön R, Ludwig K, Lienhart R, et al. The STOIC2021 COVID-19 AI challenge: applying reusable training methodologies to private data. Med Image Anal. 2024;97:103230.
45. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 3149–57.
46. Chen T, Guestrin C. XGBoost: a Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'16. New York, NY, USA: Association for Computing Machinery; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
47. Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G. Deep neural networks and tabular data: a survey. In: IEEE Transactions on Neural Networks and Learning Systems. 2022. p. 1–21. https://doi.org/10.1109/TNNLS.2022.3229161.
48. Iba T, Connors JM, Levy JH, et al. The coagulopathy, endotheliopathy, and vasculitis of COVID-19. Inflammation Res. 2020;69(12):1181–89. https://doi.org/10.1007/s00011-020-01401-6.
49. Wong K, Farooq Alam Shah MU, Khurshid M, Ullah I, Tahir MJ, Yousaf Z. COVID-19 associated vasculitis: a systematic review of case reports and case series. Ann Med Surg. 2012;74:103249. https://doi.org/10.1016/j.amsu.2022.103249.
50. Delli Pizzi A, Chiarelli AM, Chiacchiaretta P, Valdesi C, Croce P, Mastrodicasa D, et al. Radiomics-based machine learning differentiates "ground-glass" opacities due to COVID-19 from acute non-COVID-19 lung disease. Sci Rep. 2021;11(1):17237.https://doi.org/10.1038/s41598-021-96755-0.
51. Shiri I, Mostafaei S, Haddadi Avval A, Salimi Y, Sanaat A, Akhavanallaf A, et al. High-dimensional multinomial multiclass severity scoring of COVID-19 pneumonia using CT radiomics features and machine learning algorithms. Sci Rep. 2022;12(1):14817. https://doi.org/10.1038/s41598-022-18994-z.
52. Kim AY, Gandhi RT. Wolters Kluwer UpToDate. COVID-19: management in hospitalized adults. https://www.uptodate.com/contents/covid-19-management-in-hospitalized-adults?sectionName=COVID-19-SPECIFICTHERAPYtopicRef=127454anchor=H3855514466source=see_link#. 14 Jul 2023.
53. National Institutes of Health. COVID-19 treatment guidelines panel. Coronavirus disease 2019 (COVID-19) treatment guidelines. https://files.covid19treatmentguidelines.nih.gov/guidelines/covid19treatmentguidelines.pdf. 14 Jul 2023.
54. Cascella M, Rajnik M, Aleem A, Dulebohn SC, Napoli RD. Features, evaluation, and treatment of coronavirus (COVID-19). https://www.statpearls.com/ArticleLibrary/viewarticle/52171. 14 Jul 2023.

Publisher's Note