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

Technical advance

Keywords: COVID-19, Disease progression mining, Outcome early prediction, Deep learning, Long short-term memory

Posted Date: August 24th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-44308/v1

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.

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. Specifically, our method predicted the outcome of COVID-19 patients by T-LSTM, 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 risks, such as mortality and length of illness. 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.

Full Text

This preprint is available for download as a PDF.

Figures

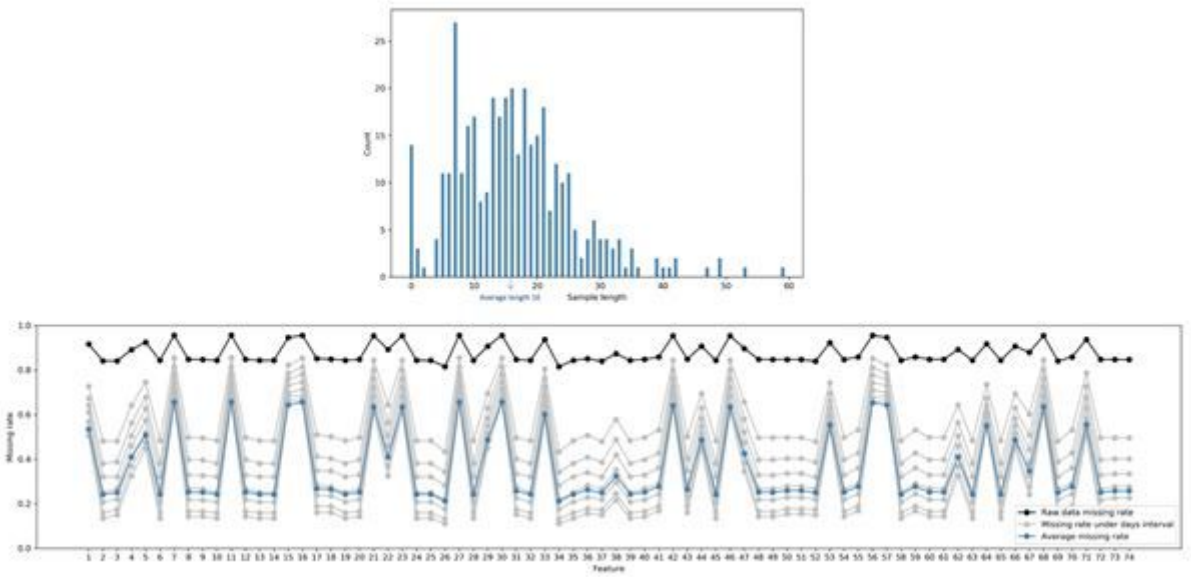

Figure 1

Lengths of record sequences and the missing rate of different features in the dataset for different sampling intervals The first graph shows the number of different data lengths in the dataset. Most of the data have lengths below 30; the average length is 16. The second graph shows the missing rate of different features of the dataset for different sampling intervals. The black line is the missing rate of different features in the raw data, above 80% on average. The gray line is the missing rate using 1-7 days granularity. The blue line is the average missing rate of granularities. We use 3 days as the basic sampling granularity, for which the average missing rate is below 30%.

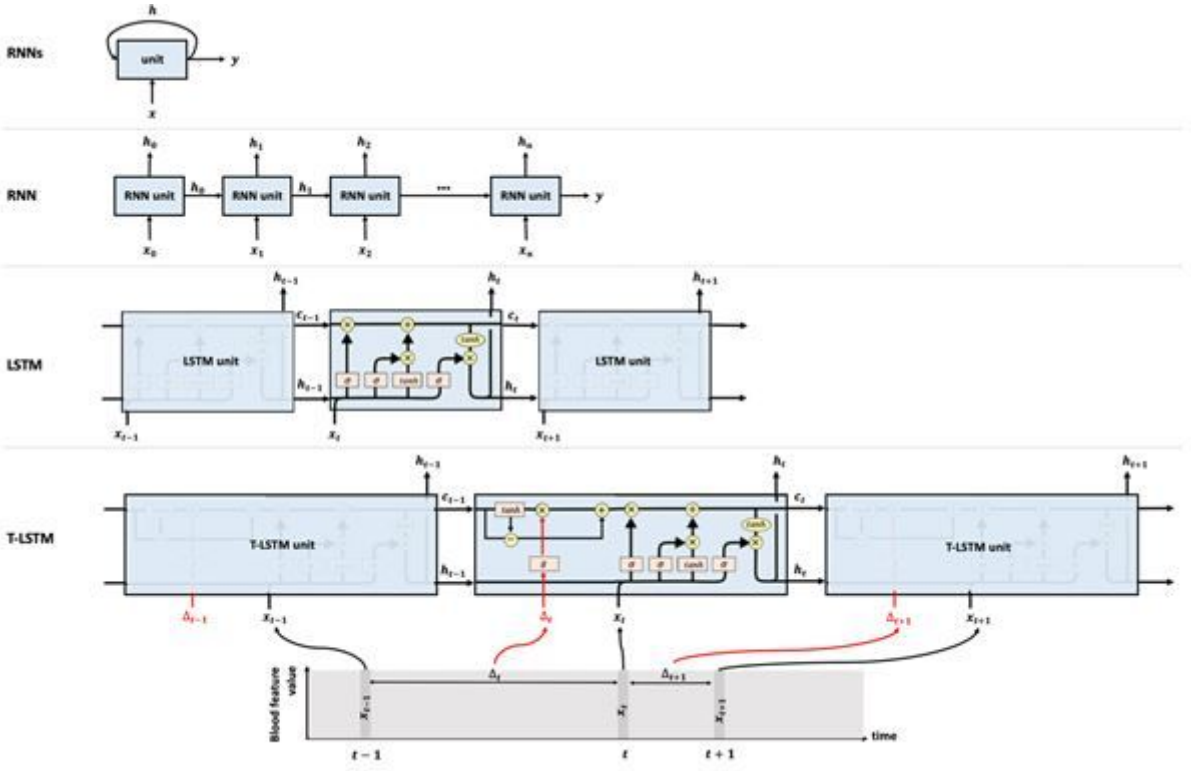

Figure 2

Recurrent Neural Networks (RNNs) structures RNNs: Recurrent Neural Networks (RNNs) are a general description of the following structures. RNN: It is the basic structure in RNNs. 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=σ(Wx_t+Uh_(t-1)+b). LSTM: Long Short-Term Memory is a variant of RNN. It uses calculations, shown in the figure as yellow icons, to form the forget gate, the input gate, the memory cell and the output gate. The formal descriptions of the different gates are in Equation 2. T-LSTM: Time-aware LSTM is the model used in this paper. It applies the memory discount, described in Equation 2, to incorporate the elapsed time information into LSTM. Its inputs are data vectors and time intervals, as shown on the gray time axis.

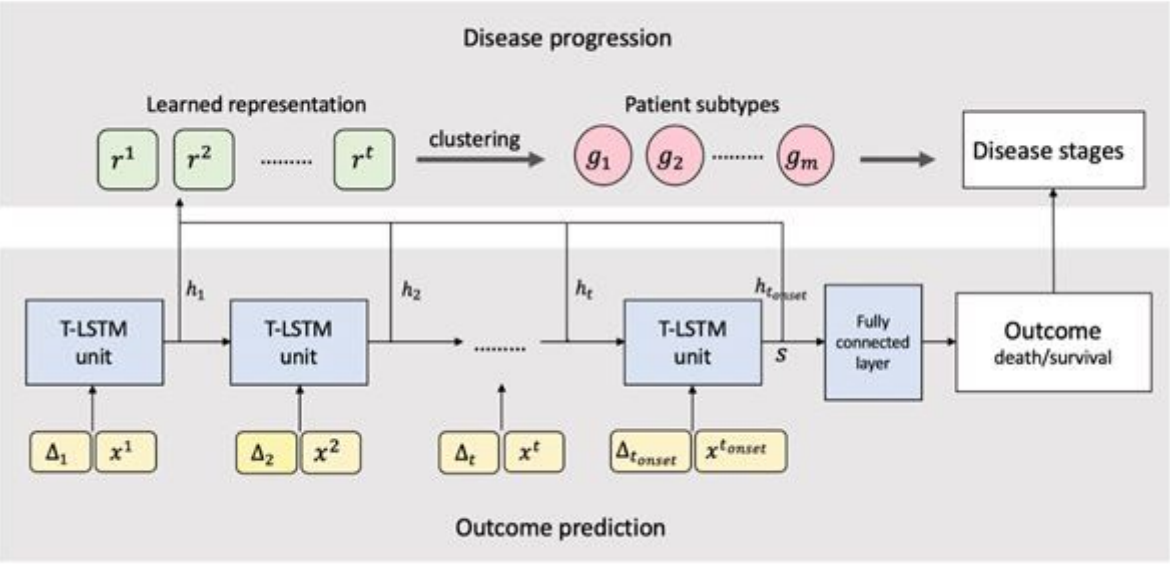

Figure 3

Process of our method The first gray box shows the process of identifying the disease progression. The specific method is divided into four steps. The details are presented in the ‘Analysis strategy’ section. The second gray box shows the process of predicting the outcome. The method uses T-LSTM as the learning model.

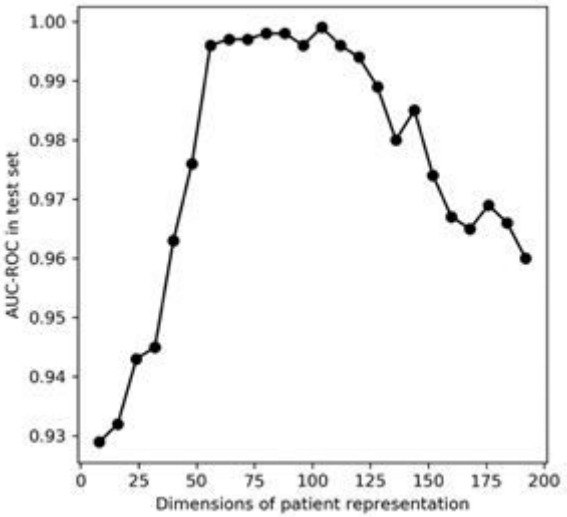

Figure 4

The relation of representation dimensions in T-LSTM and outcome prediction results The figure shows the outcome prediction results change as the representation dimensions of patients/hidden states dimensions of T-LSTM change. The results are evaluated by the area under the curve of the receiver operating characteristic (AUC-ROC). We decided to use 64 dimensions for higher accuracy and lower complexity.

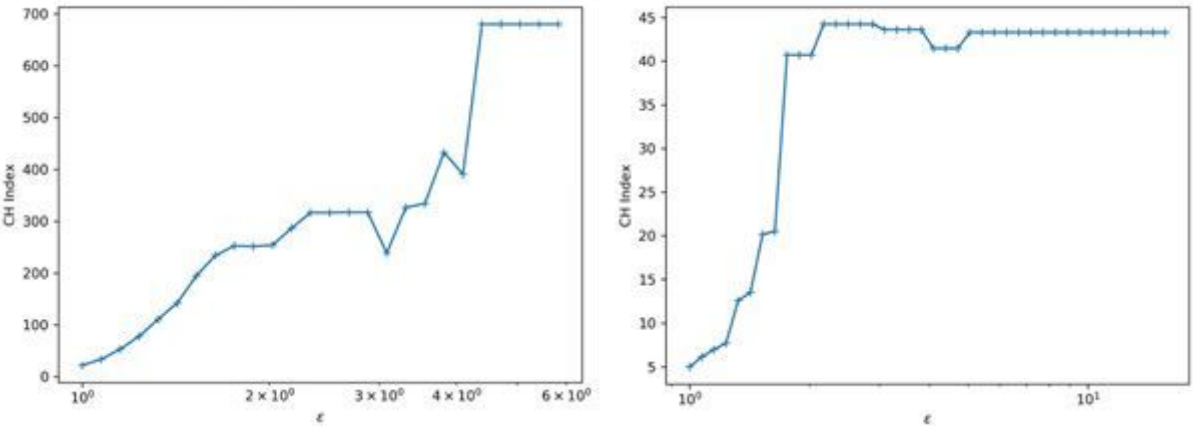

Figure 5

The change in the Calinski-Harabaz (CH) index value over the cluster radius parameter ε ε represents the cluster radius parameter. Changing ε will result in different clustering performances, which can be evaluated by the CH index. A larger CH value indicates better clustering; the best CH value for the death class and the survival class are 680.07 and 44.24, respectively.

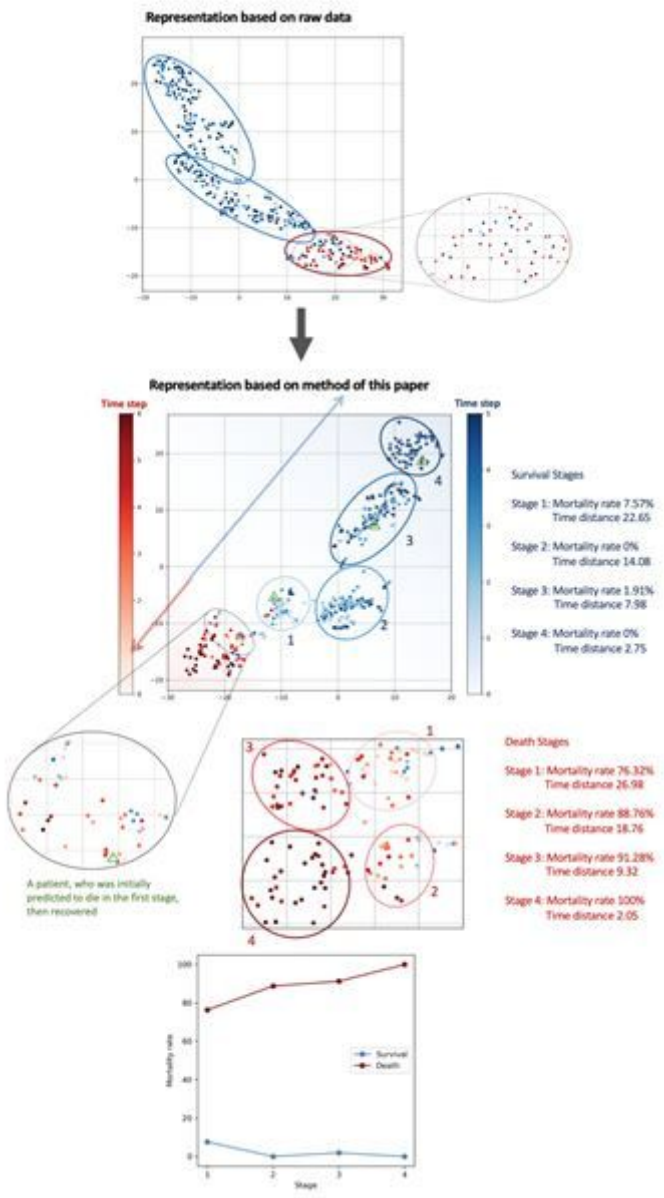

Figure 6

Four stages of disease progression in COVID-19 patients The first clustering is obtained based on the raw data representation. The second clustering is obtained based on the representations using the method in this study. The second clustering has the subtypes over time, which form 4 stages of disease progression. The 4 stages are closely related to mortality and time of illness. For example, the mortality rate decreased with stages in the survival class and increased in the death class, as the lower line chart shows.