Long Short-Term Memory Based Framework for Longitudinal Assessment of COVID-19 Using CT Imaging and Laboratory Data

CHONG CHEN, RUI LI, HONG SHEN, AND LIMING XIA

1Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China
2Shanghai United Imaging Intelligence Healthcare Company Ltd., Shanghai 200232, China
Corresponding author: Liming Xia (xialiming2017@outlook.com)

ABSTRACT Automatic longitudinal assessment of the disease progression of coronavirus disease 2019 (COVID-19) is invaluable to ensure timely treatment for severe or critical patients. An artificial intelligence system that combines chest computed tomography (CT) and laboratory examinations may provide a more accurate diagnosis. To explore an artificial intelligence solution to longitudinally assess the condition of COVID-19 using CT imaging and laboratory findings, from January 27, 2020, to April 3, 2020, multiple follow-up examinations of COVID-19 inpatients were retrospectively collected. CT imaging features were automatically extracted using a deep learning method and combined with laboratory tests. The progression sequences were generated with two follow-ups, each of which contained 60 imaging and 24 laboratory features. Pearson’s correlation was conducted to rank the importance of each univariate feature, and multivariate logistic regression was adopted for feature selection. The selected features were used to train a 2-layer long short-term memory network (LSTM) with pulse oxygen saturation ($\text{SpO}_2$) as an indicator of disease progression in three classes: alleviated, stable, and aggravated. The performance of models trained on various feature subsets was compared with five-fold cross validation. 559 patients with 1734 examinations were collected, and 1450 progression sequences were generated. Of the 559 patients, 262 (46.9%) were male. The mean age of the patients was $60 \pm 14$ years. The mean hospitalization duration was $31 \pm 12$ days. Based on the ranking of importance, 26 features from the imaging and laboratory tests were selected, achieving the best accuracy of 0.85 for progression assessment. The comparisons demonstrated that CT features outperformed laboratory features. The best sensitivities for alleviated and aggravated obtained with CT features alone were 0.83 and 0.85, respectively, while laboratory features improved the assessment precision by about 3%. Longitudinal assessment using deep learning with combined features from CT imaging and laboratory tests better predicts the progression of COVID-19 than either of them.

INDEX TERMS Artificial intelligence, COVID-19, computed tomography, image analysis.

I. INTRODUCTION AND BACKGROUND
Coronavirus virus disease 2019 (COVID-19) has developed into a worldwide disease with high transmission [1]. With the rapid increase in the number of admitted patients, a quick assessment of the progression of COVID-19 is required for the management of medical resources. Further, follow-up of COVID-19 patients are necessary to assess prognosis [2]. Therefore, an objective longitudinal assessment of the disease condition is needed as it may provide a monitoring factor for the pandemic.

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

Risk factors associated with COVID-19 have been identified from clinical symptoms, laboratory tests, and computed tomography (CT) scans since the outbreak of the epidemic [3]. These factors have been investigated through statistical analysis or CT imaging phenomena, such as ground-glass opacification [3]–[6]. Complex and manual procedures are time-consuming, and the sample size is limited. As an increasing number of patients have been confirmed, the available samples can be used, making it possible to adopt artificial intelligence (AI) technology and play an important role in the diagnosis of COVID-19 [3], [7]. Several AI-based methods have been exploited as assistant tools to automatically segment both the lung and infected regions [8]. They
helped radiologists not only accurately detect the presence of infection but also distinguish COVID-19 from community-acquired pneumonia (CAP) [8]–[11].

In addition to the diagnosis of COVID-19, immediate triage of the disease is invaluable for ensuring timely treatment for severe or critical patients. Clinical data, including age, fever, lactate dehydrogenase, C-reactive protein, and lymphocytes, have proven to be useful for differentiating non-severe from severe patients [11]–[13]. Meanwhile, the consolidation pattern in chest CT imaging has been used to assess progression to severe stages [6], [14]–[17]. However, most of the existing research has focused on cross-sectional analyses. This means that only the clinical information from the first available examination after symptom onset or hospital admission was used. In fact, the majority of patients would undergo multiple examinations, especially in severe and critical cases, during their disease course. Proper longitudinal comparisons between follow-up examinations are necessary to dynamically evaluate the disease condition. The features extracted from the laboratory tests and CT scans were analyzed separately in the published literature. The incorporation of laboratory and CT imaging features in a pairwise manner is essential for more reliable assessment. Although deep learning has been applied in the longitudinal follow-up of interstitial lung diseases with CT [18], its applications in the progression of COVID-19 solutions are rare.

Our study aimed to represent multiple follow-up examinations of patients with a time series to longitudinally assess disease progression using a deep learning technique. To take advantage of information from multiple CT and laboratory examinations, we proposed an end-to-end AI algorithm that uses CT imaging features generated with deep learning and laboratory tests from two longitudinal exams. Quantitative CT imaging and paired laboratory features were used as inputs for a long short-term memory (LSTM) network for more reliable assessments. Instead of subjective evaluation, SpO\textsubscript{2} was used to label the severity of COVID-19 for each examination, as patients with lower SpO\textsubscript{2} were clinically regarded as severe [19], [20]. The main contributions of this study are as follows: First, our study focused on the longitudinal assessment of the disease progression of COVID-19. In contrast with the existing cross-sectional diagnosis, the proposed framework can handle all historical examinations after admission and provide an in-time assessment of the disease condition during the disease course. It is important to identify patients who progress to severe or critical disease. Second, our work formulates multiple follow-up examinations of patients with time series, making it possible to longitudinally assess the disease progression of COVID-19 with LSTM. Compared with traditional machine-learning methods, the over-fitness caused by an imbalance in the samples can be improved. Moreover, the subtraction operation of features used in the existing methods becomes more difficult when dealing with clinical data extracted from more than two follow-ups. With LSTM, the examination from each follow-up can be taken as a time node, resulting in a time series after collecting all follow-ups during the disease course. Finally, our method combines the quantitative features obtained from both CT and laboratory tests, which avoids the potential inaccuracy generated when single-modality information is used alone.

II. METHODS
A. METHODOLOGY
The assessment of disease progression is based on multiple examinations that need to be performed using sequence data. In addition, to take advantage of different examinations, it is necessary to combine the CT imaging features and laboratory test results in a neural network. In this study, one-dimensional features (i.e., infection volume and infection ratio of the whole lung, lung lobes, and lung segments) extracted from CT imaging were combined with laboratory test results for the longitudinal assessment of disease progression. Thus, the proposed method consists of two neural networks: one for the segmentation of the infection area and lung, and one for the prediction of disease progression.

As we know, convolutional neural networks are widely used in object detection, classification and segmentation of imaging. U-net and V-net have excellent performance in the segmentation of medical images, and accurate segmentation is the basis of the following steps: classification of the infection area and extraction of radiomics features [3], [8], [21]. Therefore, software based on VB-net (a derivation of the well-known V-Net) is used for the segmentation of infection volume and infection ratio to generate one-dimensional CT features.

A recurrent neural network (RNN) is widely used for the processing of sequence data, which consists of a sequence of directed nodes to perform the forward on the corresponding input vector and the previous state at each time step. Compared with the conventional RNN, LSTM (a variant of RNN) adopts a cell unit to select important information and exclude irrelevant features [22], [23] by introducing three gates: input gate, output gate, and forget gate. Owing to the capability of memory of past inputs, long short-term memory networks provide better performance for solving sequence learning tasks by alleviating what is called the vanishing gradient problem. In addition, to reduce the influence of confounding factors and multicollinearity, Pearson’s correlation and multivariate logistic regression were used for the selection of features.

B. PATIENTS AND DATA COLLECTION
This retrospective study was approved by the Ethics Committee of Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China. De-identified data were collected, and the requirement for informed consent was waived. From January 27, 2020, to April 3, 2020, patients with confirmed COVID-19 according to the severe acute respiratory syndrome coronavirus 2 nucleic acid test at Tongji Hospital were included. The
exclusion criteria were as follows: (1) patients without follow-up CT, (2) age younger than 18 years, and (3) isolated CT scans without laboratory tests or SpO₂ examination performed within 3 days. Deidentified data, including chest CT imaging, laboratory tests, and medical record information, were collected from the hospital’s information system. Chest CT imaging with paired laboratory tests within 3 days was included. (Fig. 1).

C. PROGRESSION ASSESSMENT

Multiple follow-up examinations of each patient were sorted by date in ascending order. Progression was assessed based on imaging and laboratory features extracted from two adjacent follow-ups. Progression labels were assigned to the examination pair based on the severity labels of the two examinations.

According to COVID-19 guidelines [24], we categorized the severity of an examination into two types: non-severe (mild/moderate) and severe (severe and critical). Therefore, three classes of progression were obtained after comparing the two follow-up examinations: alleviated (class 0), stable (class 1), and aggravated (class 2). There were many fewer samples in classes 0 and 2; therefore, nonadjacent examinations were also included for these two classes to achieve a better sample balance with class 1. The longitudinal assessment of COVID-19 progression was formulated as a sequence classification problem (Fig. 2).

With the automatic platform developed by our cooperative institution, both the lung and infected regions were segmented using convolutional neural networks (CNNs) from chest CT images, and quantitative imaging features were calculated. Laboratory tests within three days of CT were collected and concatenated with imaging features to generate complete clinical features. The complete clinical features from multiple follow-ups formed the input sequence and were fed into the sequence classification network. A 3-dimensional vector was finally output, in which each element corresponded to the probability of three disease progression statuses: alleviated, stable, and aggravated.

Pulse oximetry is an objective measure to evaluate the severity of patients instead of subjective scoring, and is widely used for monitoring patients [19], [20]. Therefore, pulse oxygen saturation (SpO₂) was used to classify disease progression into three types: alleviated, stable, and aggravated.

D. EXPLAINABLE IMAGING FEATURES CALCULATION WITH CNN

Considering its clinical significance, the parameters involved in the AI model should be explainable. Therefore, quantitative imaging features related to COVID-19 were calculated from the CT images based on the segmentation results of the lung and infection. All imaging features were obtained according to the clinical experience of the radiologist and could describe the severity of the infection from CT data to some degree.

Quantitative analysis of lung volumes infected by COVID-19 was performed using uAI Discover-2019nCoV from Shanghai United Imaging Intelligence Healthcare Co., Ltd. [25]. Our software utilized a deep learning-based network VB-Net, a derivation of the well-known V-Net [26] by combining a bottle-neck structure [27] to segment the infected regions as well as the lung anatomies from chest CT images. Using the segmentation algorithm, 26 substructures of the lung were obtained: 1) the whole lung (WL), left lung (LL), right lung (RL), and five lung lobes, that is, superior/middle/inferior lobes of the right lung (LBS/I, RBS/M/I), and superior/inferior lobes of the left lung (SL, IL); and 2) 18 lung segments, that is, eight segments of the left lung and 10 segments of the right lung (LS₁₈, RS₁_1₀). In addition, four additional regions associated with pathology were included: normal lung regions (HU (−∞, −750)), ground glass opacity (GGO) regions (HU [−750, −300]), consolidation regions (HU [−300, 50]), and regions of calcification (HU (50, +∞)) [28], resulting in a total of 30 regions of interest.

Apart from the segmentation of lung substructures, the uAI Discover-2019nCoV platform automatically outputs masks from infected regions. To measure the infection of the extracted 30 regions of interest, the infection volume (IV) and infection ratio (IR) were calculated for each region as follows:

$$V(x) = N(x) \times \text{VPV}, \quad \text{IV}(x) = V(x_{\text{infect}})$$

$$\text{IR}(x) = \frac{\text{IV}(x)}{V(x)} = \frac{V(x_{\text{infect}})}{V(x)}$$

where \(N(x)\) is the number of voxels in region \(x\) and volume per voxel (VPV) was the volume of a voxel. \(x_{\text{infect}}\) indicates the infected region within region \(x\).

As a result, a total of 60 quantitative CT features were obtained for each patient, which are summarized in Table 1.

E. LABORATORY FEATURES

Laboratory tests, including routine blood tests and C-reactive protein (CRP) and lactate dehydrogenase (LDH) levels, were collected from the laboratory information system. Laboratory tests were conducted within three days before or after CT scans were performed. The SpO₂ for each patient was monitored daily using a pulse oximeter, and the minimal SpO₂ value was used in this study. Missing laboratory test values for some examinations were padded with a weighted sum of the available values in other examinations of the same patient and in the examinations of other patients with the same severity level (Appendix).

F. COMBINING IMAGING AND LABORATORY FEATURES

Each CT scan of a patient was paired with laboratory tests performed within three days of CT imaging. If a laboratory test was conducted multiple times during the available period, the test closest to the CT scan was chosen.

G. TRAINING OF LSTM

Our aim was to longitudinally assess disease progression by comparing the clinical data obtained from two follow-up...
examinations. In traditional methods, the subtraction of features is often used to train a machine learning model, such as the linear regression model [29] and XGBoost [12].

In our experiment, however, the sample number of aggravation (class 2) was usually smaller than the others because most patients gradually recovered during the treatment. The imbalance in the samples easily leads to over-fitness of these methods. This would result in a lower sensitivity for aggravation, which is of great significance for clinical use. However, it is believed that more reliable results will be produced when an AI model receives as much historical clinical data as possible. The subtraction operation of features becomes more difficult when dealing with clinical data extracted from more than two follow-ups. In such cases, the examination from
each follow-up can be considered as a time node, resulting in a time series after collecting all follow-ups during the disease course. Inspired by the fact that LSTM [30] has been proven to perform well in processing time-series data from electronic medical records [31], [32], we formulated a longitudinal assessment as a sequence classification with LSTM.

Considering the limited data, a 2-layer LSTM followed by two fully connected layers, was designed to perform the task. The length of the input sequence was two; that is, the (i-1)th and ith follow-ups were used because of the limited sample size. The hyperparameters of the network were empirically set. Specifically, the dimension of the hidden layer in LSTM was 256. Equal sampling and Focalloss [33] were adopted to avoid overfitting. The Adam optimizer was used in this study. Epoch was set to 300. All the parameters of the network and training setup were kept constant throughout the experiments. Five-fold cross-validation was performed to account for sample imbalance, which was repeated 20 times, and the average performance was reported for evaluation. CT imaging segmentation and LSTM model training were performed using a high-performance computer with two NVIDIA GTX 1080Ti GPUs.

**H. STATISTICAL ANALYSIS**

Statistical analyses were performed using Python (version 3.7.5). The Pearson correlation coefficient [34] between each independent feature and SpO2 was calculated to evaluate its importance. Irrelevant features (Pearson’s correlation coefficient, $r < 0.1$) were removed. The relevant features were categorized into three groups: infection volume (CT_IV), infection ratio (CT_IR), and laboratory findings (Labs).
TABLE 2. Patient characteristics and 24 laboratory findings.

| Variables                  | Normal range | Non-severe | Severe |
|----------------------------|--------------|------------|--------|
| Patients                   | 205          | 354        |        |
| Age                        | 58 (45-68)   | 65 (55-71) |        |
| Sex                        |              |            |        |
| Male                       | 101 (49%)    | 161 (45%)  |        |
| Female                     | 104 (51%)    | 193 (55%)  |        |
| Days in hospital           | 20 (17-25)   | 33 (25-42) |        |

Laboratory findings

|                          |              |            |        |
|--------------------------|--------------|------------|--------|
| PLT Distribution Width (fL) | 9.8-21.9    | 12.10 (10.90-13.50) | 12.20 (10.70-13.75) |
| RBC Distribution Width CV | <14.9        | 13.00 (12.30-13.80) | 12.80 (12.20-13.65) |
| RBC Distribution Width SD (fL) | 39.0-46.0 | 42.50 (39.80-45.70) | 41.40 (39.15-44.90) |
| Neutrophil count (*10^3/L)  | 1.80-6.30    | 3.37 (2.54-4.90) | 4.58 (3.16-7.09) |
| Neutrophil percentage (%)  | 40.0-75.0    | 61.50 (53.70-71.00) | 73.50 (61.10-83.60) |
| Monocyte count (*10^3/L)    | 0.10-0.60    | 0.52 (0.40-0.65) | 0.52 (0.40-0.68) |
| Monocyte percentage (%)     | 3.0-10.0     | 9.00 (7.30-11.00) | 8.10 (5.80-9.80) |
| Basophil count (*10^3/L)     | 0.00-0.10    | 0.02 (0.01-0.04) | 0.02 (0.01-0.03) |
| Basophil percentage (%)     | 0.0-1.0      | 0.40 (0.20-0.60) | 0.20 (0.10-0.50) |
| Eosinophil count (*10^3/L)   | 0.02-0.52    | 0.11 (0.04-0.19) | 0.05 (0.01-0.14) |
| Eosinophil percentage (%)   | 0.4-8.0      | 1.90 (0.80-3.50) | 0.80 (0.10-2.45) |
| Platelet large cell ratio (%) | 16.1-55.4   | 29.30 (24.80-34.80) | 30.10 (24.05-35.75) |
| Mean PLT Volume (fL)        | 9.0-13.8     | 10.60 (10.00-11.20) | 10.70 (9.95-11.40) |
| Mean RBC Volume (fL)        | 82.0-100.0   | 90.70 (87.80-93.50) | 90.00 (87.00-93.45) |
| Mean hemoglobin content (pg) | 27.0-34.0   | 30.80 (29.70-32.00) | 30.80 (29.70-32.00) |
| Mean hemoglobin concentration (g/L) | 316-354 | 340.00 (333.00-347.00) | 342.00 (335.00-350.00) |
| Lymphocyte count (*10^3/L)  | 1.10-3.20    | 1.40 (1.03-1.77) | 1.04 (0.69-1.42) |
| Lymphocyte percentage (%)   | 120.0-50.0   | 25.60 (17.90-32.50) | 15.60 (10.05-25.60) |
| Hematocrit (%)               | 35.0-45.0    | 35.30 (32.40-38.60) | 36.00 (32.45-39.05) |
| Thrombocytocrit (%)          | 0.14-0.39    | 0.23 (0.18-0.28) | 0.23 (0.18-0.29) |
| Platelet count (*10^3/L)     | 125.0-350.0  | 213.00 (168.00-271.50) | 217.00 (157.50-284.00) |
| Hemoglobin (g/dL)            | 115.0-150.0  | 120.00 (110.00-132.00) | 123.00 (109.50-135.50) |
| C-reactive protein (mg/L)    | < 1          | 4.40 (1.40-10.90) | 15.00 (4.10-46.04) |
| Lactate dehydrogenase level (U/L) | 135-214 | 205.38 (177.00-242.00) | 238.00 (198.00-301.50) |

Data are presented as median (interquartile range, IQR). PLT: platelets, RBC: red blood cell.

In each group, multivariate logistic regression discriminating an exam as severe or non-severe was performed to select the optimal feature subset using a forward search strategy [35]. Confusion matrix and receiver operating characteristic (ROC) analyses were performed to evaluate the performance of progression assessment. Differences were considered statistically significant at $P < 0.05$.

III. RESULTS

A total of 559 patients with complete imaging and laboratory findings were included. Of the 559 patients, 262 (46.9%) were male (Table 2). The mean age of the patients was 60±14 years. The mean hospitalization duration was 31±12 days. The minimal, median, and maximal time intervals between the follow-ups were 2, 9, and 49 days, respectively. These patients were categorized into three clinical types based on the 2019-nCoV guidelines (trial version 7): non-severe (205 patients), severe (331 patients), and critical (23 patients) [24]. Older patients are more likely to develop more severe or critical conditions. No significant differences were observed in sex among the three clinical types.

Each patient underwent at least two examinations, 148 patients underwent three examinations, and 139 patients underwent more than four examinations. A total of 1734 examinations were obtained from 559 patients, 179 of which were determined to be severe. 60 quantitative imaging features and 24 laboratory findings were extracted for each examination (Table 2). In total, 1450 pairwise disease progression samples were obtained for training and testing, including 333 alleviated, 1010 stable, and 107 aggravated samples. (Table 3).
A. UNIVARIATE CORRELATION ANALYSIS

The Pearson correlation in Table 4 demonstrates that each individual feature is weakly correlated with \( \text{SpO}_2 \) (\( r < 0.4 \)). Generally, CT features exhibit a stronger correlation with \( \text{SpO}_2 \) than with laboratory findings. Most infection ratio features yielded a similar coefficient \( r \) ranging from 0.15 to 0.35, while the coefficient \( r \) for laboratory features ranged from 0.01 to 0.26. Specifically, only 4 of the 24 laboratory features were correlated with \( \text{SpO}_2 \) with Pearson coefficients greater than 0.2, that is, LDH (\( r = 0.26 \)), lymphocyte percentage (\( r = 0.22 \)), and CRP (\( r = 0.23 \)). In contrast, 18 of the 30 features in CT_IR and 12 of the 30 features in CT_IV were relevant. The infection volume and infection ratio in \( \text{HU}_{[-300,-50]} \) showed a higher correlation with \( \text{SpO}_2 \) than those in \( \text{HU}_{[-750,-300]} \). This is consistent with previous findings that GGO occurs in the initial stage and consolidation occurs with a peak in the total CT severity score [17]. The correlation coefficients for both infection volume and infection ratio were higher in the left lung than in the right lung, and higher in the superior lobes than in the inferior lobes (\( r \) (LL) > \( r \) (RL), \( r \) (LB_S) > \( r \) (LB_I), \( r \) (RB_S) > \( r \) (RB_I)).

The relationship between different features and the time after symptom onset for non-severe and severe examinations was plotted. Typical features with higher correlation coefficients (e.g., \( r > 0.2 \)) and lower coefficients (e.g., \( r < 0.1 \)) are shown in Fig. 3. The features with higher correlation coefficients with \( \text{SpO}_2 \), such as IV (\( \text{HU}_{[-300,50]} \)), IR (LB_S), and lymphocyte percentage (Fig. 3 (a)-(c)), generally demonstrated better discrimination between severe and non-severe examinations. In contrast, limited discrimination power was observed for hemoglobin, eosinophil percentage, and platelet count, as shown in Fig. 3 (c)–(f).

B. FEATURE SELECTION USING MULTIVARIATE LOGISTIC REGRESSION

After filtering out irrelevant features (\( r < 0.1 \)) from the 84 features, 25 features in the infection volume (CT_IV) group, 29 features in the infection ratio (CT_IR) group, and eight features in the laboratory findings (Labs) group were obtained.

Fig. 4 shows the prediction accuracy for three metrics (i.e., accuracy, true positive rate, and false positive rate) while using multivariate logistic regression with different numbers of features in the three groups. In total, 26 features, including 10 CT_IV, 10 CT_IR, and 6 Labs features, were used to train the developed LSTM model after feature selection. Detailed information on these 26 features is presented in Table 4. The area under the curve (AUC) was 0.85 after feature selection, and the detailed ROC curves are shown in Fig. 5.
TABLE 4. Pearson correlation coefficients between each feature and SpO₂.

|                  | r   | P value |                  | r   | P value |                  | Labs            | r   | P value |
|------------------|-----|---------|------------------|-----|---------|------------------|-----------------|-----|---------|
| IV (HUₜ₃₀₅₀)     | 0.36| <10⁻⁴   | IV (HUₜ₅₀₁₀₀₀)  | 0.34| <10⁻⁴   |                 | LDH             | 0.26| <10⁻⁴   |
| IV (HUₜ₅₀₁₀₀₀)  | 0.31| <10⁻⁴   | IV (HUₜ₅₀₁₀₀₀)  | 0.29| <10⁻⁴   |                 | CRP             | 0.23| <10⁻⁴   |
| IV (LBₘ)        | 0.26| <10⁻⁴   | IV (LBₘ)        | 0.27| <10⁻⁴   |                 | Lymphocyte (%)  | 0.22| <10⁻⁴   |
| IV (LL)         | 0.23| <10⁻⁴   | IV (LL)         | 0.27| <10⁻⁴   |                 | Neutrophil (%)  | 0.21| <10⁻⁴   |
| IV (LSₘ)        | 0.23| <10⁻⁴   | IV (LSₘ)        | 0.26| <10⁻⁴   |                 | Neutrophil (%)  | 0.16| <10⁻⁴   |
| IV (RSₘ)        | 0.23| <10⁻⁴   | IV (LSₘ)        | 0.26| <10⁻⁴   |                 | Lymphocyte (%)  | 0.16| <10⁻⁴   |
| IV (RBₘ)        | 0.22| <10⁻⁴   | IV (RBₘ)        | 0.24| <10⁻⁴   |                 | Basophil (%)    | 0.14| <10⁻⁴   |
| IV (LSₘ)        | 0.22| <10⁻⁴   | IV (LSₘ)        | 0.24| <10⁻⁴   |                 | Monocytes (%)   | 0.12| <10⁻⁴   |
| IV (WL)         | 0.21| <10⁻⁴   | IV (LSₘ)        | 0.23| <10⁻⁴   |                 | Basophil (%)    | 0.1  | <10⁻³   |
| IV (LSₘ)        | 0.21| <10⁻⁴   | IV (LSₘ)        | 0.23| <10⁻⁴   |                 | Mean PLT Volume | 0.07| 0.01    |
| IV (LSₘ)        | 0.2  | <10⁻⁴   | IV (LSₘ)        | 0.21| <10⁻⁴   |                 | Platelet Large Cell Ratio | 0.07| 0.02 |
| IV (HUₜ₇₅₀₅₀₀)  | 0.2  | <10⁻⁴   | IV (RSₘ)        | 0.2  | <10⁻⁴   |                 | Eosinophils (%) | 0.06| 0.03    |
| IV (RL)         | 0.18| <10⁻⁴   | IV (HUₜ₇₅₀₅₀₀)  | 0.2  | <10⁻⁴   |                 | PLT Distribution Width | 0.06| 0.05 |
| IV (RSₘ)        | 0.18| <10⁻⁴   | IV (RBₘ)        | 0.2  | <10⁻⁴   |                 | Eosinophils (%) | 0.03| 0.29    |
| IV (RSₘ)        | 0.17| <10⁻⁴   | IV (RSₘ)        | 0.2  | <10⁻⁴   |                 | Mean RBC Volume | 0.03| 0.32    |
| IV (LSₘ)        | 0.17| <10⁻⁴   | IV (RL)         | 0.2  | <10⁻⁴   |                 | Hemoglobin      | 0.03| 0.33    |
| IV (LBₘ)        | 0.16| <10⁻⁴   | IV (LSₘ)        | 0.2  | <10⁻⁴   |                 | RBC Distribution Width CV | 0.03| 0.34 |
| IV (LSₘ)        | 0.14| <10⁻⁴   | IV (RSₘ)        | 0.19| <10⁻⁴   |                 | RBC Distribution Width SD | 0.03| 0.34 |
| IV (RBₘ)        | 0.14| <10⁻⁴   | IV (RSₘ)        | 0.19| <10⁻⁴   |                 | Mean Hemoglobin Concentration | 0.02| 0.4  |
| IV (RSₘ)        | 0.13| <10⁻⁴   | IV (LSₘ)        | 0.19| <10⁻⁴   |                 | Hematocrit      | 0.02| 0.43    |
| IV (RSₘ)        | 0.13| <10⁻⁴   | IV (RBₘ)        | 0.18| <10⁻⁴   |                 | Platelet (%)    | 0.02| 0.59    |
| IV (RSₘ)        | 0.12| <10⁻⁴   | IV (RBₘ)        | 0.17| <10⁻⁴   |                 | Mean Hemoglobin Content | 0.01| 0.62 |
| IV (RSₘ)        | 0.12| <10⁻⁴   | IV (RSₘ)        | 0.17| <10⁻⁴   |                 | Monocytes (%)   | 0   | 0.97    |
| IV (LSₘ)        | 0.12| <10⁻⁴   | IV (RSₘ)        | 0.17| <10⁻⁴   |                 | Thrombocytocrit | 0   | 0.98    |
| IV (RBₘ)        | 0.11| <10⁻³   | IV (LSₘ)        | 0.17| <10⁻⁴   |                 |                 |                 |        |
| IV (LSₘ)        | 0.09| <10⁻²   | IV (RSₘ)        | 0.16| <10⁻⁴   |                 |                 |                 |        |
| IV (RSₘ)        | 0.08| <10⁻²   | IV (RSₘ)        | 0.15| <10⁻⁴   |                 |                 |                 |        |
| IV (RSₘ)        | 0.07| 0.01    | IV (RSₘ)        | 0.14| <10⁻⁴   |                 |                 |                 |        |
| IV (HUₜ₇₅₀₅₀₀)  | 0.04| 0.17    | IV (RSₘ)        | 0.14| <10⁻⁴   |                 |                 |                 |        |
| IV (RSₘ)        | 0.01| 0.64    | IV (HUₜ₇₅₀₅₀₀)  | 0.01| 0.63    |                 |                 |                 |        |

Features were ranked by correlation coefficient in descending order. The selected relevant features after feature selection were highlighted. For laboratory tests, count and percentage of biomarkers were denoted with (%) and (#), respectively. LDH: lactate dehydrogenase, CRP: C-reactive protein (CRP), PLT: platelets, RBC: red blood cell, r: Pearson correlation coefficient.

C. LONGITUDINAL PROGRESSION ASSESSMENT BASED ON TWO FOLLOW-UPS

Using the selected 26 features, the proposed LSTM network was trained on the defined progression dataset. Seven models, that is, CT_IV, CT_IR, Labs, CT_IV + CT_IR, CT_IV + Labs, CT_IR + Labs, and CT_IV + CT_IR + Labs, were generated using different combinations of the three feature groups. The average sensitivity for each class and classification accuracy was shown in Fig. 6. The best sensitivity for alleviated and aggravated obtained when using CT features alone were 0.83 and 0.85, respectively. The difference in classification accuracy between CT_IV and CT_IR when using features was small (0.79 vs. 0.8). In contrast, the model using laboratory features generated the lowest value, 0.7. The model trained on the combined 26 features yielded the best classification accuracy of 0.85, with sensitivities of 0.91, 0.85, and 0.94 for alleviated, stable, and aggravated, respectively. Moreover, performance improvement was achieved when laboratory features were combined into CT feature groups. By comparing CT_IV vs. CT_IV + Labs, CT_IR vs. CT_IR + Labs, and CT_IV+CT_IR vs. CT_IV + CT_IR + Labs, the accuracy was improved by 1-3%. This indicated that laboratory features were helpful in improving the overall performance of the progression assessment models.

Table 5 shows the confusion matrix generated by the model when using the combined 26 features during the
cross-validation. The total classification accuracy was 0.89, with sensitivities of 0.85, 0.91 and 0.86 for alleviated, stable, and aggravated, respectively. Errors mainly occurred between contiguous progression tendencies, whereas misclassifications across levels (i.e., alleviated vs. aggravated) were negligible. This indicates the capability of our model to distinguish between aggravated and alleviated cases.

In clinical practice, patients who progress to disease aggravation require more attention and instant treatment. Therefore, the probabilities for the three progression classes output by the last fully connected layer of the network were recorded. Regarding the aggravated as positive, and the others as negative, ROC was plotted to evaluate the identification performance for Aggravated from others. Different feature subsets are also compared, as shown in Fig. 7. The combined 26 features achieved the best AUC of 0.91. Even though there is an obvious difference between Labs and the other feature subsets, the addition of laboratory features improves the identification capability (dotted curves vs. the solid curves with the same color).
FIGURE 8. Cases in which our model made incorrect classifications The model was trained on the selected 26 features (i.e., the model marked with the green dotted line in Figure 7), as shown in the red box in (a), the lymphocyte percentage returned to the normal range (20%-50%), and the disease remained non-severe from the second examination (14 days after admission) to the third examination (18 days after admission). However, the model output was alleviated, which might be due to the decrease in the infection ratio in HU (−300, 50). Similarly, the purple box in (b) demonstrated that the model was influenced by the increasing infection ratio in HU (−300, 50) and outputted aggregated, in which the patient was still in severity and the progression was stable. These observations suggest that the model learned more information from CT features and thus was more sensitive to CT features than laboratory findings.

Notably, our model was more sensitive to CT features when analyzing cases with incorrect classifications (Fig. 8). Such situations mostly occurred with an apparent change in the infection ratio in HU (−300, 50), implying that the model learned more information from CT features and thus made mistakes.

IV. DISCUSSION
This study retrospectively investigated an LSTM network to longitudinally assess the progression of COVID-19 via time-sequence classification. The proposed method combines features obtained from CT imaging and laboratory tests to predict progression. The LSTM network trained with the selected 26 features produced the best classification accuracy (0.85 for progression assessment. The proposed framework formulated clinical data from multiple follow-ups as a sequence and could handle all historical examinations, even after admission. Therefore, it can capture dynamic clinical information over time for each patient. In addition, the end-to-end framework can provide a dynamic assessment not only during hospitalization, but also during regular follow-up analysis after discharge.

In clinical practice, real-time reverse transcriptase-polymerase chain reaction (RT-PCR) is routinely performed to diagnose COVID-19 [24]. Although some studies have suggested that CT is a reliable approach for this purpose [36], [37], it is not recommended as a screening tool instead of RT-PCR because of the lack of etiological credibility and radiation injury [38]. As an indispensable tool for COVID-19 patients, CT imaging plays an important role in assessing both the clinical status and disease progression, which is critical for the stratification and treatment of patients. CO-RADS has also been developed as a categorical CT assessment scheme for suspected patients [36]. Recent studies have suggested that CT has predictive power in assessing clinical outcomes such as severity, intubation, and mortality [14], [15], [39], [40]. Changes in chest CT were associated with COVID-19 during the disease course; for example, the consolidation pattern that occurred in the third stage (9-13 days) was associated with a peak in the total CT severity score [17]. Our results in this study revealed that the CT pattern and distribution of the lesion in the lung were related to clinical status, which is consistent with the literature [15], [19], [37]. Furthermore, CT imaging features were superior to key laboratory tests, including widely admitted risk factors (LDH, CRP, and lymphocyte percentage) in predicting disease status. This was also validated when more laboratory test results were included [28]. The results of our study further showed that the combined CT imaging features and laboratory tests yielded the best performance in predicting disease progression compared with CT imaging features or laboratory tests alone.

Our study also implies a correlation of some key CT and laboratory findings with SpO2 in diagnosed patients. In accordance with previous studies [6], [17], [20], [41], [42], our findings revealed that the CT pattern in the lung (including GGO and consolidation), distribution of lesions (infection in superior lobes), and laboratory findings (including CRP, lymphocyte percentage, and LDH) were related to SpO2. Therefore, significant pulmonary injury accompanied by pronounced abnormalities in biomarkers may result in acute deterioration of clinical status and severe hypoxemia. Early intervention to attenuate this abnormality may improve oxygenation. With the proposed method, intervention can be performed earlier to prevent the deterioration of the disease. For example, if a moderate patient is more likely to progress to severe disease, proactive treatment (e.g., respiratory support) could be performed. If many patients are likely to experience deterioration, the deployment of medical resources need to be increased accordingly.

In our study, the imaging features were extracted from CT scans, since the medical practices during the course of COVID-19 in Tongji Hospital employed multiple CT scans for each patient and therefore accumulated ample retrospective CT data. Our framework is flexible because it changes the designed features for other purposes. For example, given that chest X-ray (CXR) is more widely used in monitoring disease progression over time, the imaging features extracted from CXR can be used as an alternative to CT imaging features as an input framework for assessing disease progression. More importantly, bedside chest radiography is the main examination method for patients in intensive care units (ICU) owing to restricted movement. It is necessary to train a neural network for the segmentation of the infection area in chest radiography.

In summary, the main contribution of this work is the longitudinal assessment disease progression of COVID-19 using CT imaging and laboratory features. Based on these results, it is feasible to predict disease progression using artificial intelligence based on multimodal tests.
A. LIMITATIONS OF THE STUDY

This study had several limitations. First, most patients in our study were admitted to the hospital in the severe stage. We attempted to distinguish cases with an aggravated clinical status from stable or alleviated cases. Deep analysis for those with stable clinical condition, which might be accompanied by potential signs of the tendency to be severe, was lacking. Second, imaging features were extracted using the automatic segmentation of AI. Even with good performance for both lung and infection segmentation [25], improper segmentation may result in an inaccurate quantitation of these features. In the next step, it is necessary to continue training the neural network for segmentation. In addition, a more reliable but time-consuming solution is to correct the segmentation results of qualified radiologists. Third, testing of datasets in different regions is necessary to address regional variations and general applicability, which should be carried out in the next stage of research. Fourth, multiple CT images were compared using the infection volume and infection ratio of each region (i.e., whole lung, lung lobe, lung segment). In fact, the infected volume may not change significantly when new lesions emerge, as old lesions are absorbed. Calculating the infected volume of each lung segment helps use local information; however, the boundaries of the lung segments are not as clear as the lobes. Further research is needed to explore the role of different methods (e.g., image registration and radiomics) in comparing the details of the two CT images. Finally, only paired CT and laboratory tests were used in the proposed method, which is quite reasonable for the algorithm; however, isolated CT scans or laboratory tests are also essential for the assessment of disease. Based on these results, CT should be more important in the assessment of disease in the absence of laboratory examinations within three days.

V. CONCLUSION

The proposed LSTM model using combined features from CT imaging and laboratory tests could accurately predict the progression of COVID-19 better than when using CT imaging or laboratory features alone. The proposed framework handles the sequence of historical follow-ups for each patient and extracts dynamic clinical information over time to make more reliable assessments.

APPENDIX

Among these examinations, several missed at least one C-reactive protein and lactate dehydrogenase level. These values were padded using the following formula:

\[
\text{Padding value} = \alpha \times \frac{P(f_i - 1) + P(f_i + 1)}{2} + (1 - \alpha) \times \text{mean} \tag{3}
\]

where \(P(f_i)\) is the value of feature \(f\) obtained in the \(i^{th}\) examination in ascending order by the date of the patient, and \(\text{mean}(f)\) denotes the average value of those examinations with the same SpO\(_2\) label. In our study, \(\alpha\) is the weight and was set to 0.8.

Fig. 8 Cases in which our model made incorrect classifications. The model was trained on the selected 26 features (i.e., the model marked with the green dotted line in Fig. 7). As shown in the red box in (a), the lymphocyte percentage returned to the normal range (20%-50%), and the disease remained non-severe from the second examination (14 days after admission) to the third examination (18 days after admission). However, the model output was alleviated, which might be due to the decrease in the infection ratio in HU\([−300,50]\). Similarly, the purple box in (b) demonstrated that the model was influenced by the increasing infection ratio in HU\([−300,50]\) and outputted aggregated, in which the patient was still in severity and the progression was stable. These observations suggest that the model learned more information from CT features and thus was more sensitive to CT features than laboratory findings.

ACKNOWLEDGMENT

(Chong Chen and Rui Li contributed equally to this work.)

REFERENCES

[1] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, and P. Niu, “A novel coronavirus from patients with pneumonia in China, 2019,” New England J. Med., vol. 382, pp. 727–733, Jan. 2020, doi: 10.1056/NEJMojo2001017.
[2] C. Liu, L. Ye, R. Xia, X. Zheng, C. Yuan, Z. Wang, R. Lin, D. Shi, Y. Gao, J. Yao, Q. Sun, X. Wang, and M. Jin, “ Chest computed tomography and clinical follow-up of discharged patients with COVID-19 in wenzhou city, Zhejiang, China,” Ann. Amer. Thoracic Soc., vol. 17, no. 10, pp. 1231–1237, Oct. 2020, doi: 10.1513/AnnalsATS.202004-324OC.
[3] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. Bonten, D. L. Dahly, and J. A. Damen, “Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal,” Brit. Med. J., vol. 369, Apr. 2020, Art. no. m1328, doi: 10.1136/bmj.m1328.
[4] Y. Sun, V. Koh, K. Marinuthu, O. T. Ng, B. Young, S. Vasoo, M. Chan, V. J. Lee, P. P. De, T. Barkham, and R. T. Lin, “Epidemiological and clinical predictors of COVID-19,” Clin Infectious Diseases, vol. 71, no. 15, pp. 786–792, Jul. 2020, doi: 10.1093/cid/ciaa322.
[5] Y. Cao, Z. Xu, J. Feng, C. Jin, X. Han, H. Wu, and H. Shi, “Longitudinal assessment of COVID-19 using a deep Learning–based quantitative CT pipeline: Illustration of two cases,” Radiol., Cardiothoracic Imaging, vol. 2, no. 2, Apr. 2020, Art. no. e200082, doi: 10.1148/rct.2020020082.
[6] A. Bernheim, X. Mei, M. Huang, Y. Yang, Z. A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li, S. Li, H. Shan, A. Jacobi, and M. Chung, “Chest CT findings in coronavirus disease-19 (COVID-19): Relationship to duration of infection,” Radiology, vol. 295, no. 3, Jun. 2020, Art. no. 200463, doi: 10.1148/radiol.2020200463.
[7] X. Mei, H. C. Lee, K. Y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie, M. Chan, P. M. Robson, M. Chung, and A. Bernheim, “Artificial intelligence-enabled rapid diagnosis of patients with COVID-19,” Nature Med., vol. 26, no. 8, pp. 1224–1228, Aug. 2020, doi: 10.1038/s41591-020-0931-3.
[8] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, K. Cao, D. Liu, G. Wang, Q. Xu, X. Fang, S. Zhang, J. Xie, and J. Xia, “Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy,” Radiology, vol. 296, no. 2, pp. E65–E71, Aug. 2020, doi: 10.1148/radiol.2020200905.
[9] S. Hu, G. Yang, H. Ye, J. Xia, W. Menpes-Smith, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, and E. F. Fang, “Weakly supervised deep learning for COVID-19 infection detection and classification from CT images,” IEEE Access, vol. 8, pp. 118869–118883, 2020, doi: 10.1109/ACCESS.2020.3005510.
CHONG CHEN received the M.D. degree from the Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, in 2019. He is currently a Postdoctoral Fellow with the Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. His current research interests include chest CT imaging, coronary artery CT imaging, and computer-aided diagnosis.

RUI LI received the B.S. degree in measurement and control technology and instrument from Chongqing University, Chongqing, China, in 2013, and the M.S. degree in instrument science and technology from Beihang University, in 2016. She is currently an AI Algorithm Engineer at Shanghai United Imaging Intelligence Healthcare Company Ltd. Her research interest includes deep learning in medical image analysis.

HONG SHEN received the B.S. degree in electrical and electronics engineering from Tsinghua University, in 1994, and the Ph.D. degree in electrical and computer engineering from the Rensselaer Polytechnic Institute, in 2000. He is currently an VP of AI and the GM of Shanghai United Imaging Intelligence Healthcare Company Ltd. His research interests include multi-modality full-stack AI hardware and software solution covering the whole process of diagnosis and therapy.

LIMING XIA received the Ph.D. degree from the Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, in 2004. He is currently the Head of the Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. His current research interests include chest CT/MRI imaging, cardiac MRI imaging, breast MRI imaging, and fetus MRI imaging.