Predicting COVID-19 Disease Progression and Patient Outcomes based on Temporal Deep Learning

Chenxi Sun
Peking University
Shenda Hong
Peking University
Moxian Song
Peking University
Hongyan Li
Peking University
Zhenjie Wang (zhenjie.wang@pku.edu.cn)
Peking University https://orcid.org/0000-0003-1717-5790

Technical advance

Keywords: COVID-19, Disease progression, Outcome early prediction, Irregularly sampled time series, Time-aware long short-term memory

Posted Date: November 6th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-44308/v2

Version of Record: A version of this preprint was published on February 8th, 2021. See the published version at https://doi.org/10.1186/s12911-020-01359-9.

Predicting COVID-19 Disease Progression and Patient Outcomes based on Temporal Deep Learning

Chenxi Sun[1,2], Shenda Hong[3,4], Moxian Song[1,2], Hongyan Li[1,2], and Zhenjie Wang[5,*]

[1]School of Electronics Engineering and Computer Science, Peking University, Beijing, People's Republic of China.

[2]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, People's Republic of China.

[3]National Institute of Health Data Science, Peking University, Beijing, People's Republic of China.

[4]Institute of Medical Technology, Health Science Center of Peking University, Beijing, People's Republic of China.

[5]Institute of Population Research, Peking University, Beijing, People's Republic of China.

*Corresponding Author: Zhenjie Wang, Peking University, No.5 Yiheyuan Road, Beijing 100871, People's Republic of China. Email: zhenjie.wang@pku.edu.cn

Abstract

Background: The coronavirus disease 2019 (COVID-19) pandemic has caused health concerns worldwide since December 2019. From the beginning of infection, patients will progress through different symptom stages, such as fever, dyspnea or even death. Identifying disease progression and predicting patient outcome at an early stage helps target treatment and resource allocation. However, there is no clear COVID-19 stage definition, and few studies have addressed characterizing COVID-19 progression, making the need for this study evident.

Methods*:* We proposed a temporal deep learning method, based on a time-aware long short-term memory (T-LSTM) neural network and used an online open dataset, including blood samples of 485 patients from Wuhan, China, to train the model. Our method can grasp the dynamic relations in irregularly sampled time series, which is ignored by existing works. Specifically, our method predicted the outcome of COVID-19 patients by considering both the biomarkers and the irregular time intervals. Then, we used the patient representations, extracted from T-LSTM units, to subtype the patient stages and describe the disease progression of COVID-19.

Results: Using our method, the accuracy of the outcome of prediction results was more than 90% at 12 days and 98%, 95% and 93% at 3, 6, and 9 days, respectively. Most importantly, we found 4 stages of COVID-19 progression with different patient statuses and mortality risks. We ranked 40 biomarkers related to disease and gave the reference values of them for each stage. Top 5 is Lymph, LDH, hs-CRP, Indirect Bilirubin, Creatinine. Besides, we have found 3 complications - myocardial injury, liver function injury and renal function injury. Predicting

which of the 4 stages the patient is currently in can help doctors better assess and cure the patient.

Conclusions: To combat the COVID-19 epidemic, this paper aims to help clinicians better assess and treat infected patients, provide relevant researchers with potential disease progression patterns, and enable more effective use of medical resources. Our method predicted patient outcomes with high accuracy and identified a four-stage disease progression. We hope that the obtained results and patterns will aid in fighting the disease.

Keywords: COVID-19, Disease progression, Outcome early prediction, Irregularly sampled time series, Time-aware long short-term memory.

Background

Coronavirus disease 2019 (COVID-19) outbreaks have caused health concerns worldwide since December 2019; the disease was declared a pandemic by the World Health Organization (WHO) on 11 March 2020 [1]. Over seven million cases of COVID-19 have been reported worldwide, including more than 400,000 deaths (as of 15 June 2020) [2]. Even though the disease has been controlled in certain countries, the WHO director warns the pandemic is still 'Speeding Up' [25]. Because of its sudden onset, many hospitals are still facing medical resource shortages. For example, news in [26] reported a lack of medical resources in New Delhi. In [27], Arizona has experienced record-high hospital capacity as coronavirus cases climb. A reasonable allocation of resources according to patient condition is needed.

The solution to this problem involves determining the stages of disease progression by subtyping and predicting the outcome of COVID-19 patients. Then, targeted treatment and medical resource allocation can be carried out for patients in different stages. Recent studies [3-5,14-16] have used statistical methods to analyze COVID-19 progress by inpatient symptoms. However, different statistical results were obtained by considering different patient groups and different symptoms. At present, there is no clear division of the stages of COVID-19 progression.

Longitudinal disease analysis is the key to understanding disease progression, designing prognoses and developing early diagnostic tools. The time dynamics of disease can provide more information than static symptom observation [46]. Considering the complex patient states, the amount of interventions and the real-time requirement, the data-driven machine learning approaches by learning from electronic health records are the desiderata to help clinicians [44].

Many existing works have used machine learning methods for COVID-19 prediction tasks. We have summarized them in Table 1. For example, in most method of [45] and in [1,32,36-38,40,42], authors used non-deep learning methods, such as k-NN, LR, Cox, SVM and DT to classify CT/X-ray images and predict the outcomes of COVID-19 patients. However, in terms of prediction accuracy, non-deep learning is not as good as deep learning methods. Deep learning methods can train the parameters with complex nonlinearity to learn the data structures and have achieved state-of-the-art in many medical prediction tasks [19,20,47]. Thus, many current works apply deep learning methods for COVID-19 prediction tasks [28-36]. However, these methods either use the simple multi-layer perceptron for predicting or use the convolutional structures for image classification. Both the above methods ignored the temporal development of patient's status. In the real-world patient records, except for the basic information, vital signs, test values and diagnoses are both time series, especially for the blood samples of COVID-19 patients, the data we used in this paper.

Recently, a deep learning method, recurrent neural network (RNN) [18] can efficiently model temporal sequences. It uses recursion in the direction of sequence evolution to learning the relations among past, present and future. But the basic RNN has the long-term dependency problems [7]. Meanwhile, RNN only process uniformly distributed longitudinal data while COVID-19 patient blood samples are distributed nonuniformly with irregular time intervals between observations. Thus, a method that can model this irregular time series of COVID-19 patients is needed.

In this paper, we retrospectively analyzed the blood samples of 485 patients from the region of Wuhan, China. The medical records collected with standard case report forms, including

epidemiological, demographic, clinical, laboratory and mortality outcome information, from an online open dataset under an MIT license. We applied a temporal deep learning method Time-aware Long Short-term Unit (T-LSTM) to model the irregular time series of COVID-19 patients. T-LSTM can predict the mortality with more than 98% accuracy before 3 days. Meanwhile, we have discovered four stages of COVID-19 patients. According to the different stages, we gave the analysis of the patient's state and found the related biomarkers and complications.

Methods

In this section, we first introduce the COVID-19 dataset and the data preprocessing process. Then, we describe the methods for mortality prediction and disease progression in detail.

Dataset description

Blood index values can reflect a COVID-19 patient's physical condition [15]. COVID-19 patients' blood samples were collected between 10 January and 18 February 2020 at Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China [6]. The dataset contains 80 characteristics from 375 patients with 6120 records as a training set and 110 patients with 757 records as a test set. A case of sample is shown in Figure 1. It draws lines of the time series of LHD, lymph and hs-CRP of a 70-year-old female patient during hospitalization. We can see the time intervals between two observations are irregular, which could be a few minutes or even days.

The detailed statistical information of demographic and 74 clinical laboratory test features is listed Table 2. For example, in the dataset, the average age of patients is 58.83, the survival

rate is 53.6% and the ratio of male to female is about 1.5:1. We also list the range and mean value of each feature. In Figure 1, we display the distributions of some features (age, gender, LHD, lymph and hs-CRP) of survival class (0) and death class (1).

This COVID-19 blood test data is publicly available at https://github.com/HAIRLAB/Pre_Surv_COVID_19.

Dataset preprocessing

First, we attempted to find a suitable time measurement granularity. In the raw dataset, the lengths of sequences are unequal and different sampling times result in missing data, with an 85% missing rate (mr in Equation 1) on average. The presence of vacancies has a large impact on data quality, resulting in unstable predictions and other unpredictable effects [17]. We used 3 days as the basic sampling interval, reducing the average mr below 30%. The time series length of raw data, the average missing rate and the missing rate for each feature are shown in Figure 1.

$$mr = \frac{\#\ of\ records}{sequence\ length} \tag{1}$$

Meanwhile, for feature selection, using all 74 laboratory test features is unrealistic. To address the high missing rate, repeated features and collection difficulties, we considered three key features: lactic dehydrogenase (LDH), lymphocytes (lymph) and high-sensitivity C-reactive protein (hs-CRP). These features contain specific research biomarkers of COVID-19 patients [6] and can be easily collected in any hospital. Considering that only three features may not achieve high prediction accuracy, we also select 40 features (listed in Table 7) with missing rate less than 30% for comparative experiment.

T-LSTM

Recurrent neural networks (RNNs) [18] (the first structure in Figure 2) are deep network architectures designed to model temporal sequences. They take sequence data as input, recursion occurs in the direction of sequence evolution, and all units are chained together. In basic RNN (the second structure in Figure 2), the current state h_t is affected by the previous state h_{t-1} and the current input x_t and is described as $h_t = \sigma(Wx_t + Uh_{t-1} + b)$, where σ is an activation function, and W, U and b are learnable parameters. Long Short-Term Memory (LSTM) [7] (the third structure in Figure 2) is a variant of RNN that is adept at solving long-term dependency problems. A standard LSTM unit consists of a forget gate, an input gate, a memory cell and an output gate.

However, RNNs only process uniformly distributed longitudinal data by assuming that the sequences have an equal distribution of time differences. COVID-19 patient blood samples are distributed nonuniformly. For example, the time gap between two sequential records could be hours or days. Time-aware Long Short-Term Memory (T-LSTM) [8] (the fourth structure in Figure 2) incorporates the elapsed time information into LSTM. It applies a memory discount to capture the irregular temporal dynamics. T-LSTM can be formulated as:

(2)

$$C_{t-1}^{S} = \tanh(W_d C_{t-1} + b_d) \quad \text{Short-term memory}$$

$$\hat{C}_{t-1}^{S} = C_{t-1}^{S} * g(\Delta_t) \quad \text{Discounted short-term memory}$$

$$C_{t-1}^{T} = C_{t-1} - C_{t-1}^{S} \quad \text{Long-term memory}$$

$$C_{t-1}^{*} = C_{t-1}^{T} - \hat{C}_{t-1}^{S} \quad \text{Adjusted previous memory}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad \text{Forget gate}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad \text{Input gate}$$

$$\widetilde{C}_t = tanh(W_c x_t + U_c h_{t-1} + b_o) \quad \text{Candidate memory}$$

$$C_t = f_t * C_{t-1}^{*} + i_t * \widetilde{C}_t \quad \text{Current memory}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad \text{Output gate}$$

$$h_t = o_t * \tanh(C_t) \quad \text{Current hidden state}$$

Here, we use a log calculation for the elapsed time function. Δ_t describes the time gap between two records at two sequential time steps t and $\mathrm{t}-1$. T_t is the actual time at time step t.

$$g(\Delta_t) = \frac{1}{\log(e+\Delta_t)} \tag{3}$$

$$\Delta_t = T_t - T_{t-1}$$

Analysis strategy

We first describe the two tasks in this study and then introduce the specific methods. The whole method process is shown in Figure 3.

Task 1 (Outcome prediction) *A set of labeled patient data is represented as* $\mathcal{D} = \{(x_i, c_i) \in (X, C) | i = 1, \dots, n\}$. X *is a time series set of patients, where* $x_i = \{x_i^t | t = 1, \dots, t_{onset}\}$ *represents a patient's records over* t *time steps; specifically,* x_i^t *is multivariate data, and each dimension is a clinical record represented by a numeric vector.* $C \in \{0,1\}$ *is the outcome, where class* 0 *means death and class* 1 *means survival. The outcome prediction task aims to*

predict patient outcomes by the prediction function $f: X \rightarrow C$.

Task 2 (Temporal patient subtyping / Disease progression mining) *The goal is to find patient groups* $G = \{g_i | i = 0, \dots, m\}$ *with similar feature representation* $R = \{r_i^t | i = 0, \dots, n; t = 0, \dots, t_{onset}\}$. r_i^t *is the representation of clinical record* x_i^t *at time* t. *Then, the patient groups* G *distributed over time are used to analyze the stages of disease progression*.

In COVID-19 patient outcome prediction task, T-LSTM is used to handle patient record sequences and then make the prediction. The process is displayed in the proposed method of Figure 2, in the lower gray area.

For a patient i, the input of T-LSTM at time step t is a three-dimensional feature vector $x_i^t = \left[v_{LDH}, v_{lymphocytes}, v_{hs-CRP},\right]$ with time gap Δ_t. The output is the state representation s_i at the last time step. We apply this outcome prediction task as a binary classification task, with two classes: death and survival.

The cross-entropy [43] is mainly used to measure the difference between two probability distributions. We expect our predicted distribution of patient outcomes to be closer to the true distribution. Thus, we use the cross-entropy loss function in Equation 4. Besides, when using sigmoid active function, this loss can avoid the reduced learning rate causing by traditional mean square error loss when gradient decreases.

$$L = L_{CE}\left(C, \hat{C}\right) = -\sum_x p(x) log q(x) = -\sum_{i=1}^{n} \widehat{c_i} log c_i + (1 - \widehat{c_i}) log(1 - c_i) \quad (4)$$

$p(x)$ is the prior probability (true label vector) and $q(x)$ is the prediction probability (predicted results vector). Correspondingly, $\hat{C}$ is the real class of input data, and C represents the prediction class.

In COVID-19 patient disease progression task, temporal patient subtyping can uncover the

dynamic characteristics of diseases by significantly nuanced subtyping, which leads to the potential stages of disease progression. We addressed the issue by building a time stage reference and providing a low-dimensional representation of each subject, encoding his or her position with respect to this reference.

The method structure is displayed in the upper gray area of proposed method in Figure 2. It has 4 steps: 1) Acquisition of patient representation r^t. We used the hidden state h_t, extracted from every T-LSTM unit, as the patient's representation r^t at time step t. 2) Dimension reduction of r^t. For better demonstration, we used the t-distributed Stochastic Neighbor Embedding (t-SNE) [9] method to reduce these high-dimensional vectors r^t into two dimensions. 3) Obtaining the patient group set G. As prior information about the patient groups was not available, we acquired patient groups by applying an unsupervised clustering method, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [10], on r^t. 4) Analysis of G and stages of disease progression. The mortality rate, MR, and the average time distance, TD, were calculated as the analysis criteria:

$$MR = \frac{\#\ of\ death}{\#\ of\ patients} \tag{5}$$

$$TD = \frac{1}{|g_k|}\sum_{x_i^t \in g_k}(T_{t_{onset}} - T_t) \tag{6}$$

Evaluation metrics

The prediction results were evaluated by assessing the area under the curve of the Receiver Operating Characteristic (AUC-ROC). The ROC is a curve of the True Positive Rate (TPR) and the False Positive Rate (FPR). TN, TP, FP and FN represent true positives, true negatives, false positives and false negatives, respectively.

$$TPR = \frac{TP}{TP+FN} \quad (7)$$

$$FPR = \frac{FP}{TN+FP} \quad (8)$$

The patient groups obtained by unsupervised clustering were evaluated by the Calinski-Harabaz Index (CH), which measures the covariance of data within a class and between classes. A larger CH value indicates a better clustering performance. In Equation 9, m is the number of data and k is the number of groups. B_k and W_k respectively represent the covariance matrices between groups and within groups.

$$CH = \frac{tr(B_k)}{tr(W_k)} \frac{m-k}{k-1} \quad (9)$$

When we get the stages of COVID-19 patients, we used Kullback-Leibler Divergence (KL divergence) to analyze patient characteristics through each laboratory test feature. KL divergence can measure the degree of difference between two probability distributions. For each feature, we first establish the Gaussian distribution at each stage. Then, we calculate the average KL divergence of the distribution of adjacent stages. If the average KL divergence of a feature is large, it more likely is a biomarker to distinguish different stages. The basic KL divergence of distribution $p(X)$ and $q(X)$ and the KL divergence of two univariate Gaussian distributions are in Equation 10 and 11.

$$KL(p(X)||q(X)) = \sum_{x_i \epsilon X} p(x_i) log \frac{p(x_i)}{q(x_i)} \quad (10)$$

$$KL(\mathcal{N}(\mu_1, \sigma_1^2)||\mathcal{N}(\mu_2, \sigma_2^2)) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \quad (11)$$

For measure and evaluate each feature, we use the average KL divergence (Average KL) between neighbor stages g_i, g_{i+1}.

$$Average\ KL = \frac{1}{m} \sum_{i=0}^{m-1} KL_{g_i, g_{i+1}} \quad (12)$$

Results

We used the records of 375 patients as a training set; the ratio of the training set to the verification set was 0.8:0.2. The records of 110 patients made up the test set. This experiment was conducted on 5-fold cross-validation. The code implementation is publicly available at https://github.com/scxhhh/COVID-19.

Baselines

We use the related works summarized in Table 1 as comparison methods. Related works are divided into non-deep learning methods and deep learning methods. We use Cox [32], k-NN [42], SVM [36], DT [1], BPNN [28], PNN [29], RNN, LSTM and T-LSTM for COVID-19 mortality prediction. T-LSTM is our method.

Outcome prediction results.

Table 3 shows the results of COVID-19 mortality prediction using baselines. The AUC-ROC is evaluated at 0, 3, 6, 9, 12, 15, and 18 days early. Here, the results are obtained when the patient's representations are 64 dimensional. The results indicate that our method T-LSTM performed better than all of baselines no matter how early before the onset times of patients. More precisely, using T-LSTM, the outcome prediction accuracy is above 90% at 12 days early and is approximately 97% accurate when predicting 3 days before the disease outcome. More detailed results of train, validation and test sets using T-LSTM are listed in Table 4.

The first four figures in Figure 3 are the visualizes of prediction results. The first two figures are the AUC-ROC of prediction results of baselines and T-LSTM in different earliness.

The third figure is the changes of prediction accuracy and cross-entropy loss when training the model. The fourth figure represents the relation of patient representation dimension and AUC-ROC of prediction using T-LSTM. Too few dimensions lead to incomplete feature learning, while too many dimensions lead to redundant calculations and easy over-fitting. Considering result accuracy, computational complexity and ease of representation use in the following task, we decided to use 64 dimensional vectors to represent patients.

Further, we also select 40 features (listed in Table 7) as the input of T-LSTM for comparative experiment. The results in Table 5 indicate that learning a large number of patient characteristics does not necessarily contribute to COVID-19 patient mortality prediction task. The three biomarkers, LDH, lymph and hs-CRP can make the results better. The AUC-ROC of using 3 features is 3% higher than using 40 features on average.

Disease progression results.

By implementing the four steps of disease progression mining, we obtained the 4 stages in both the death class (critical) and the survival class (general), shown in Figure 4.

For better visualization, we reduced the dimension of the patient's representation vector, the fifth figure in Figure 3 is the dimension reduction effect. We chose 2 dimensions due to low representation loss and clear observation. Besides, the DBSCAN clustering effect evaluated by the CH index is shown in the sixth and seventh figures in Figure 3. Different clustering effects can be obtained by changing the cluster radius parameter ε. The best CH index values for the death class and the survival class are 680.07 and 44.24, respectively.

In this case, both classes have four groups. Four stages of COVID-19 patients are shown

in Figure 4. For each stage, we calculate the mortality rate MR and the average time distance TD. For the death class, MR increases over stages and is 100% at stage 4. For the survival class, MR decreases over stages and is 0% in stage 4. TD in both classes decreases, meaning that the 4 stages are distributed over time. Meanwhile, as the CH index of the survival class is higher than that of the death class, the survival class stages are relatively loosely distributed.

In Figure 4, the first clustering is obtained by using biomarkers directly and shows that reasonable stages could not be found. In the first clustering, no stage is clustered in the death class and the 2 stages in the survival class have similar mortality rates and no time difference, as the shade of blue indicates. However, using our method, different stages have obvious differences, such as the data point color deepening with the stages. Meanwhile, as shown in the two insets, the class boundary is clearer based on our method.

Meanwhile, we calculated the mean values of 40 laboratory test features in each stage, the feature values vary with stages. Table 6 lists 10 of these features - Lymph, LDH, hs-CRP, Indirect Bilirubin, Creatinine, INR, Serum Sodium, eGFR, Serum Chlorine and Albumin. The changes of values through 4 stages are visualized in Figure 5. Under different classes, the trends of features are different.

Further, we calculated the average KL divergence between adjoint stages of each features in 40 clinical laboratory tests data. We ranked the average KL values. The higher the ranking, the better the biomarkers can be used to distinguish different stages. The top five are: Lymph, LDH, hs-CRP, Indirect Bilirubin and Creatinine. Among them, Lymph is the best biomarker to distinguish stages.

Discussion

In recent years, deep learning (DL) technology has been widely used because of its superior performance in various medical applications [19, 20], such as medical image recognition [21] and medication recommendations [22]. In the past year, the spread of COVID-19 has had a peripheral effect on the global economy and health. Therefore, we expect to combine DL methods to study and fight COVID-19.

The states of COVID-19 patients in hospital are dynamic time sequence processes. In addition to the basic information of patients, the vital signs, diagnoses and other lab tests are all time series. Existing many works [28-42,45] have achieved good results for COVID-19 prediction tasks. But they paid little attention to analyze and model the characteristics of COVID-19 patients' time series. Dynamic time series modeling can grasp the relationship between historical observations and current observations, and learn the potential development mode of sequence, which is conducive to more accurate prediction and representation. Besides, we have found that the time series of COVID-19 patients is irregularly sampled - Different time intervals exist in adjacent observations. Every possible test is not regularly measured during an admission. When a certain symptom worsens, corresponding variables are examined more frequently; when the symptom disappears, the corresponding variables are no longer examined. These time intervals will add a time sparsity factor when the intervals between observations are large [44]. Therefore, it is necessary not only to deal with time series, but also to deal with irregular time series according to the characteristics of COVID-19 patients.

This study has two basic contributions. First, we can predict patient outcomes with higher accuracy than baselines. The method can effectively predict whether the infected patient will

die or survive 12 days prior to disease outcome with over 90% accuracy. The prediction accuracies at 3, 6, and 9 days prior are 98%, 95% and 93%, respectively. Second, we identified four stages of COVID-19 progression. The stages are closely related to mortality and time of illness and can help analyze the status of infected patients. Through the analysis of each stage, we ranked 40 biomarkers according to the degree of correlation with COVID-19.

Currently, certain studies have identified suitable predictive biomarkers, such as the 3 biomarkers in [6], which are regarded to have a significant impact on patient mortality. If a doctor predicts survival or death only by observing the biomarkers and using a threshold, the accuracy is at or below 80% for early predictions. However, the clinical reference value of inaccurate results is very low [11, 12]. Finding the appropriate prediction-related biomarkers is important, but it is equally important to create an appropriate and high-accuracy prediction method. The DL method has better performance and the time-aware aspect enables higher accuracy, as shown in Table 3. Decision trees [6] are easy to explain, but do not have good accuracy. T-LSTM can better grasp the relationship between time than basic LSTM. It not only improves the prediction accuracy but also contributes to the identification of disease stages, as the disease progression is distributed over time.

However, there are some concerns about the use of DL methods in the high-risk tasks of healthcare. First, we recognize that it may be risky to apply predictive methods directly to clinical practice [13]. DL methods may be assistive tools for doctors but not used to make decisions directly. It is challenging for doctors to make optimal decisions; a data-driven, high-accuracy prediction method could help. Second, the DL method is troubled by poor interpretability [23], and clinical settings prefer interpretable models. Our approach has this

flaw, as do other black-box models. We feel it is best to keep a complex black-box model to gain 3 percent higher accuracy, compared with interpretable methods [6]. We also wish to provide interpretable results to help doctors better understand the model's result. Thus, we ranked 40 biomarkers according to the degree of correlation with COVID-19 in disease progression task.

To help relevant researchers better study COVID-19 patients, we have uncovered the disease progression of COVID-19 patients and obtained 4 stages. This interesting finding cannot be distinguished simply by the value of biomarkers, as shown as the comparison of two clustering results in Figure 4. Our method found stages closely related to mortality and time of illness. This shows that the DL method can explore new patterns in multidimensional space that cannot be demonstrated by a simple variable value [24].

The division of stages contains the potential characteristics of COVID-19. Here, we present three findings. First, at the time of initial diagnosis, the COVID-19 infected patients' physical conditions are similar, regardless of final survival or death. In Figure 4, the distance between stage 1 for the death class and the survival class is small, and the two even overlap. This indicates that outcome prediction is likely not accurate at the time of infection. Second, the physical condition of patients who eventually die changes less than that of those who eventually survive. We conclude this from CH index values, where the CH value of the survival class is larger than that for the death class. Third, mortality rate varies by stage. For example, if the patient is classified into the death class and is at stage 1, there is still hope of survival, as shown by the green triangle sample in Figure 4. However, if the patient is in stage 3 or 4, he or she is very likely to die. Based on estimating the current stage of a patient, doctors will be given a

reference, which can help them assess a patient's current situation. Based on that, doctors can carry out targeted treatment and reasonable resource allocation more easily. Thus, the ultimate goal of this study, helping improve the quality of medical care, can be achieved.

Meanwhile, by ranking 40 biomarkers according to the degree of correlation with COVID-19 (Table 7), we have found the biomarkers which are more relevant to COVID-19. The top 10 are: Lymph, LDH, hs-CRP, Indirect Bilirubin, Creatinine, INR, Serum Sodium, eGFR, Serum Chlorine and Albumin. For each marker, we gave its reference value in each stage, shown in Table 6. Different markers have unique trends in different stages. Combining the correlation analysis with the reference value analysis, we found that the critical COVID-19 patients are usually accompanied by low values of lymph, eGFR, albumin and Serum Sodium, high values of LDH, hs-CRP, indirect bilirubin, creatinine and INR. For example, in the critical stage 4, the average lymph (%) is just 4 and the average LDH (U/l) is up to 499. Besides, there are three major complications of COVID-19 patients - myocardial injury, liver function injury and renal function injury. We got the conclusions separately through the value of 1) LDH, 2) albumin and indirect bilirubin, 3) serum sodium, serum chlorine and creatinine in different stages.

Further, there is room for further improvement. First, because of the data limitations, our method may face risk of bias, because data-driven methods are easily influenced by different source of data. For example, the results may vary when using different datasets [13]. Second, we hope to design a more interpretable model, rather than using a complex DL black-box model. Meanwhile, we hope to enlighten the relevant researchers to further study these 4 stages and present more clinical explanations. In particular, we expect to be able to give specific treatments for different stages. Targeted treatment is significant for both patient rehabilitation and the

reasonable allocation of medical resources.

Conclusions

The sudden outbreak and epidemic of COVID-19 has led to worldwide suffering and shortages of medical resources. In this paper, we propose T-LSTM to predict patient outcomes with high accuracy - 98%, 95% and 93% at 3, 6, and 9 days, which will enable reasonable allocation of medical resources. T-LSTM can effectively model the irregular sampled time series in blood test samples of COVID-19 patients and predict more accurately than existing baselines. Meanwhile, we identified four COVID-19 stages. We ranked biomarkers according to correlations to the outcomes of patients, gave the reference values of top 10 biomarkers for each stage and have found 3 complications. The top 10 biomarkers are: Lymph, LDH, hs-CRP, Indirect Bilirubin, Creatinine, INR, Serum Sodium, eGFR, Serum Chlorine and Albumin. The complications are myocardial injury, liver function injury and renal function injury. By analyzing patients' life conditions at different stages, doctors can choose specific, targeted treatments. Future work will focus more on the treatments in different stages. Meanwhile, more real clinical data are expected to be available for model validation.

Abbreviations

COVID-19 (Corona Virus Disease 2019)

WHO (World Health Organization)

LDH (Lactic Dehydrogenase)

hs-CRP (High-sensitivity C-reactive Protein)

lymph (Lymphocytes)

DL (Deep Learning)

RNN (Recurrent Neural Network)

LSTM (Long Short-Term Unit)

T-LSTM (Time-aware Long Short-Term Memory)

PNN (Probabilistic Neural Network)

RBFNN (Radial Basis Function Neural Network)

GRNN (Generalized Regression Neural Network)

BPNN (Back Propagation Neuron Network)

DT Decision Tree)

RF (Random Forest)

XGBoost (eXtreme Gradient Boosting)

SVM (Support Vector Machines)

Cox (Cox's Proportional Hazards Regression)

LR (Linear Regression)

NB (Naive Bayes)

LDA (Linear Discriminant Analysis)

t-SNE (t-distributed Stochastic Neighbor Embedding)

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

MR (Mortality Rate)

TD (average Time Distance)

AUC-ROC (the Area Under the Curve of the Receiver Operating Characteristic)

CH (Calinski-Harabaz Index)

KL divergence (Kullback-Leibler Divergence)

Declarations

Ethics approval and consent to participate

The data is under an MIT license (https://doi.org/10.5281/zenodo.3758806).

Consent for publication

Not applicable.

Availability of data and material

The code implementation is publicly available at https://github.com/scxhhh/COVID-19. The data is at https://github.com/HAIRLAB/Pre_Surv_COVID_19.

Competing interests

No financial competing interests

Funding

This work was supported by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry and UKRI's Global Challenge Research Fund (ES/P011055/1). The funders had no role in study design, data collection, analysis, the writing

of the manuscript, or the decision to submit this article for publication.

Authors' contributions

C.S. and S.H. conceptualized the idea. S.H., Z.W. and H.L. initialized, conceived and supervised the project. C.S, S.H. and M.S. collected data and implemented the experiments. C.S, S.H, M.S. and Z.W. drafted the manuscript. All authors provided a critical review of the manuscript and approved the final draft for publication.

Acknowledgments

This paper is dedicated to those who want to fight COVID-19.

References

1. World Health Organization. Coronavirus Disease 2019 (COVID-19) Situation Report 68, 28 March 2020 (2020); https://www.who.int/docs/default-source/ coronaviruse/situation-reports/20200328-sitrep-68-covid-19.pdf
2. World He alth Organization. Coronavirus Disease 2019 (COVID-19) Situation Report 147, 15 June 2020 (2020); https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200615-covid-19-sitrep-147.pdf?sfvrsn=2497a605_2
3. Huang, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395, 497–506 (2020).
4. 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).

5. 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 Resp. Med. 8, 475–481 (2020).

6. Yan L, Zhang H T, Goncalves J, et al. An interpretable mortality prediction model for COVID-19 patients[J]. Nature, Machine intelligence, VOL 2, 2020.

7. S Hochreiter and J Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735– 1780.

8. Baytas I M, Xiao C, Zhang X, et al. Patient Subtyping via Time-Aware LSTM Networks[C] the 23rd ACM SIGKDD International Conference. ACM, 2017.

9. Laurens V D M, Hinton G. Visualizing Data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(2605):2579-2605.

10. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD 1996: 226-231.

11. Zhibo Wang, Zhezhi He, Milan Shah, Teng Zhang, Deliang Fan, Wei Zhang:Network-based multi-task learning models for biomarker selection and cancer outcome prediction. Bioinform. 36(6): 1814-1822 (2020).

12. Luchen Liu, Haoran Li, Zhiting Hu, Haoran Shi, Zichang Wang, Jian Tang, Ming Zhang. Learning Hierarchical Representations of Electronic Health Records for Clinical Outcome Prediction. AMIA 2019.

13. Wynants L, Calster B V, Bonten M M J, et al. Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal[J]. BMJ (online), 2020, 369:m1328.

14. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study[J]. The Lancet, 2020, 395(10229).

15. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China[J]. Jama the Journal of the American Medical Association, 2020.

16. Yang, X., Yu, Y., Xu, J., Shu, H., Xia, J., Liu, H., et al. (2020) Clinical Course and Outcomes of Critically Ill Patients with SARS-CoV-2 Pneumonia in Wuhan, China: A Single-Centered, Retrospective, Observational Study. The Lancet Respiratory Medicine, 1-7.

17. X.Chu, I.F.Ilyas, S.Krishnan, and J.Wang, Data cleaning: Overview and emerging challenges, in Proc. Int. Conf. Manage. Data SIGMOD Conf., San Francisco, CA, USA, Jun./Jul. 2016, pp. 2201–2206.

18. Williams R J, Zipser D. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks[J]. Neural Computation, 1998, 1(2)

19. Adam, G., Rampášek, L., Safikhani, Z. et al. Machine learning approaches to drug response prediction: challenges and recent progress. npj Precis. Onc. 4, 19 (2020).

20. Jalali, A., Lonsdale, H., Do, N. et al. Deep Learning for Improved Risk Prediction in Surgical Outcomes. Sci Rep 10, 9289 (2020).

21. Shaoqiang Wang, Shudong Wang, Song Zhang, Fangfang Fan, Gewen He. Research on Recognition of Medical Image Detection Based on Neural Network. IEEE Access 8: 94947-94955 (2020)

22. Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, Jimeng Sun. GAMENet: Graph Augmented MEmory Networks for Recommending Medication Combination. AAAI 2019: 1126-1133.

23. Christoph Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.

24. Gibney H M J. Analysis of meal patterns with the use of supervised data mining techniques—artificial neural networks and decision trees[J]. The American Journal of Clinical Nutrition, 2008, 88(6): p.1632-1642.

25. Emily Czachor. WHO Director Warns COVID-19 Pandemic is 'Speeding Up,' Here for 'Long Haul' [N]. Newsweek, News. 6/29/2020. https://www.newsweek.com/who-director-warns-covid-19-pandemic-speeding-here-long-haul-1514169.

26. Sébastien Farcis. Coronavirus: worries and worries about bed shortage in New Delhi[N]. Liberation, Reportage. 6/15/2020. https://www.liberation.fr/planete/2020/06/15/coronavirus-a-new-delhi-inquietude-et-desarroi-face-a-la-penurie-de-lits_1791300.

27. Katherine Fung. Arizona Hits Record-High Hospital Capacity As Coronavirus Cases Climb[N]. Newsweek, News. 6/29/2020. https://www.newsweek.com/arizona-hits-record-high-hospital-capacity-coronavirus-cases-climb-1511578.

28. Sujath R, Chatterjee J M, Hassanien A E. A machine learning forecasting model for COVID-19 pandemic in India[J]. Stochastic Environmental Research and Risk Assessment, 2020(1).

29. S Dhamodharavadhani, R Rathipriya and Jyotir Moy Chatterjee. COVID-19 Mortality Rate Prediction for India Using Statistical Neural Network Models. Frontiers in Public Health. 2020, 8(441).

30. Panwar H, Gupta P K , Siddiqui M K , et al. Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet[J]. Chaos Solitons & Fractals, 2020, 138:109944.

31. Harsh Panwar, P.K. Gupta, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Prakhar Bhardwaj, Vaishnavi Singh. A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images[J]. Chaos Solitons & Fractals, 2020, 140: 110190.

32. Liang W , Yao J , Chen A , et al. Early triage of critically ill COVID-19 patients using deep learning[J]. Nature Communications. 2020.

33. R. G. Babukarthik, V. Ananth Krishna Adiga, G. Sambasivam, D. Chandramohan, J. Amudhavel: Prediction of COVID-19 Using Genetic Deep Learning Convolutional Neural Network (GDCNN). IEEE Access, 8: 177647-177666. 2020

34. Wang L, Wong A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images. arXiv:2003.09871. 2020.

35. Li L et al (2020) Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology, pp 200905, 2020.

36. de Moraes Batista AF et al. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. medRxiv. 2020.

37. Li K et al. The clinical and chest CT features associated with severe and critical COVID-19 pneumonia. Investigative radiology, 2020.

38. Li C et al. Preliminary study to identify severe from moderate cases of COVID-19 using NLR&RDW-SD combination parameter. medRxiv, 2020.

39. Tang Z et al Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images. arXiv:2003.11988.

40. Farid AA, Selim GI, Khater HAA. A novel approach of CT images feature analysis and prediction to screen for corona virus disease (COVID-19). Int J Sci Eng Res 11(3):1–9, 2020.

41. Youssoufa Mohamadou, Aminou Halidou, Pascalin Tiam Kapen: A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19. Appl. Intell. 50(11): 3913-3925 (2020).

42. Kumar R (2020) Accurate prediction of COVID-19 using chest x- ray images through deep feature learning model with smote and machine learning classifiers. In: medRxiv

43. Corduneanu, C. Integral Equations and Applications[J]. 1991, 10.1017/CBO9780511569395.

44. Chenxi Sun, Shenda Hong, Moxian Song and Hongyan Li. A Review of Deep Learning Methods for Irregularly Sampled Medical Time Series Data. arXiv:2010.12493. 2020.

45. Radanliev P, Roure D D, Walton R . Data mining and analysis of scientific research data records on covid 19 mortality, immunity, and vaccine development - In the first wave of the Covid-19 pandemic[J]. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 2020.

46. Gabriel Spadon, Shenda Hong, Bruno Brandoli, Stan Matwin, Jose F. Rodrigues-Jr, Jimeng Sun. "Pay Attention to Evolution: Time Series Forecasting with Deep Graph-Evolution Learning." arXiv preprint arXiv:2008.12833 (2020).

47. Siuly, S., Zhang, Y. Medical Big Data: Neurological Diseases Diagnosis Through Medical Data Analysis. Data Sci. Eng. 1, 54–64 (2016).

Figure Legends

Figure 1 Examples and statistics of COVID-19 dataset

The first block is a line chart of an example in dataset - a 70-year-old female patient. It draws the time series of LHD, lymph and hs-CRP during hospitalization; The second block is the distributions of age, gender, LHD, lymph and hs-CRP of survival class (0) and death class (1); The third block is the statistics about dataset. It contains the counts of time series length, the statistics of overall missing rate and the statistics of each feature’s missing rate under different sampling rate.

Figure 2 Structures of the methods

The first block shows the structure of RNNs, including basic RNN, LSTM and our T-LSTM; The second block shows the structure that how to use T-LSTM to complete the outcome prediction task (lower grey area) and disease progressing task (upper grey area).

Figure 3 The results of outcome prediction

The first line's charts are the AUC-ROC of mortality prediction results using baselines; The second line's chart is the changes of accuracy and loss during training T-LSTM; The third line's charts are the dimension experiments. They show the accuracy of mortality prediction by using different representation dimensions and the effect of representation dimension reduction; The fourth line's charts are the effect when using DBSCAN.

Figure 4 The result of COVID-19 progression

This figure shows the four stages of COVID-19 patients by using T-LSTM. The upper clusters are the original clustering of data. The lower are the patient subtyping by using T-LSTM. We can find there are four clusters with distinct boundaries both in death/critical class (red) and survival/general class (blue).

Figure 5 Changes of features in different stages

This figure shows the changes of features (Mortality rate, Lymph, LDH, hs-CRP, Indirect Bilirubin, Creatinine, INR, Serum Sodium, eGFR, Serum Chlorine and Albumin) through 4 stages. Under different classes, the trends of features are different.

Figures

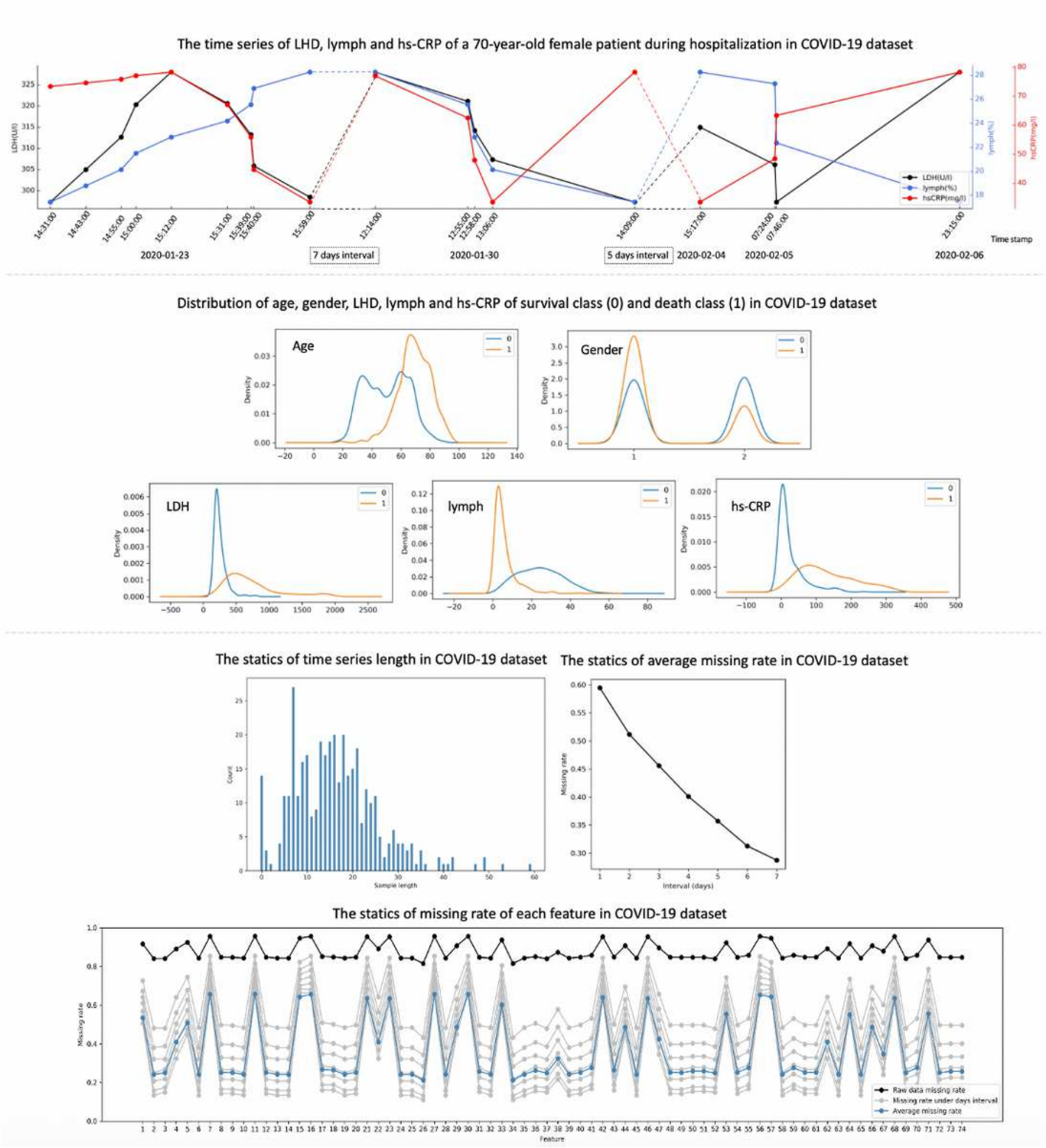

Figure 1

Examples and statistics of COVID-19 dataset The first block is a line chart of an example in dataset - a 70-year-old female patient. It draws the time series of LHD, lymph and hs-CRP during hospitalization; The second block is the distributions of age, gender, LHD, lymph and hs-CRP of survival class (0) and death

class (1); The third block is the statistics about dataset. It contains the counts of time series length, the statistics of overall missing rate and the statistics of each feature's missing rate under different sampling rate.

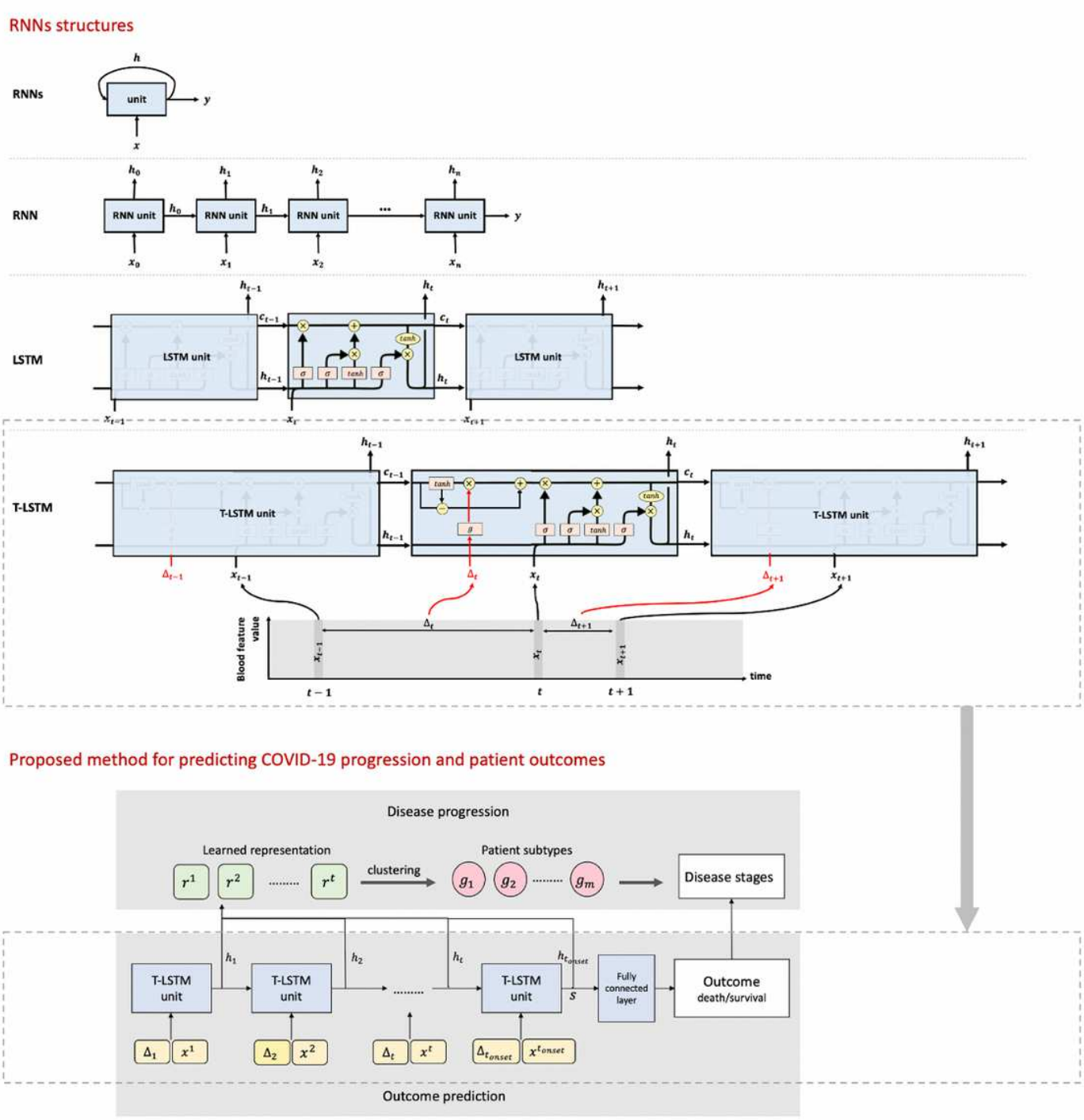

Figure 2

Structures of the methods The first block shows the structure of RNNs, including basic RNN, LSTM and our T-LSTM; The second block shows the structure that how to use T-LSTM to complete the outcome prediction task (lower grey area) and disease progressing task (upper grey area).

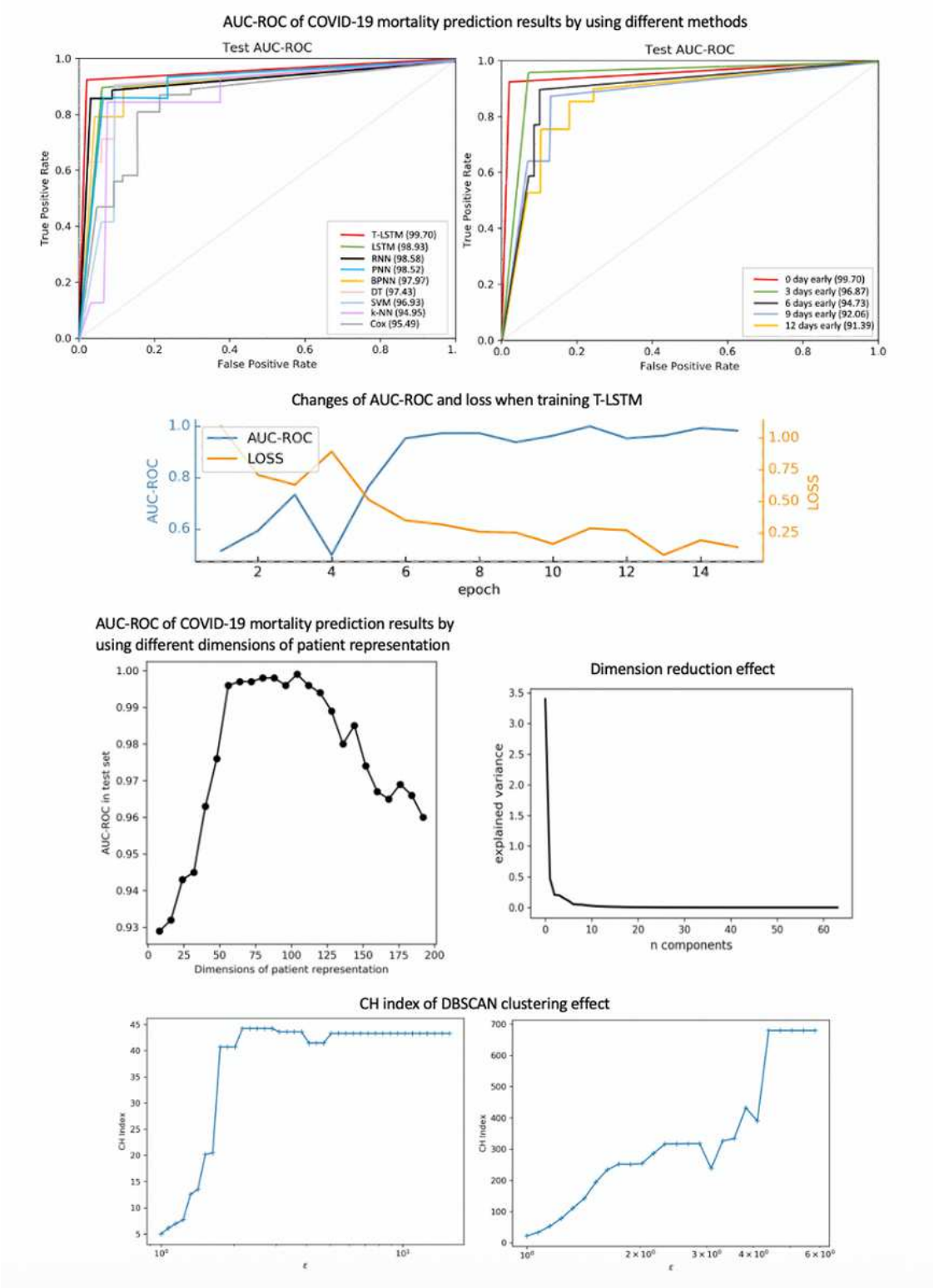

Figure 3

The results of outcome prediction The first line's charts are the AUC-ROC of mortality prediction results using baselines; The second line's chart is the changes of accuracy and loss during training T-LSTM; The third line's charts are the dimension experiments. They show the accuracy of mortality prediction by using different representation dimensions and the effect of representation dimension reduction; The fourth line's charts are the effect when using DBSCAN.

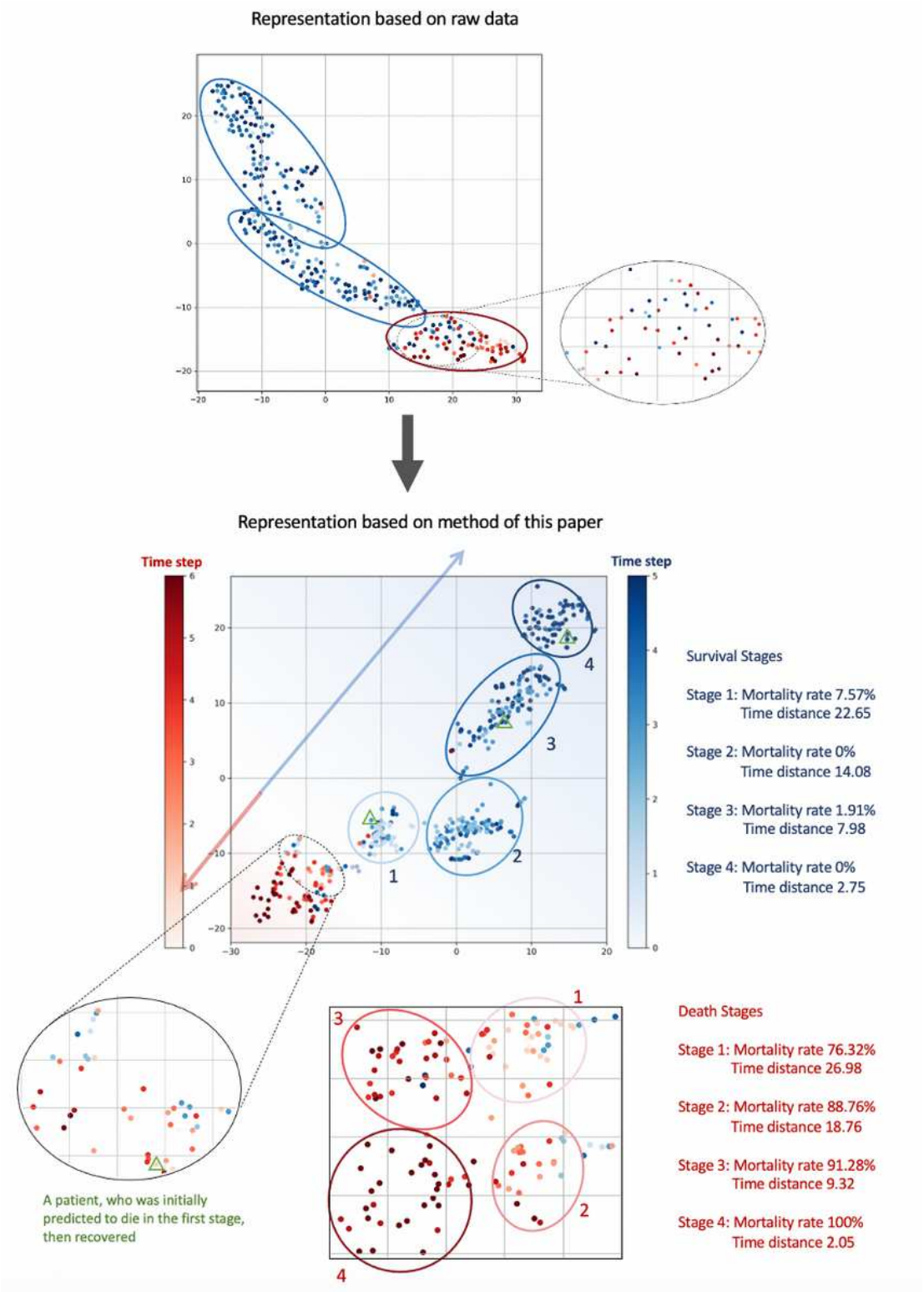

Figure 4

The result of COVID-19 progression This figure shows the four stages of COVID-19 patients by using T-LSTM. The upper clusters are the original clustering of data. The lower are the patient subtyping by using T-LSTM. We can find there are four clusters with distinct boundaries both in death/critical class (red) and survival/general class (blue).

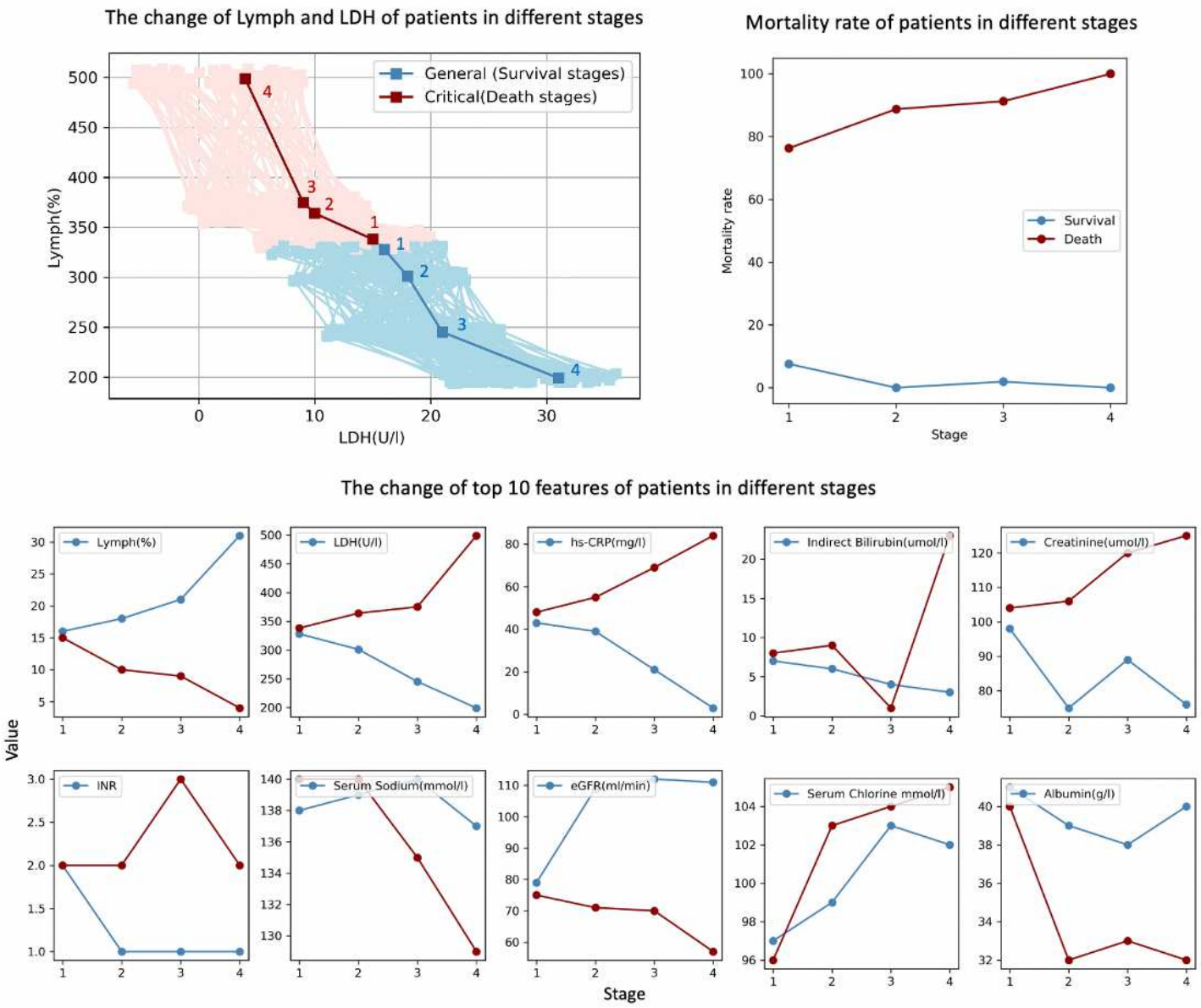

Figure 5

Changes of features in different stages This figure shows the changes of features (Mortality rate, Lymph, LDH, hs-CRP, Indirect Bilirubin, Creatinine, INR, Serum Sodium, eGFR, Serum Chlorine and Albumin) through 4 stages. Under different classes, the trends of features are different.