iCTCF: an integrative resource of chest computed tomography images and clinical features of patients with COVID-19 pneumonia

Wanshan Ning1,†, Shijun Lei2,3,†, Jingjing Yang4,†, Yukun Cao5,†, Peiran Jiang1, Qianqian Yang2, Jiao Zhang2, Xiaobei Wang2, Fenghua Chen2, Zhi Geng2, Liang Xiong6, Hongmei Zhou7, Yaping Guo1, Yulan Zeng4,*, Heshui Shi5,*, Lin Wang2,3,*, Yu Xue1,*, Zheng Wang3,8*

1Key Laboratory of Molecular Biophysics of Ministry of Education, Hubei Bioinformatics and Molecular Imaging Key Laboratory, Center for Artificial Intelligence Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China
2Department of Clinical Laboratory, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China
3Research Center for Tissue Engineering and Regenerative Medicine, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China
4Department of Respiratory and Critical Care Medicine, Liyuan Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430077, China
5Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China, Hubei Province Key Laboratory of Molecular Imaging, Wuhan 430022, China
6Department of Laboratory Medicine, Liyuan Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430077, China
7Department of Radiology, Liyuan Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430077, China
Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China

Running title: A data resource of COVID-19 pneumonia

†These authors contributed equally to this work.
*To whom correspondence should be addressed.
Zheng Wang, Tel: +86-27-85726612, E-mail: zhengwang@hust.edu.cn.
Yu Xue, Tel: +86-27-87793903, Fax: +86-27-87793172, E-mail: xueyu@hust.edu.cn.
Lin Wang, Tel: +86-27-85726612, E-mail: lin_wang@hust.edu.cn.
Heshui Shi, Tel: +86-27-85726730, Fax: +86-27-85727001, E-mail: heshuishi@hust.edu.cn.
Yulan Zeng, Tel: +86-27-86773985, Fax: +86-27-86793060, E-mail: 1989ly0551@hust.edu.cn.
Abstract

The outbreak of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was initially reported in Wuhan, China since December, 2019. Here, we reported a timely and comprehensive resource named iCTCF to archive 256,356 chest computed tomography (CT) images, 127 types of clinical features (CFs), and laboratory-confirmed SARS-CoV-2 clinical status from 1170 patients, reaching a data volume of 38.2 GB. To facilitate COVID-19 diagnosis, we integrated the heterogeneous CT and CF datasets, and developed a novel framework of Hybrid-learning for UnbiaSed predicTion of COVID-19 patients (HUST-19) to predict negative cases, mild/regular and severe/critically ill patients, respectively. Although both CT images and CFs are informative in predicting patients with or without COVID-19 pneumonia, the integration of CT and CF datasets achieved a striking accuracy with an area under the curve (AUC) value of 0.978, much higher than that when exclusively using either CT (0.919) or CF data (0.882). Together with HUST-19, iCTCF can serve as a fundamental resource for improving the diagnosis and management of COVID-19 patients.

Keywords

COVID-19 pneumonia; SARS-CoV-2; computed tomography; clinical feature; ground-glass opacity; hybrid-learning
Introduction

Since December, 2019, the outbreak of an unknown viral pneumonia has severely affected Wuhan, China. This virus was quickly identified and named by the World Health Organization (WHO) as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)\textsuperscript{1-3}, and the resulting viral pneumonia was referred to as coronavirus disease 2019 (COVID-19) pneumonia\textsuperscript{4-8}. By the end of March, 2020, nearly 200 countries and regions were affected with > 500,000 confirmed cases, which are still increasing. Such a severe situation underscores the urgency for developing effective measures to control this pandemic.

To control the outbreak, early diagnosis of patients with COVID-19 pneumonia for timely treatment is critical, especially in epidemic regions\textsuperscript{2,8-11}. However, it was challenging to be achieved. Across many COVID-19-stricken regions, it appears to be common that limited medical resources and a large number of quickly piled-up patients result in long waiting time for diagnosis and medical decisions, such as quarantine or hospitalization, which potentially increases the chance of cross-infection and leads to poor prognosis. Although COVID-19 confirmation relies on real-time polymerase chain reaction (RT-PCR) detecting existence of SARS-CoV-2, this PCR testing was found to have a high specificity ($Sp$)\textsuperscript{8} but a rather low sensitivity ($Sn$) with a reported positive rate of merely 38%-57%\textsuperscript{12}.

In addition to etiological laboratory confirmation, other key diagnostic elements that could facilitate identification of COVID-19 pneumonia include clinical features (CFs) and chest computed tomography (CT) imaging\textsuperscript{13,14}. Consistent with the importance of these elements, COVID-19 guidelines, for instance, the one followed in China, uses all these elements to define mild, regular, severe, and critically ill forms of COVID-19 pneumonia\textsuperscript{9,14-18}. Despite far from complete understanding, the limited number of relevant papers have begun to reveal relevant CFs, including symptoms of COVID-19 such as fever, dry cough, myalgia and shortness of breath\textsuperscript{10,19,20}. Other CFs, such as lymphopenia, elevated levels of inflammatory cytokines, and reduction in T cell subsets, are also frequently found\textsuperscript{10,11,15}. Moreover, chest CT imaging characteristics of infected lungs reportedly include ground glass opacity (GGO) and severity-correlated
Although the full picture regarding each element still awaits full depiction, comprehensively pooling the features of the aforementioned diagnostic elements might collectively improve diagnosis accuracy and efficacy.

Amid the ongoing COVID-19 pandemic, the availability of first-hand radiographic and clinical datasets would be essential and important to help in guiding clinical decision-making, to supply information for deepening the understanding of this viral infection, and to provide a basis for systemic modeling that may facilitate early diagnosis for timely medical intervention. A way to achieve this goal is to build an open-access and comprehensive resource containing chest CT images and CFs of individual patients, to offer a platform allowing internationally joint efforts to combat COVID-19 pneumonia.

From the accumulated data in our hospitals, we enrolled 1170 anonymous patients including 649 laboratory-confirmed, 222 COVID-19-negative/control and 299 suspected patients, and collected their corresponding chest CT images, CFs and SARS-CoV-2 laboratory testing results if available. Then, we developed a patient-centric resource named integrative CT images and CFs for COVID-19 (iCTCF) to archive and share the rich data. Through integration of the highly heterogeneous CT and CF datasets, we built a novel framework of Hybrid-learning for Unbiased Prediction of COVID-19 patients (HUST-19) to predict negative cases (Control), mild/regular (Type I) and severe/critically ill (Type II) patients, respectively. The 10-fold cross-validation was conducted for model training, parameter optimization and performance evaluation. From the results, we found that both CT images and CFs are informative in predicting patients with or without COVID-19 pneumonia, whereas CFs are considerably better in recognizing Type II patients. The integration of CT and CF datasets in HUST-19 generated a higher accuracy with area under the curve (AUC) values of 0.978, 0.921 and 0.931 for predicting control, Type I and Type II patients, respectively. Using HUST-19, we conducted a retrospective analysis of 299 suspected cases, and predicted 207 and 71 cases to be potential Type I and II patients, respectively. In conclusion, this medical resource can be a fundamental resource not only for the purpose of
retrospective analysis, but also provide a useful tool for improving diagnosis and treatment of COVID-19 patients. Together with HUST-19, iCTCF is freely available for academic research at http://ictcf.biocuckoo.cn/.
Results

A summary of collected data sets

We enrolled 1170 patients admitted from January 25 to February 20, 2020, including 775 from Union Hospital (HUST-UH) and 395 from Liyuan hospital (HUST-LH). Among these 1170 patients, 978 patients had both CT and CF data, and 192 patients had only CF data. The 1170 patients include (i) 649 COVID-19-confirmed patients with pneumonia severity ranging from mild (23 cases, 2.0%), regular (417 cases, 35.6%), severe (146 cases, 12.5%), to critically ill form (63 cases, 5.4%), (ii) 222 COVID-19-negative cases (regarded as the control), and (iii) 299 COVID-19- suspected cases (Fig. 1a). Due to limited data regarding mild and critically ill forms, here we defined mild and regular forms as Type I, and severe and critically ill forms as Type II.

In total, 256,356 CT images were collected from 978 patients who had chest CT data, including 53,993 images (25.9%) from 208 control cases, 3879 images (1.5%) from 19 patients with mild form, 85,924 images (33.5%) from 353 patients with regular form, 35,352 images (13.8%) from 110 patients with severe form, 7017 images (2.7%) from 29 patients with critically ill form, and 70,191 images from 259 patients with suspected COVID-19 (Fig. 1b). For each form of COVID-19 pneumonia, one representative chest CT image was shown (Fig. 1c). In contrast to the normal lung CT image for the control group showing no pneumonia signs in bilateral lungs, the major CT image characteristics of mild form of COVID-19 was GGO in limited regions of lungs. Chest CT images of regular form of COVID-19 were often featured with multiple lesions, heterogeneous GGO, consolidation and linear shadows, while severe form of COVID-19 could lead to extensive GGO and linear consolidation in bilateral lung parenchyma. Moreover, critically ill form of COVID-19 often presented radiographic signs of extensive consolidation, internal pulmonary septal thickening, and air branchogram. Suspected patients could have one or more radiographic features similar to any form of COVID-19.
In our data set, there were 1170 patients with CF data, including 222 control, 23 mild, 417 regular, 146 severe, 63 critically ill and 299 suspected cases. The CF data contained 127 types from 9 categories, including basic information, routine blood test, inflammation test, blood coagulation test, biochemical test, immune cell typing, cytokine profile test, autoimmune test, and routine urine test (Supplementary Table 1). The information regarding underlying diseases (Udis) of patients from the two hospitals were also present (Supplementary Table 2). Compared to the controls, 5 types of CFs in COVID-19 patients were statistically higher, including body temperature (average 38.1°C vs. 36.8°C in controls, \( p \)-value = 2.67E-39, two-sided unpaired t-test), globulin (GLB, \( p \)-value = 1.78E-05, two-sided unpaired t-test), lactate dehydrogenase (LDH, \( p \)-value = 2.11E-05, two-sided unpaired t-test), neutrophil percent (NEP, \( p \)-value = 2.18E-05, two-sided unpaired t-test), and age (average 55.1 vs. 49.3 in controls, \( p \)-value = 3.01E-05, two-sided unpaired t-test) (Fig. 1d, Supplementary Table 3). In COVID-19 patients, 11 CFs significantly decreased, such as eosinophil counts (EO, \( p \)-value = 2.22E-17, two-sided unpaired t-test), eosinophil percent (EOP, \( p \)-value = 3.99E-15, two-sided unpaired t-test), and calcium (CA, \( p \)-value = 2.75E-12, two-sided unpaired t-test) (Fig. 1d, Supplementary Table 3). We also compared Type I and II patients. Compared to Type I patients, 19 and 11 CFs were dramatically increased and decreased in Type II patients, respectively (Fig. 1d, Supplementary Table 3). Thus, the number of CFs different between Type II and Type I patients are more than that between COVID-19 and control cases, indicating that CF data might be informative for classification of Type II and I patients.

Compared to Type I patients, two CFs, body temperature (average 38.4°C vs. 37.9°C, \( p \)-value = 1.23E-07, two-sided unpaired t-test) and age (average 62.8 vs. 50.8, \( p \)-value = 7.98E-17, two-sided unpaired t-test), were significantly higher in Type II patients (Fig. 1d, Supplementary Table 3), suggesting an association between elder age and illness severity, consistent with recent reports\textsuperscript{22,23}. The proportions of patients with Udis were similar in the controls and in COVID-19 patients (Type I and Type II) (Supplementary Table 3), suggesting a general susceptibility. Of note, Type II cases had a higher
proportion of patients with Udis than Type I cases (Two-sided chi-squared test, $p$-value = 2.73E-09, Supplementary Table 3), revealing a link of Udis with COVID-19 disease severity, consistent with a previous study\textsuperscript{24}.

**An integrative resource of COVID-19 pneumonia**

For 1170 patients with or without COVID-19 pneumonia, we developed an easy-to-use resource named iCTCF to maintain and share their corresponding CT images, CFs, and SARS-CoV-2 laboratory-testing results. On the resource page (http://ictcf.biocuckoo.cn/Resource.php), multiple select options could be chosen for database searching, including hospital (HUST-UH, and/or HUST-LH), age (< 40 years, 40-60 years, and/or > 60 years), gender (female or male), SARS-CoV-2 laboratory-confirmed status (positive and/or negative), CT evidence (positive, negative, and/or N/A), and forms of COVID-19 (critically ill, severe, regular, mild, suspected, and/or control) (Fig. 2a). Desired selections could be customized. By clicking the button “Submit”, the results could be displayed in a tabular list with 20 patients’ information per page (Fig. 2a).

For convenience, we also provided an “example” button that could be clicked to automatically load pre-configured selections, followed by the representation of several typical cases (Fig. 2b). Here, we selected “Patient 4” as an example to show the annotations in iCTCF. Patient 4 had intermittent fever (highest as 38.5°C), fatigue, shortness of breath, and myalgia ten days prior to admission. He coughed occasionally with sputum. On Feb 3, the RT-PCR for SAS-CoV-2 nucleic acid test on his throat swab specimens was positive. He was admitted on Feb 4 with finger blood oxygen saturation of 90% in ambient air. It reached 98% through face mask oxygen support (3 L/min). According to the Guideline of China (6\textsuperscript{th} edition), he was diagnosed as severe form of COVID-19 and regarded as a Type II case in iCTCF.

By clicking “Patient 4”, detailed information on the anonymous patient would be shown (Fig. 2c). On patient page, a brief summary of the patient was presented, whereas 5
representative CT images would be displayed (Fig. 2c, d). All numerical forms of CFs were provided in a tabular list, and the laboratory-confirmed SARS-CoV-2 status was also shown (Fig. 2c, e). Consistent with his brief clinical summary described above, users would find the age of this patient (73), gender (male), his body temperature (38.5℃), the positive SARS-CoV-2 infection status, and his Udis (aorta calcification). His CT examination at clinics suggested possible bilateral viral pneumonia. His tabular list contained 82 numerical CFs, including but not limited to the decrease of EC, lymphocyte count (LY), EOP and lymphocyte percent (LYP), and the increased levels of NEP, erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP) (Fig. 2e). He was diagnosed as the severe form of COVID-19 (Fig. 2c).

A hybrid-learning framework to predict COVID-19 patients

To exemplify the usefulness of iCTCF, we developed a computational method named HUST-19 that integrated the data sets of CT images and CFs for prediction of patients with (positive) or without COVID-19 pneumonia (control). The full procedure of HUST-19 contained four steps, including the classification of individual CT images, the CT image-based prediction of patients, the CF-based prediction of patients, and the integration of CT- and CF-based predictions to classify patients into three types: control, Type I and Type II (Fig. 3).

First, we classified individual CT images into three types, (i) non-informative CT (NiCT) images where lung parenchyma was not captured for any judgement, (ii) positive CT (pCT) images where imaging features associated with COVID-19 pneumonia could be unambiguously discerned, and (iii) negative CT (nCT) images where imaging features in both lungs were irrelevant to COVID-19 pneumonia. To enable the image-based prediction, we adopted a deep learning framework of 13-layer convolutional neural networks (CNNs), containing one input layer, 3 sets of dual convolutional and pooling layers (3 × 3), 2 fully connected (dense) layers, and one output layer (Fig. 3). This 13-layer CNNs were utilized to classify individual CT images into three types: NiCT, pCT, and nCT. Second, an additional framework of 13-layer
CNNs was implemented to transform the individual CT image-based prediction into the patient-based prediction (Fig. 3). For each patient, ten mostly probable pCT images were reserved as the representative images, which were inputted into the secondary 13-layer CNNs to classify the patient as a control, Type I or Type II case. Third, the CF-based classification of patients into three types was implemented in a framework of 7-layer deep neural networks (DNNs), including one input layer, 5 dense layers, and one output layer (Fig. 3). In contrast to CNNs, DNNs did not have convolutional and pooling layers. Finally, the predictions using CT images or CFs were integrated through the penalized logistic regression (PLR) algorithm to output final predictions on patient classification (control, Type I and Type II).

The prediction accuracy of HUST-19

For training individual image-based models in HUST-19, we manually labelled 19,685 CT images (5705 NiCT, 4001 pCT and 9979 nCT images) randomly selected from 61 COVID-19 patients and 43 control cases. Each CT image was confirmed by ≥3 researchers, including at least one clinician from HUST-UH and one clinician from HUST-LH. The 10-fold cross-validations were conducted for model training, parameter optimization and performance evaluation. For the image-based prediction, HUST-19 achieved an AUC value of 0.994 in distinguishing pCT and nCT images from NiCT images, and an AUC value of 0.991 in predicting pCT images (Fig. 4a).

For training patient-centered models, we used 186,165 CT images and 127 types of CFs data from 222 controls, 440 Type I patients, and 209 Type II patients. Again, the 10-fold cross-validations were conducted for model training, parameter optimization and performance evaluation. For the CT image-based prediction, HUST-19 recognized COVID-19 patients with an AUC value of 0.919, whereas the AUC values for predicting Type I and II patients were 0.804 and 0.838, respectively (Fig. 4b). Compared to the CT image-based prediction, the CF-based prediction had a lower accuracy in general for identifying COVID-19 patients, as evidenced by an AUC value of 0.882 (Fig. 4c). However, CFs were more effective than CT images in classifying
Type I and II COVID-19 patients (Fig. 4c). Thus, CT images and CFs had their unique advantages, providing a justification for integration of these two types of datasets. Indeed, their combination produced much higher AUC values of 0.978, 0.921 and 0.931 in predicting controls, Type I and II patients, respectively (Fig. 4d).

For a general prediction, the type of a CT image or a patient was determined based on the highest probability score of three output values. Under this threshold, confusion matrices were generated from the 10-fold cross-validations to visualize the average agreement between actual and predicted results (Fig. 4e-h). It was found that NiCT, pCT and nCT images can be correctly recognized with a high accuracy (Fig. 4e). For patient-based predictions, CT image-based and CF-based models achieve a similar performance in recognizing control cases and Type I patients, whereas CF-based prediction exhibited a higher efficiency in correctly predicting Type II patients (Fig. 4f, g). The integration of CT images and CFs considerably improve the prediction efficiency on recognizing control cases and Type II patients (Fig. 4h). The results of other performance measurements under the general threshold were shown in Supplementary Table 4.

In HUST-19, the best model with the highest 10-fold cross-validation AUC value was reserved for each of the four types of predictions. To maximize the ability on correctly predicting true COVID-19 patients, we chose a sensitive threshold and summarized the corresponding confusion matrices on individual CT image-based prediction (Fig. 4i) as well as CT image-based (Fig. 4j), CF-based (Fig. 4k) and HUST-19-based predictions (Fig. 4l), respectively. Additional performance values under this threshold were also shown in Supplementary Table 4.

**Computational annotations of suspected cases**

In iCTCF, there were 299 suspected cases without definitive SARS-CoV-2 laboratory confirmation at the time of enrollment. Here, we used HUST-19 with the sensitive threshold to predicted 21,207 and 71 patients of 299 suspected cases to be COVID-19
negative cases, Type I cases, and Type II cases, respectively (Fig. 5a). For each patient, the six intermediate scores generated from CT image-based and CF-based prediction of patients were retrieved and analyzed by the t-distributed stochastic neighbor embedding (t-SNE) in a 2-dimensional (2D) plot. The t-SNE result demonstrated the suspected cases were dispersed in the three types, and the predicted Type I and II cases were highly approximal to COVID-19 confirmed cases (Fig. 5b). For example, Patients 324 and 610 were predicted as Type I and II cases, respectively (Fig. 5b).

Patient 324, female, 34-year old, was admitted to the HUST-UH on Feb 9, because of "fever for seven days" and ground glass lesion in left lower lung suggested by CT imaging (Fig. 5c). Her highest body temperature was 38.3°C, accompanied by mild cough, diarrhea and fatigue. On Feb 10 (Day 2 after admission), the blood biochemical examination showed eosinopenia, lymphopenia, elevated high-sensitivity C-reactive protein (HSCRP), and increased interleukin-6 (IL-6). However, the results of three SARS-CoV-2 nucleic acid tests on Feb 3 (at clinics), 11 (Day 2 after admission) and 13 (Day 4 after admission) were all negative. Given that she was afebrile since Feb 12 (Day 3 after admission) and CT indicated a significant improvement in lung lesions on Feb 17 (Day 8 after admission), this patient was discharged with the diagnosis of “suspected COVID-19 regular form” on Feb 20 (Day 11 after admission). Based on her CT and CF data, our HUST-19 model predicted this patient to be Type I COVID-19. Her COVID-19 infection was definitively verified by her positive test for COVID-19 serum IgM and IgG when she came back to the hospital for follow-up examination on Mar 6 (Day 15 after discharged). Thus, her diagnosis was eventually corrected to be “COVID-19 regular form” (Fig. 5c).

Another example was from Patient 610, female, 73-year old (Fig. 5c). She was admitted to the HUST-UH on Feb 6, because of "fever, headache, and dizziness for a week" and multiple GGO lesions in her both lungs’ CT images. Her body temperature was 38.4°C, accompanied by shortness of breath, coughing and phlegm. She had hypertension that was under control. Upon admission, her laboratory tests showed eosinopenia, an
elevated level of D-Dimer (DD), and an increased ESR. During hospitalization, the patient had dyspnea with finger blood oxygen saturation < 93% in ambient condition, which was corrected by oxygen support (10 min/L, 86%) through face mask. However, the results of four SARS-CoV-2 nucleic acid tests on Feb 5 (Day 1 before admission), Feb 7 (Day 1 after admission), Feb 23 (Day 17 after admission), and Feb 25 (Day 19 after admission) were all negative. Her COVID-19 was eventually confirmed by the positive result for serum SARS-CoV-2 IgM and IgG tests on Mar 2 (Day 25 after admission). Thus, her diagnosis was severe form of COVID-19 (Fig. 5c). Given her CT and CF data, the HUST-19 also accurately classified this case to be Type II COVID-19 even without knowing her SARS-CoV-2 infection status (Fig. 5b). Taken together, our annotation on suspected cases suggests that HUST-19 can be a helpful tool for identifying true COVID-19 patients, and rapidly provide useful information towards further diagnosis and treatment of the disease.
**Discussion**

During the last few months, the outbreak of COVID-19 pneumonia has affected nearly 200 countries and regions, and endangered millions of lives. Since the early stage of this outbreak, our local hospitals at the COVID-19 epicenter had been accepting COVID-19 patients, and accumulated considerable amount of first-hand COVID-19 related radiographic and clinical data. We believe that timely and consistent curation of CT imaging, clinical and laboratory information of COVID-19 is important to help understand scope of disease impact, spectrum of pathophysiological features, and risk factors for progressive worsening, and to provide the reliable resource for diagnostic modelling and retrospective analysis that can inform screening, triage, diagnosis, facilitate risk management, and treatment efforts to effectively combat this disease. Further, detailed information provided at large scale in real time can be critical for revealing robust clinical findings, deciding where to prioritize research or therapeutic efforts, and building consensus for clinicians.

Large datasets that are freely available for sharing during outbreaks are rare. Several published papers and some preprints have reported the use of cutting-edge artificial intelligence (AI) in predicting COVID-19 patients based on CT imaging and relevant clinical information\textsuperscript{15,25-33}. Although these studies highlighted the potential value of machine learning in COVID-19 diagnosis, their datasets unitized to build their algorithms were relatively small with no free public access. It would be difficult for other researchers to utilize those data for further analysis. In contrast, our iCTCF database is an openly-accessible computational infrastructure for public. It systemically integrated CT images and CFs from patients with or without COVID-19 pneumonia. To the best of our knowledge, iCTCF is not only a comprehensive repository for COVID-19 pneumonia, but also the largest CT imaging and CF characteristics database of COVID-19 thus far.

Given the rapid surge in COVID-19 infected cases globally, the number of patients has been overloading local medical systems in many regions. A key to control this epidemic is to diagnose COVID-19 as early as possible for applying timely medical interventions,
such as isolation or treatment, with a goal of reducing cross-infection, and blocking illness progression in individual patients. To facilitate COVID-19 diagnosis and also demonstrate the usefulness of iCTCF, we developed a novel framework of HUST-19 by capitalizing on CT and CFs data, and achieved an effective identification of true COVID-19 patients with a striking accuracy with an AUC value of 0.978, much higher than that when using either CT data alone (0.919) or CFs data only (0.882).

As the pandemic is ongoing, we hope to continuously archive relevant COVID-19 information into the iCTCF system for broad sharing. With abundant and reliable information, the iCTCF database can be a valuable platform to converge global intelligence towards further improving diagnosis and clinical management, and identifying effective cures for COVID-19 during this outbreak.
Methods

Data collection and preparation

The collection, use, and retrospective analysis of CT images, CFs and SARS-CoV-2 nucleic acid PCR results of patients were approved by the institutional ethical committees of HUST-UH and HUST-LH. The data were from (i) patients receiving PCR nucleic acid testing and hospitalized between Jan 25 and Feb 20, 2020 in HUST-UH and HUST-LH; and (ii) 112 patients admitted to HUST-UH between Nov 14 and Nov 30, 2019, and diagnosed with community-acquired pneumonia; (iii) 14 healthy cases. For each patient, her/his medical records were manually checked and confirmed by ≥ 3 researchers, which included at least one clinician from HUST-UH and one clinician from HUST-LH. In total, 1170 patients were included, 775 patients from HUST-UH and 395 patients from HUST-LH (Fig. 1a). There were 222 control cases consisting of 112 patients with community-acquired pneumonia, 14 healthy cases, and 96 patients whose SARS-CoV-2 nucleic acid testing was negative and CT showed no signs of COVID-19 infection. 649 laboratory-confirmed COVID-19 patients was composed of 23 mild, 417 regular, 146 severe and 63 critically ill. The remaining 299 patients were the suspected cases (Case definitions described below). Among these 1170 patients, 978 patients had CT data that contained 256,356 CT images. This study was approved by both HUST-UH (IRB ID: [2020] IEC (A001)) and HUST-LH (IRB ID: [2020] IEC (A001)).

COVID-19 case definitions

The Guidelines of the National Health Commission of China for COVID-19 (6th edition) described suspected COVID-19 cases would have the following features (i) recent travel history to Wuhan City or Hubei Province or close contact with a confirmed or probable case; (ii) fever and/or respiratory symptoms; (iii) laboratory findings of normal or decreased number of leukocytes and/or lymphopenia; (iv) radiographic CT evidence showing signs of pneumonia. The suspected cases were defined if they met the following criteria: i plus any two of ii, iii and iv, or three of ii, iii and iv at the same time. The confirmed cases need etiological evidence showing SARS-CoV-2 positive
for specimens (such as throat swab, nasopharyngeal swab, sputum, peripheral blood, lower respiratory tract, or alveolar lavage fluid) by RT-PCR. Thus, suspected cases would have COVID-19-associated CT features but without definitive SARS-CoV-2 RT-PCR positive results; confirmed cases would have COVID-19 radiographic evidence and SARS-CoV-2 RT-PCR positive evidence.

The procedure of clinical examinations

Nine classes of clinical examinations that patients received included basic information, routine blood tests, inflammation tests, blood coagulation tests, biochemical tests, immune cell typing, cytokine profile tests, autoimmune tests, and routine urine tests. These tests were performed in the departments of clinical laboratory in HUST-UH and HUST-LH, and the results were collectively as CFs (Supplementary Table 1). The basic information included age, gender, body temperature (°C), and Udis, which were derived from patients’ medical records.

In HUTS-UH, routine blood tests, such as hemoglobin (HGB), were carried out by a Sysmex XE-5000 automatic blood analyzer (Sysmex Corporation, Kobe, Japan). For inflammation, ESR was detected by a Monitor 100 (Vital Diagnostics, Italy), and CRP was tested by a BN II (Siemens, Germany). Blood coagulation tests were carried out by a STA-R Evolution (Stago, France). Biochemical tests were done by an AU5800 (Beckman Coulter, USA). B-type brain natriuretic peptide precursor (BNP) was detected by an Architect i2000 (Abbott, USA). Quantification and typing of immune cells were conducted using a flow cytometry (Cyromics FC 500, Beckman Coulter, USA). The cytokine profile tests, such as detection of interleukin-2, 4, 6, 8, 10, TNF-α, and IFN-γ, were also quantitatively determined by a Cyromics FC 500 (Beckman Coulter, USA). Autoimmune tests of complements (C1q, C3 and C4) and immunoglobulins (A, M and G) were conducted in an IMMAGE 800 (Beckman Coulter, USA), while Anti-Streptolysin O (ASO) and rheumatoid factor (RF) were tested by BN II (Siemens, Germany). Routine urine tests were performed using a Sysmex UF-1000i (Sysmex Corporation, Kobe, Japan).
In HUST-LH, routine blood tests were performed by a BC5390 automatic blood cell analyzer (Mindray, China). For inflammation, CRP were also tested by a BC5390 automatic blood cell analyzer (Mindray, China). Procalcitonin (PCT) was tested by a Pylon immunoassay (ET Healthcare, China). HSCRP were tested by an AU5800 (Beckman Coulter, USA). Blood coagulation tests were carried out by a Sysmex CS5100 automatic blood coagulation analyzer (Sysmex Corporation, Kobe, Japan). For biochemical tests and autoimmune tests, an AU5800 (Beckman Coulter, USA) was used. Fungi (1-3)-β-D-glucan (FDG) was tested with an LKM series dynamic test tube detector (Labkinetics LLC, USA). BNP was analyzed using a Cobas e601 full-automatic electrochemical luminescence immunoassay system (Roche, Switzerland). Routine urine tests were performed by an FUS-2000 fully automatic urine analysis workstation (DIRUI, China).

Chest CT image acquisitions
In HUST-UH, all patients underwent CT examinations in the supine position on one of the three CT systems: SOMATOM Definition AS+ (Siemens Healthineers, Germany), Discovery 750HD (GE Medical Systems, Milwaukee, WI) and TOSHIBA Activion 16 (Toshiba, Tokyo, Japan). The scanning range was set from the thoracic inlet to the diaphragm. The scan parameters were 128 × 0.6 mm or 64 × 0.6 collimation, 120 kV tube voltage, and 350 × 350 mm field of view. All datasets were reconstructed with a slice thickness of 1.5-2 mm and an increment of 1.5-2 mm. To ensure the anonymity, rims of original CT images were cropped if containing personal information.

In HUST-LH, the chest CT scan of the patient was performed with a uCT510 spiral CT scanner (United Imaging, China) with a collimator width of 16 × 1.25 mm, a pitch of 1.375, a tube voltage of 120 kV, a tube current of 250 mV, a layer thickness of 1.25 mm, and a layer spacing of 1.25 mm. All patients were in the supine position, and the patients were trained to breathe before the scan. During scanning, patients were asked to hold their breath. The scan ranged from the tip of the lungs to the lower edge of the costal angle. The original data were reconstructed into an image with 1.25-mm layer thickness.
and 1.25-mm layer spacing. Before using, rims of original CT images were cropped if containing personal information.

**Laboratory confirmation of COVID-19 pneumonia**

The etiological confirmation of SARS-CoV-2 infection was conducted using RT-PCR through amplifying ORF1ab gene and N gene of SARS-CoV-2 (BioGerm, Shanghai, China) using throat or nasopharyngeal swab specimens from patients. The sequences for amplifying ORF1ab gene and detecting products were: forward primer 5’-CCCTGTGGGTTTACACTTAA-3’, reverse primer 5’-ACGATTGTGACATCAGCTGA-3’, fluorescence probe 5’-FAM-CCGTCTGCCTGTTATGTGGAAAGTTATGC-BHQ1-3’; the sequences for amplifying N gene and detecting products were: forward primer 5’-GGGGAACTTCTCCTGCTAGAAT-3’, reverse primer 5’-CAGACATTTTGCTCTCAAGCTG-3’, fluorescence probe 5’-FAM-TTGCTGCTGCTTGACAGATT-TAMRA-3’. A cycle threshold value < 35 (or > 35 but less than 38 for two times) was defined as positive.

**Performance evaluation**

To evaluate the accuracy of HUST-19, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values were counted. Then, we calculated six measurements, including Sn, Sp, accuracy (Ac), positive predictive value (PPV), negative predictive value (NPV), Matthews correlation coefficient (MCC) as below:

\[
Sn = \frac{TP}{TP + FN}, \quad Sp = \frac{TN}{TN + FP}, \quad Ac = \frac{TP + TN}{TP + FP + TN + FN},
\]

\[
PPV = \frac{TP}{TP + FP}, \quad NPV = \frac{TN}{TN + FN},
\]

\[
MCC = \frac{TP \times TN - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}
\]

For each method, the 10-fold cross-validation was repeatedly performed 10 times, and average Sn, Sp, Ac, PPV, NPV and MCC values were calculated. The receiver operating characteristic (ROC) curve was illustrated based on final Sn and 1-Sp scores, and the AUC value was computed. For the computational annotation of suspected cases, the t-SNE analysis was implemented in Scikit-learn 0.21.2 (https://scikit-learn.org/stable/), a powerful package for data mining and analysis.
Chest CT image pre-processing
In order to extract the lung parenchyma from chest CT images, we implemented an integrative pipeline for CT image pre-processing. First, the adaptive threshold segmentation (ATS) algorithm was applied to convert gray-scale CT images into binary CT images. Then, we removed the noise of CT images by only keeping the largest connected architecture, and human body profile. After such a manipulation, the background of a CT image was black, and the black lung parenchyma was surrounded by the white human body profile. In order to retain only the black lung parenchymal area, the flood fill algorithm was used to fill the CT image background with white color started from the border of CT images. Finally, the lung parenchyma in the original CT image was extracted. For non-squared CT images, we cropped these images into squared CT images before image pre-processing. In order to rapidly process and analyze such a large dataset of CT images and unify the size of CT images, we re-scaled all CT images into 200 × 200 pixels and effectively avoided distortion by using bilinear interpolation. OpenCV 3.4.2 (https://opencv.org/) and Scikit-image 0.15.0 (https://scikit-image.org/), two powerful computer vision libraries, were adopted for CT image pre-processing.

The 13-layer CNNs
We used the 2 sets of 13-layer CNNs for the image-based and patient-based predictions, respectively. In each CNN framework, there were one input layer, 3 sets of dual convolutional and pooling layers, 2 dense layers and one output layer (Fig. 3). In the 11 hidden layers, neurons were the basic computation units, and both internal feature coding and computational outcome were connected and propagated by neurons inside each layer. The convolutional layers were used for feature extraction and presentation, and a widely-used rectified linear unit (ReLU) function was adopted to activate the outcome of a neuron and defined as below:

\[
\text{ReLU}(x) = \begin{cases} 
  x, & x \geq 0 \\
  0, & x < 0
\end{cases}
\]

Where \(x\) was the weighted sum of a neuron.
In the pooling layers, feature selection and information filtering were performed by the max pooling strategy. The last 2 hidden layers were dense layers for generating prediction outcomes. In order to prevent overfitting that frequently occurs in deep learning algorithms, we used a simple dropout method to randomly select a number of nodes from the 2 dense layers and set their corresponding scores to 0 if the average $Ac$ value went up. In the output layer, 3 sigmoid nodes were set to separately calculate 3 scores for an inputted CT image shown as below:

$$Score(y) = \frac{1}{1 + e^{-y}}$$

Where $y$ was the input of a sigmoid node derived from the dense layer. In the CNN model for the image-based prediction, a $Score(y)$ was a 0~1 value to represent the probability of a CT image classified as a NiCT, pCT or nCT image. For the patient-based prediction, a $Score(y)$ was a 0~1 value to reflect the probability of a patient to be a control case, Type I patient, or Type II patient.

**Normalization of CF data and the 7-layer DNNs**

For each patient, a diagnosed value of a CF was $f$, and $f$ was normalized as below:

$$F = \frac{f - Min}{Max - Min}$$

Where $F$ was the normalized value of $f$, the normal range of the CF was Min to Max. If $f$ was an unavailable value, we set $F$ to 0.5. For two CFs of gender and Udis, we used 0 or 1 to encode males or females, and adopted 0 and 1 to encode patients with and without Udis, respectively.

To enable the classification of patients based on normalized CFs, we used 7-layer DNNs, including one input layer, 5 dense layers and one output layer (Fig. 3). Again, to avoid over-fitting, the dropout method was used by randomly dropping nodes from the 5 hidden layers if the average $Ac$ value went up. In the first step, the input layer received numerical values of CFs for each patient. The 5 hidden layers were mainly adopted for feature extraction and representation. The ReLU activation function was used to transform data for each node. Again, the output layer contained 3 sigmoid neurons to
individually calculate 3 values ranging from 0 to 1 for each patient. Finally, the DNN model was trained for classification of patients into the 3 types.

The PLR algorithm
The integration of predictions from CT images and CFs were performed by the PLR algorithm, which was implemented in Python 3.7 with Scikit-learn 0.21.2. For each patient, CNN models and DNN models were individually used to calculate 3 scores, respectively. Then, the 6 intermediate values were taken as secondary features, and the weight score of each value was initially set to 1. The ridge regression (L2 regularization) penalty was adopted to optimize the weight scores if the average $Ac$ value went up. Finally, the PLR model calculated 3 scores for predicting control cases, Type I or Type II patients.

Model training and parameter optimization
To train the 13-layer CNN models for individual CT image-based prediction, we randomly generated a training data set and a testing data set with a ratio of approximately 9:1, in which the labelled NiCT, pCT and nCT images were proportionally distributed. For the training data set, we further randomly split it into 10 parts, in which 9 of them were used for model training. Then, we used the remaining one part to calculate the average $Ac$ value for predicting the three types of CT images, and the process of parameter optimization was stopped until the $Ac$ value was not increased any longer. The randomization and parameter optimization on the training data set was repeatedly performed 10 times, and the model with the highest $Ac$ value was reserved. Using the determined parameters, the final model was trained on the full training data set. The testing data set was not used for training, and it was only used to count $TP$, $TN$, $FP$ and $FN$ values and calculate the performance measurements. To avoid any bias, the 10-fold cross-validation on the full data set was randomly repeated 10 times. For the 10 best models, the average $Sn$, $Sp$, $Ac$, $PPV$, $NPV$, $MCC$ and AUC values were computed, and a confusion matrix was visualized. Such a 10-fold cross-validation was also adopted for the CT image-based prediction of patients, the CF-based
prediction of patients, and the integration of predictions from CT images and CFs.

For model training, we used a lab computer with an Intel(R) Core(TM) i7-6700K@4.00 GHz central processing unit (CPU), 32 GB of RAM and a NVIDIA GeForce GTX 1070 core. The Keras version 2.2.4 (http://github.com/fchollet/keras), a highly useful neural networks API that was written in Python and developed based on the TensorFlow 1.13.1 (https://github.com/tensorflow), was adopted for parallel computing. For CNNs, the Adam optimizer in Keras was adopted, by using learning rates of 0.001 and 0.0007 for CT image-based and patient-based predictions, respectively, with an additional parameter of 64 for mini-batch size. For DNNs, transient parameters such as the learning rate, mini-batch size, and dropout probability were simultaneously optimized to achieve optimal performance.

**Data Availability**

All source data sets including chest CT images, CFs and laboratory confirmations were archived and maintained in iCTCF at http://ictcf.biocuckoo.cn/. All source codes and the computational model of HUST-19 is available at: http://ictcf.biocuckoo.cn/HUST-19.php. Manually labelled 19,685 original CT images including 5705 NiCT, 4001 pCT and 9979 nCT images were also downloadable at: http://ictcf.biocuckoo.cn/HUST-19.php.
References

1. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5**, 536-544 (2020).
2. Zhu, N., *et al.* A novel coronavirus from patients with pneumonia in china, 2019. *N. Engl. J. Med.* **382**, 727-733 (2020).
3. Wu, F., *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269 (2020).
4. Kannan, S., Shaik Syed Ali, P., Sheeza, A. & Hemalatha, K. COVID-19 (Novel Coronavirus 2019) - recent trends. *Eur. Rev. Med. Pharmacol. Sci.* **24**, 2006-2011 (2020).
5. Remuzzi, A. & Remuzzi, G. COVID-19 and Italy: what next? *Lancet* doi: 10.1016/S0140-6736(20)30627-9 (2020).
6. Zu, Z.Y., *et al.* Coronavirus disease 2019 (COVID-19): A Perspective from China. *Radiology* 200490 (2020).
7. Zhou, P., *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273 (2020).
8. Li, Q., *et al.* Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199-1207 (2020).
9. Lu, R., *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565-574 (2020).
10. Huang, C., *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497-506 (2020).
11. Chan, J.F., *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* **395**, 514-523 (2020).
12. Liu, R., *et al.* Positive rate of RT-PCR detection of SARS-CoV-2 infection in 4880 cases from one hospital in Wuhan, China, from Jan to Feb 2020. *Clin. Chim. Acta.* **505**, 172-175 (2020).
13. Rodriguez-Morales, A.J., *et al.* Clinical, laboratory and imaging features of COVID-19: A systematic review and meta-analysis. *Travel Med. Infect. Dis.* **101623** (2020).
14. Shi, H., *et al.* Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *Lancet Infect. dis.* doi: 10.1016/S1473-3099(20)30086-4 (2020).
15. Yan, L., *et al.* A machine learning-based model for survival prediction in patients with severe COVID-19 infection. *medRxiv* 2020.02.27.20028027 (2020).
16. Xu, Y.H., *et al.* Clinical and computed tomographic imaging features of novel coronavirus pneumonia caused by SARS-CoV-2. *J. Infect.* **80**, 394-400 (2020).
17. Li, K., *et al.* The clinical and chest CT features associated with severe and
critical COVID-19 pneumonia. *Invest. Radiol.* doi: 10.1097/RLI.0000000000000672 (2020).

18. Xia, W., *et al.* Clinical and CT features in pediatric patients with COVID-19 infection: Different points from adults. *Pediatr. Pulmonol.* doi: 10.1002/ppul.24718 (2020).

19. Chen, N., *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* **395**, 507-513 (2020).

20. Yu, J., Ouyang, W., Chua, M.L.K. & Xie, C. SARS-CoV-2 transmission in patients with cancer at a tertiary care hospital in Wuhan, China. *JAMA Oncol.* doi: 10.1001/jamaoncol.2020.0980 (2020).

21. Liu, K.C., *et al.* CT manifestations of coronavirus disease-2019: A retrospective analysis of 73 cases by disease severity. *Eur. J. Radiol.* **126**, 108941 (2020).

22. Guan, W.J., *et al.* Clinical characteristics of coronavirus disease 2019 in china. *N. Engl. J. Med.* doi: 10.1056/NEJMoa2002032 (2020).

23. Yang, X., *et al.* Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir. Med.* doi: 10.1016/S2213-2600(20)30079-5 (2020).

24. Wu, C., *et al.* Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 Pneumonia in Wuhan, China. *JAMA intern. med.* (2020).

25. Santosh, K.C. AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data. *J. Med. Syst.* **44**, 93 (2020).

26. Li, D., *et al.* False-negative results of real-time reverse-transcriptase polymerase chain reaction for severe acute respiratory syndrome coronavirus 2: Role of deep-learning-based CT diagnosis and insights from two cases. *Korean J. Radiol.* **21**, 505-508 (2020).

27. Li, L., *et al.* Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology* 200905 (2020).

28. Qi, X., *et al.* Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study. *medRxiv* 2020.02.29.20029603 (2020).

29. Song, Y., *et al.* Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. *medRxiv* 2020.02.23.20026930 (2020).

30. Wang, S., *et al.* A deep learning algorithm using CT images to screen for Coronavirus Disease (COVID-19). *medRxiv* 2020.02.14.20023028 (2020).

31. Zheng, C., *et al.* Deep learning-based detection for COVID-19 from chest CT using weak label. *medRxiv* 2020.03.12.20027185 (2020).

32. Song, C.Y., Xu, J., He, J.Q. & Lu, Y.Q. COVID-19 early warning score: a multi-parameter screening tool to identify highly suspected patients. *medRxiv* 2020.03.05.20031906 (2020).

33. Feng, Z., *et al.* Early prediction of disease progression in 2019 novel
coronavirus pneumonia patients outside Wuhan with CT and clinical characteristics. medRxiv 2020.02.19.20025296 (2020).

34. van der Walt, S., et al. scikit-image: image processing in Python. PeerJ 2, e453 (2014).
Acknowledgements

This work was supported by grants from the Frontier Exploration Program of Huazhong University of Science and Technology (2015TS153), the program for HUST Academic Frontier Youth Team, and the Fundamental Research Funds for the Central Universities (2018KFYYXJ107, 2019kfyRCPY043, 2020kfyXGYJ019).
Author information

These authors contributed equally: Wanshan Ning, Shijun Lei, Jingjing Yang, and Yukun Cao.

Affiliations

Key Laboratory of Molecular Biophysics of Ministry of Education, Hubei Bioinformatics and Molecular Imaging Key Laboratory, Center for Artificial Intelligence Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China
Wanshan Ning, Peiran Jiang, Yaping Guo & Yu Xue

Department of Clinical Laboratory, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China
Shijun Lei, Qianqian Yang, Jiao Zhang, Xiaobei Wang, Fenghua Chen, Zhi Geng & Lin Wang

Research Center for Tissue Engineering and Regenerative Medicine, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China
Shijun Lei, Lin Wang & Zheng Wang

Department of Respiratory and Critical Care Medicine, Liyuan Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430077, China
Jingjing Yang & Yulan Zeng

Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China, Hubei Province Key Laboratory of Molecular Imaging, Wuhan 430022, China
Yukun Cao & Heshui Shi
Department of Laboratory Medicine, Liyuan Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430077, China
Liang Xiong

Department of Radiology, Liyuan Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430077, China
Hongmei Zhou

Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China
Zheng Wang

**Contributions**
Z.W., Y.X., L.W., H.S. and Y.Z. initiated the project and oversaw all aspects of the project. W.N. designed the iCTCF database and HUST-19. S.L., J.Y., and Y.C. collected, classified, labeled, confirmed and analyzed data and patient information. P.J. and Y.G. contributed to the development of iCTCF. Q.Y., J.Z., X.W., F.C., Z.G., L.X. and H.Z. contributed to data collection, labeling and confirmation, and put forward helpful suggestions for the analysis of data. Z.W., Y.X., L.W., W.N. and S.L. wrote the manuscript with input from all the authors. All authors reviewed and approved the manuscript for publication.

**Competing interests**
The authors declare no competing interests.

**Corresponding author**
Correspondence and requests for materials should be addressed to Z.W., Y.X., L.W., H.S., or Y.Z.
Figure Legends

Fig. 1. The data statistics of the study. a The numbers of control patients, patients with mild, regular, severe, or critically ill forms of COVID-19, and suspected patients were enrolled from HUST-UH and HUST-LU. b Numbers of chest CT images of patients with or without COVID-19 pneumonia from the two hospitals. c Representative chest CT images for six types of cases. d Statistical comparisons of CFs between COVID-19 (Type I & II) and Controls, and between Type II and Type I cases (Two-sided unpaired t-test, $p$-value $< 10^{-4}$). Fold changes were also present. CF abbreviations were shown, and their full names are available in Supplementary Table 1.

Fig. 2. The resource page of iCTCF database. a Multiple select options were provided for searching the resource in a customizable manner. b After clicking “Submit”, the results returned in a tabular list. c A general summary of a given patient in the overview section. d One of five representative chest CT images was displayed for each patient. The other images could be seen by clicking horizontal scrolling buttons. e The numerical CFs of a given patient were shown in a tabular list.

Fig. 3. The hybrid-learning architecture of HUST-19, including a 13-layer CNN framework for classification of individual CT images, a second 13-layer CNN framework to transform individual CT image-based prediction into patient-based prediction, a 7-layer DNN framework to predict COVID-19 patients from CFs, and the PLR algorithm used for integration of CT- and CF-based results for predicting control, Type I or Type II patients.

Fig. 4. The performance evaluation of HUST-19 based on 10-fold cross-validations. a-d The individual CT image-based prediction of NiCT, pCT and nCT images, as well as CT-based, CF-based and HUST-19-based prediction of control, Type I or II patients. AUC values were calculated to represent the prediction accuracy. e-h Confusion matrices were derived from the 10-fold cross-validations under the general threshold
for individual CT image-based prediction, CT-based prediction of patients, CF-based prediction of patients, and HUST-19 predictions. i-i Confusion matrices of the four types of predictive models under the sensitive threshold. More details on the performance evaluation were available in Supplementary Table 4.

**Fig. 5.** Computational annotations of 299 suspected cases without SARS-CoV-2 nucleic acid laboratory confirmations at the time of enrollment. a HUST-19 was used for classifying 229 suspected cases into controls, Type I patients, or Type II patients, using the sensitive threshold (Supplementary Table 4). b The t-SNE analysis of the classification efficiency of HUST-19 on the classification described in a. c Schematics showed the clinical courses of two suspected cases, Patient 324 and Patient 610, who were predicted by HUST-19 to be Type I and Type II cases, respectively.
**Patient 4**

- **Hospital:** HUST-Union Hospital
- **Age:** 73
- **Gender:** Male
- **Body temperature:** 38.5°C
- **Underlying diseases:** Aorta calcification
- **SARS-CoV-2 nucleic acids:** Positive
- **Computed tomography (CT):** Positive
- **Type:** Severe
**a**

Suspected

HUST-19

CT

CF

Bar graph:
- Control: 21
- Type I: 207
- Type II: 71

**b**

Patient 610
Patient 324

Control
Type I
Type II
Suspected

**c**

CT: positive
SARS-CoV-2 nucleic acids: negative

Feb 2
Feb 3
Feb 10
Feb 11
Feb 13
Feb 17
Feb 20
Mar 6

- Fever (38.3°C)
- Mild cough
- Diarrhea
- Fatigue

Feb 2
Feb 3
Feb 5
Feb 6
Feb 7
Feb 23
Feb 25
Mar 2

- Eosinophil count ↓
- Lymphocyte count ↓
- C-reactive protein ↑
- Interleukin-6 ↑

CT: improvement in lung lesions

COVID-19 serum IgM and IgG: positive

CT: positive
SARS-CoV-2 nucleic acids: negative

Jan 31
Feb 3
Feb 5
Feb 6
Feb 7
Feb 23
Feb 25
Mar 2

- Fever (38.4°C)
- Headache
- Dizziness
- Cough
- Phlegm

COVID-19 serum IgM and IgG: positive

Eosinophil count ↓
D-Dimer ↑
Erythrocyte sedimentation rate ↑
Finger blood oxygen saturation: 86%