# JOINT SPATIOTEMPORAL ATTENTION FOR MORTALITY PREDICTION OF PATIENTS WITH LONG COVID

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Long COVID is a general term of Post-Acute Sequelae of COVID-19. Patients with Long COVID can endure long-lasting symptoms including fatigue, headache, dyspnea and anosmia, facing increased risk of death. Identifying the cohorts with severe long-term complications in COVID-19 could benefit the treatment planning and resource arrangement. However, due to the heterogeneous phenotypes and various duration of symptoms presented in patients with Long COVID, it is difficult to predict their outcomes from their longitudinal data. In this study, we proposed a spatiotemporal attention mechanism to weigh feature importance jointly from the temporal dimension and feature space of longitudinal medical data. Considering that medical examinations can have interchangeable orders in adjacent time points, we restricted the learning of short-term dependency with a Local-LSTM and the learning of long-term dependency with the joint spatiotemporal attention. We also compared the proposed method with several state-of-the-art methods and a method in clinical practice. The methods are evaluated on a hard-to-acquire clinical dataset of patients with Long COVID. Experimental results show the Local-LSTM with joint spatiotemporal attention achieved superior performance in mortality prediction comparing to related methods. By analyzing the critical time points identified by the joint spatiotemporal attention, we identified time-specific prognostic biomarkers for life-threatening Long COVID. The proposed method provides a clinical tool for the severity assessment of Long COVID.

## 1 INTRODUCTION

SARS-CoV-2 is a novel coronavirus previously unknown. The infection of SARS-CoV-2 can cause coronavirus disease (COVID-19) with clinical syndromes including coughing, headache, fever, etc. Among the patients infected by SARS-CoV-2, a significant number of them have sustained post-infection sequelae, which is known as Long COVID. The patients with Long COVID can present long-lasting COVID-19 syndromes for at least two months after the acute infection (Soriano et al., 2021). The Long COVID symptoms such as fatigue, dyspnea, and memory problems could relapse, keep ongoing, or emerge in the following months and even years, which could endanger their lives (Blomberg et al., 2021). Although the World Health Organization (WHO) has updated the ICD-10 code to characterize Long COVID, currently there are few tools for quantitative assessment of the severity of the disease (Pfaff et al., 2022a). Identifying the Long COVID patients with high risk of death has great potential in providing in-time medical interventions.

The progression of Long COVID has been predicted from several data modalities in previous studies. For example, Acute Physiology and Chronic Health Evaluation II (Apache II) is used to predict hospital mortality for patients with COVID-19 based on physiologic variables, age, and previous health conditions (Zou et al., 2020). More recently, studies (Sneller et al., 2022; Pfaff et al., 2022b) show the prognostic power of electronic health record in predicting the outcome of patients with Long COVID. The predictions in these studies are based on the patients' statuses at the first admission, overlooking the longitudinal information in the patients' follow-up visits. For patients with Long COVID, the data collected from the first visit only describes the patients' condition at the onset of the disease, containing limited information about disease progression. By incorporating longitudinal electronic health record data, the past trajectory about the disease could inform the outcome prediction. In addition to electronic health record, medical imaging data are also shown predictive

for COVID-19 patients in the intensive care units (Cheng et al., 2022). Effective integration of the longitudinal electronic health record and medical imaging data could further enhance the outcome prediction.

Deep learning has led to many promising applications for longitudinal modeling of medical data (Morin et al., 2021). In a Covid-19 study, CT images at different time points were first registered, segmented and then subtracted as residual values for consolidation and recovery assessment (Kim et al., 2021). A generative adversarial network (GAN) was proposed to synthesize the colorful fundus photograph to facilitate the longitudinal prediction of advanced age-related macular degeneration (AMD) (Ganjdanesh et al., 2022). These two methods accept input with only two time points. To model medical data containing more than two time points and even in various length, recursive neural network (RNN) such as long-short-term-memory (LSTM) network (Lipton et al., 2015; Wang et al., 2019a) and gated recursive unit (GRU) network (Choi et al., 2016) have been applied. RNN is well-equipped with learning the short-term dependency (e.g., ordered/sequential patterns) in a sequence. Example applications include outcome predictions in diseases such as Alzheimer's disease (Nguyen et al., 2020), AMD (Altay et al., 2021) etc.

Attention enables the deep learning models to learn the correlation among multiple time points (or features) even in a long distance (i.e., long-term dependency). For example, a temporal attention mechanism was proposed to unravel the temporal importance of the history of measurements in making predictions with RNN (Lee et al., 2019). This attention mechanism assigns feature importance based on previous longitudinal measurements, giving same importance for features collected at the same measurement time. In another word, features at the same time point receive the same weights. In other studies, a self-attention is integrated with the LSTM to extract both local and global representations from the feature space. Example applications include the modeling of sequential chest X-rays (Cheng et al., 2022) and longitudinal clinical data (Nitski et al., 2021). In these two applications, the self-attention mechanisms provide static feature importance, ignoring the temporal changes of feature importance. In some chronic diseases such as Long COVID, however, some syndromes that happen at the early stage could indicate worsening conditions in the future (Bocchino et al., 2022). A video-based attention that first organizes features with spatial attention then combines them using temporal attention shows promising result for person identification (Li et al., 2018). Customizing the weights of features jointly for different features at different time points could further leverage the prognostic information contained in the data.

In this study, we proposed a deep learning-based, joint spatiotemporal attention mechanism to enable the Long COVID mortality prediction from longitudinal electronic health records. The proposed attention assigns feature importance by jointly considering the time and features from a global perspective. The joint spatiotemporal attention can be plugged into many deep learning models such as LSTM and GRU. This enables a deep neural network to effectively capture both short-term and long-term dependencies within the multi-modal time series. We evaluated our method on a clinical trial dataset composed of patients hospitalized due to severe COVID-19 pneumonia. The patients' electronic health records (EHR) are collected at their initial admission and during their multiple follow-up visits. The proposed methods were also compared with state-of-the-art deep learning methods and a clinical method.

Our contributions can be summarized as following: 1). We integrated multiple data modalities for Long COVID mortality prediction. 2). We proposed a joint spatiotemporal attention mechanism for deep learning-based longitudinal prediction models. 3). With the proposed model, we identified critical time points and feature patterns that are key for understanding the life-threatening Long COVID.

## 2 METHODS

### 2.1 PROBLEM FORMULATION

Consider a sequence of longitudinal data $\{x_i|x_1, x_2, \cdots, x_T\}$, where $T$ is the length of the sequence and $x_i$ denotes the data at the $ith$ time point. Here $x_i$ can be a vector, a matrix or a tensor depending on the data type of the application. The longitudinal prediction is to predict $y$, a clinical outcome such as the overall survival or progression-free survival based on the information from the longitu-

dinal data. In the case of mortality prediction of Long COVID from longitudinal medical data, it is formulated as a sequence-to-one problem.

### 2.2 JOINT SPATIOTEMPORAL ATTENTION

The sequence-to-one problem for Long COVID mortality prediction is challenging in that there are various syndromes/features at different time points. And the syndromes/features at early time points might be correlated with those at later time points. An accurate prognostic model is expected to account for both feature-wise information and temporal information. When applying attention mechanisms to longitudinal analysis, previous works either maintain time-dependent feature importance that ignores feature diversity (Lee et al., 2019) or provide spatial feature importance that is static over time (Nitski et al., 2021). These assumptions may not hold as different features could have varying importance as time progresses. For example, for patients previously hospitalized due to COVID-19 pneumonia, a study shows the fibrotic-like abnormalities are common in three months but mostly will disappear after one year (Bocchino et al., 2022). And consolidations disappeared in six months. In this scenario, the feature importance of fibrosis should not only change over time but also differ from that of consolidation.

To jointly weigh feature importance over the time axis and the feature space, we proposed a spatiotemporal attention mechanism as shown in Figure 1. Specifically, when transforming the input features into key-, query- and value-space, instead of using 1-dimensional linear layers that are commonly used in machine translation (Sankaran et al., 2016), we adopted 2-dimensional convolutional layers that are commonly used in computer vision domain (Wang et al., 2018). The proposed attention mechanism computes keys, queries and values with 1x1 convolutional filters to calculate the alignment score as feature importance. The benefit of using 1x1 convolutional filters for attention calculation is the joint weighting of feature importance from two dimensions including the time dimension and the feature dimension. We term the convolutional self-attention that moves both along the time axis and across the feature space as the joint spatiotemporal attention. As shown in Equation 1, the hidden features are adjusted by attention with a weighting factor $\gamma$.

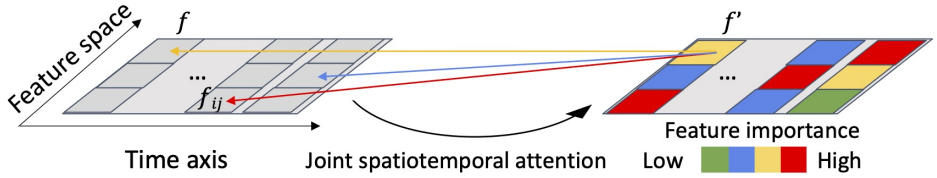


Figure 1: Illustration of the joint spatiotemporal attention. The feature importance of a feature $f$ at a specific time point is calculated by considering each feature's value at every time point $f_{ij}$.

$$f' = f + \gamma \cdot a(f), \tag{1}$$

where $f$ denotes the hidden features extracted by a deep neural network, $f'$ denotes the features' adjusted value by the joint spatiotemporal attention $a(\cdot)$. Equation 2 shows the calculation of the joint spatiotemporal attention.

$$a(f) = \sum_{i=1}^{H} \sum_{j=1}^{T} softmax(q(f_{ij}) \cdot k(f)) \cdot v(f), \tag{2}$$

where $H$ and $T$ represent the size of feature space and the number of time points, respectively; the feature importance of a given feature f is determined by other features $\{f_{ij} | i = 1, 2, \cdots, H; j = 1, 2, \cdots, T\}$. Specifically, the feature importance is measured by calculating the alignment score via the key $k()$, value $v()$ and query $q()$ operations that are implemented by $1 \times 1$ convolution.

### 2.3 LSTM WITH JOINT SPATIOTEMPORAL ATTENTION

We first applied RNN with joint spatiotemporal attention to predict the mortality of patients with Long COVID. To alleviate the vanishing gradient problem in standard RNN (Hochreiter, 1998), we adopted the LSTM network. Figure 2 shows the structure of the LSTM network with joint spatiotemporal attention. The LSTM network first extracts temporal dependency from the longitudinal input data into a feature map. The size of the feature map is $N \times H$, where $H$ is the number of features in the hidden state of one recurrent layer and $N$ is the number of stacked recurrent layers. In the stacked recurrent layers, one recurrent layer takes the latent features of previous recurrent layer as input and outputs the processed signals to the next recurrent layer. In this way, the early recurrent layers extract low-level (often short-term) dependency and the late recurrent layers extract high-level (often long-term) dependency. We use joint spatiotemporal attention to learn the time-dependent features' correlation. The adjusted feature map is fed to the multi-layer perceptron for outcome prediction. To evaluate the effectiveness the proposed method, we also compared it with a standalone LSTM model and a LSTM model with temporal attention (Lee et al., 2019).

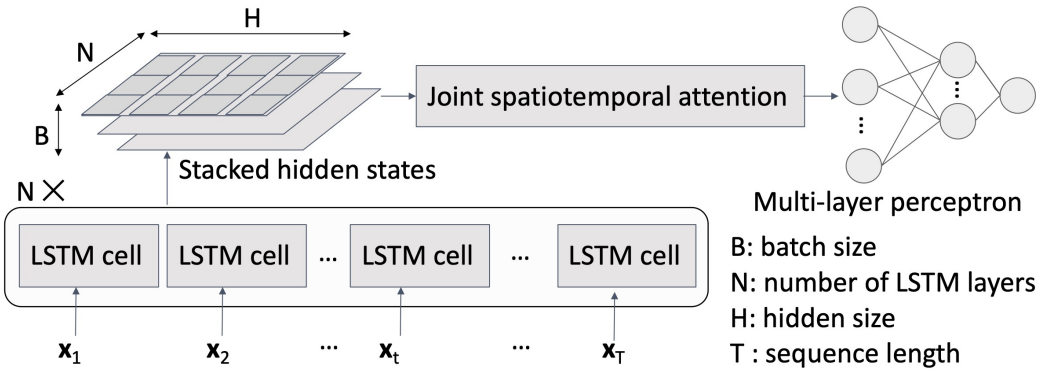


Figure 2: Model architecture of LSTM with joint spatiotemporal attention.

### 2.4 LOCAL-LSTM WITH JOINT SPATIOTEMPORAL ATTENTION

We integrated the joint spatiotemporal attention into a Local-LSTM for separate learning of short- and long-term dependency. While attention is good at learning long-term dependencies, it lacks the capability of modeling local/sequential structures in order (Vaswani et al., 2017), which usually exist in short-term dependencies. Take the electronic health record for an example. Different from an image or a sentence where the sequential ordering of elements has contextual meanings, the electronic health record within a few days does not follow strict orders (i.e., short-term randomness). For instance, a patient could undergo the lab tests on the first day and medical imaging examination on the next day or vice versa. The randomness of orders in short term makes the attention incapable of encoding the short-term dependency in electronic health record. In contrast, LSTM can account for the local randomness of ordering because signals stored in the memory cells can still propagate even if the local order is changed. To disentangle the learning of short- and long-term dependency, we restricted the learning of short-term/ordering dependency to a set of Local-LSTMs and the learning of long-term dependency to joint spatiotemporal attention. The Local-LSTM only learns the sequential patterns within a window size and extracts the local patterns as hidden states.

As shown in Figure 3, the Local-LSTM sequentially processes longitudinal data in a sliding window of a given length $t$. And each time point and its $t-1$ neighbors are represented by a latent feature vector of size $H$. For a sequence of length $T$, the shape of the sequence's latent feature map would be $T \times H$. After concatenating the hidden states from a batch of size $B$, the stacked hidden states become a tensor of size $B \times T \times H$. Then the joint spatiotemporal attention refines the hidden states by mining the feature-wise and long-term dependencies and outputs adjusted hidden states. The adjusted hidden states are fed to a multi-layer perceptron for mortality prediction. Our model differs from the R-Transformer (Wang et al., 2019b) in that we applied joint spatiotemporal attention on Local-LSTM while the previous work used only spatial attention on Local-RNN. Our work is also

different from the transformer models in computer vision (Dosovitskiy et al., 2020) and natural language understanding (Vaswani et al., 2017) where the temporal information are encoded with the positional embeddings while we encode the temporal information with Local-LSTM. Due to the short-term randomness of the sequences in EHR, we did not experiment with the methods based on positional embeddings.

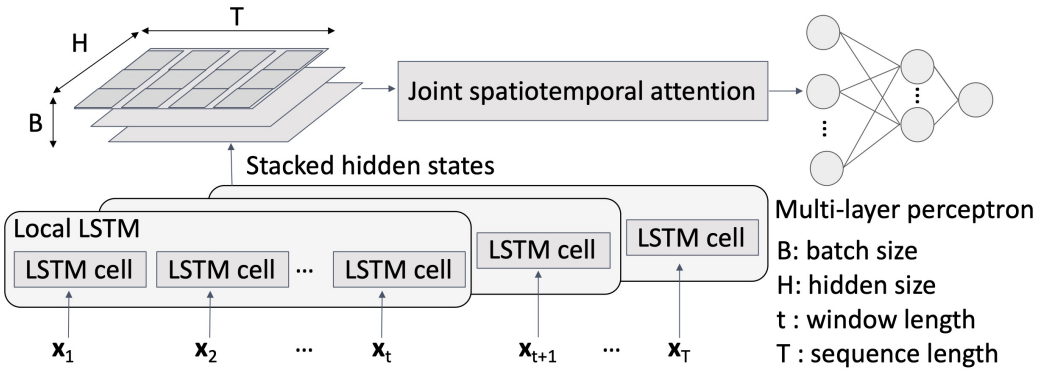


Figure 3: Model architecture of Local-LSTM with joint spatiotemporal attention. The model uses Local-LSTM to encode the short-term dependency and joint spatiotemporal attention to encode the long-term dependency.

### 2.5 The Clinical Model

Apache II system is a clinical nomogram that has been widely used in clinical practice to quantify the disease severity (Knaus et al., 1985). By measuring the initial values of 12 routine physiologic measurements, age, and previous health status, Apache II gives a score between 0 to 71 where higher scores indicate a higher risk of death. A previous study shows Apache II score is a significant prognostic biomarker of hospital mortality for patients with COVID-19 (Zou et al., 2020). This method is initially designed for patients admitted to ICU and does not account for longitudinal changes. As a result, Apache II only models patients' data at the admission to the hospital. We applied this model on our study cohort and used the Apache II score to predict the patients' mortality and to interpret the joint spatiotemporal attention. We term this model as the clinical model.

## 3 Experiments

### 3.1 Cohort and Dataset

Initially a total of 396 hospitalized patients with severe COVID-19 pneumonia were identified for this study. Patients enrolled in this study had been assigned to a clinical trial (anonymized for review purpose). The exclusion criteria included: i) patients without chest X-ray available; ii) patients whose contacts are lost in the follow-up visits; iii) patients with sparse data. Finally, 365 patients were included for the subsequent analysis. Patient data include demographic information, medical history, chest X-ray data that are collected at their initial admission to the hospital, and longitudinal data such as laboratory test and vital signs that are collected during patients' multiple visits to the clinics. The average number of time points for each patient is 10 with a standard deviation of 4. The patient survival statuses are collected at the 60th day after their initial admission to the hospital. We used radiographic assessment of lung oedema (RALE) (Warren et al., 2018) to quantify the disease severity reflected by the chest x-ray. Table 3 (in Appendix) shows the patient characteristics when they are first admitted to the hospital.

Table 4 (in Appendix) shows the laboratory test variables and their values from an example patient on day one. To reduce the sparse data issue, we selected prevalent variables with selection details in Data Preprocessing section. The selected laboratory test variables include Fibrinogen, C Reactive Protein, Prothrombin International Normalized Ratio, Prothrombin Time, Lactate Dehydrogenase,

D-Dimer, Albumin, Ferritin, Alanine Aminotransferase, Aspartate Aminotransferase, Chloride, Protein, Alkaline Phosphatase, Bilirubin, Calcium, Creatinine, Glucose, Hematocrit, Hemoglobin, Potassium, Platelets, Erythrocytes, Sodium, and Leukocytes.

### 3.2 Data Preprocessing

Considering the sparsity of some variables, we first calculated the percentage of patients having each variable. If a variable was tested by more than 95% of the cohort, we kept it for the following analysis. For longitudinal data with missing values, we used forward-filling as the imputation method; for non-longitudinal data, we used mean-filling. We selected important medical history variables including hypertension, obesity, hyperlipidemia, and diabetes mellitus based on previous studies (Pfaff et al., 2022a; Sneller et al., 2022). We calculated the score of radiographic assessment of lung oedema (RALE) to characterize the disease severity from chest X-rays (Warren et al., 2018). Measuring the degree of acute respiratory distress syndrome (ARDS), RALE has been widely used to describe the chest radiographic findings in patients positive for COVID-19. For numeric variables, min-max normalization was employed to put each variable on the same scale of zero to one. For binary variables, we represented them with zero or one. At each time point, we concatenated the longitudinal data at that time point and the static/non-longitudinal data into a vector.

### 3.3 Implementation Details

We implemented the proposed networks using PyTorch framework (Paszke et al., 2017). Stochastic gradient optimization (Ruder, 2016) was used for training. We tuned the hyperparameters on the validation set. For Local-LSTM, in light of the number of time points available, we experimented window sizes of from two to six. We performed five-fold cross validation. Based on the performance on validation set, we set window size as six; for both the LSTM model and the Local-LSTM model, the size of hidden features was set to 32. With the optimized hyperparameters, all models are trained for 50 epochs with a batch size of two. We scheduled an annealed learning rate that gradually decreased from 1e-3 to 1e-5. The model with the best performance on the validation set was evaluated on the testing set.

### 3.4 Experiment Settings

We first evaluated the prognostic values of different data modalities by incrementally incorporating them into a LSTM model. To evaluate the performance of joint spatiotemporal attention, we compared LSTM with joint spatiotemporal attention and a previous work, i.e., LSTM with temporal attention (Lee et al., 2019). The previous method has been used for survival prediction of cystic fibrosis disease. We adapted the previous method to model the mortality prediction of Long COVID. To maintain the focus of spatiotemporal attention on long-term dependencies, we replaced LSTM with Local-LSTM and formed the new model named Local-LSTM with spatiotemporal attention. The performance for mortality prediction are evaluated using the area under the receiver operating curve (AUC). We reported the average AUC after five-fold cross validation.

To interpret the proposed model's mortality prediction, we first visualized the joint spatiotemporal attention that represents the hidden features' importance over time. Then we compared the critical time points identified by joint spatiotemporal attention (i.e., time points with the most important features) and the critical time points identified by the Apache II system (i.e., time points with the highest increase of Apache II score). Then we performed Mann-Whitney U test to find the common critical time points that are identified by both methods.

## 4 Results

### 4.1 Comparison of Different Input Combinations

Table 1 shows the AUC when using LSTM model to combine differnt data modalities for mortality prediction of Long COVID. As shown in Table 1, when only using the laboratory test data (in longitudinal format), the LSTM model achieves an AUC of 0.63 on the testing set. With the incorporation of vital signs (in longitudinal format), the LSTM model's performance is increased to

Table 1: Prediction performance of Long COVID patients' mortality using LSTM model with different data modalities.

| Lab tests | Vital signs | Demographic | Medical history | Medical images | AUC |
|---|---|---|---|---|---|
| X | | | | | 0.63 |
| X | X | | | | 0.70 |
| X | X | X | | | 0.73 |
| X | X | X | X | | 0.75 |
| X | X | X | X | X | 0.76 |

Table 2: Prediction performance of Long COVID patients' mortality using different models.

| Model name | AUC |
|---|---|
| Clinical model (Knaus et al., 1985) | 0.61 |
| LSTM | 0.76 |
| LSTM with temporal attention (Lee et al., 2019) | 0.77 |
| LSTM with joint spatiotemporal attention | 0.80 |
| Local-LSTM with joint spatiotemporal attention | 0.87 |

0.70. These two results demonstrate the effectiveness of LSTM in modeling longitudinal data. Then we incorporated the non-longitudinal data collected at the patients' initial admission to the hospital. The static data include demographic data, medical history data, and chest x-ray data (in the form of RALE score). The addition of static data further increases the LSTM model's AUC to 0.73, 0.75 and 0.76, respectively.

### 4.2 Comparison of Different Models

Table 2 shows the performance of models with different network architectures for mortality prediction of Long COVID. The clinical model achieves an AUC of 0.61. We use the LSTM model with a combination of all data modalities as the baseline model (AUC = 0.76). After adding temporal attention, the LSTM model's AUC is slightly increased from 0.76 to 0.77. After replacing the temporal attention with the joint spatiotemporal attention, the LSTM model's AUC is further increased to 0.80. By replacing LSTM with Local-LSTM, the AUC is improved to 0.87.

### 4.3 Identifying Key Prognostic Biomarkers for Severe Long COVID

The feature maps adjusted by joint spatiotemporal attention contain non-linear interactions of multiple variables. Explaining the attention with the hidden features is less straightforward than explaining variables with physical meanings such as heart rate, temperature, etc. To fill the gap between attention and clinically interpretable phenotypes, we used Apache II system as a bridge to explain where the attention is looking at. Figure 4 shows the increase of Apache II score decomposed by different physiologic variables of an example patient. Figure 5 shows the Local-LSTM's joint spatiotemporal attention on the feature maps of the same patient. As can be seen, a feature's importance varies from time point to time point and can be different among features at the same time point.

Based on the statistical testing of critical time points identified by the joint spatiotemporal attention and Apache II system, we discovered the following patterns: 1) the features at the initial and final time points tend to have higher importance than features at other time points, which correspond to the disease onset and the most recent condition of the patient, respectively. The findings about initial time points' prognostic value align with a previous study showing that experiencing more than five symptoms during the first week of illness is associated with Long COVID (Sudre et al., 2021); 2). Significantly ($p \leq 0.05$) correlated critical time points between joint spatiotemporal attention and Apache II system include: i) when the respiratory rate is abnormal at the initial and 60-day time points, and ii) when the heart rate and creatinine level are both abnormal at any time point. With

the Apache II system as an interpreter, we identified these biomarkers and temporal patterns that the Local-LSTM model focused its attention on.

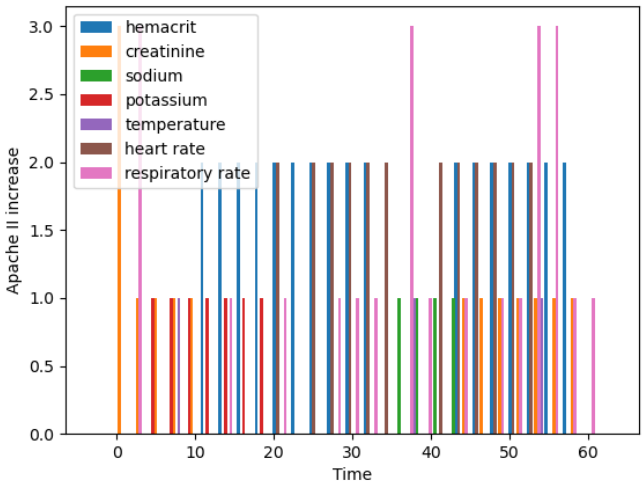


Figure 4: Increase of Apache II score of an example patient at different time points. The physiologic variables' ticks on Time axis are slightly shifted for presentation purpose.

## 5 Discussion

In this study, we proposed a joint spatiotemporal attention mechanism to enable the simultaneous learning of feature importance across the temporal and feature space. The integration of joint spatiotemporal attention into the LSTM model and the Local-LSTM model demonstrated the effectiveness of the proposed attention mechanism. We further evaluated the LSTM model and the Local-LSTM model by comparing with related methods. The experiments on a real-world and hard-to-acquire clinical dataset demonstrated the proposed models' effectiveness for mortality prediction of patients with Long COVID.

### 5.1 Modeling with longitudinal data

We proposed a joint spatiotemporal attention to assign dynamic feature importance both along the time axis and across the feature space. Previous work either calculates the feature importance independent of time or calculates the temporal importance being unaware of the inter-feature differences. In longitudinal modeling, we argue that the same feature at different time points have varied prognostic values and the feature importance should vary according to both time and features. The proposed joint spatiotemporal attention addresses this problem by learning two-dimensional convolutional filters. Our experimental results show that the LSTM with joint spatiotemporal attention (AUC=0.80) outperformed LSTM with temporal attention (AUC=0.77) and LSTM without attention (AUC=0.76). This demonstrates the effectiveness of the proposed attention mechanism.

| | | Feature index | | | |
|---|---|---|---|---|---|
| | | 0 | ... | 158 | 159 |
| Time | 10/16/20 | 0.10815899 | 0.07561667 | 0.0275702 | 0.08629946 |
| | 10/17/20 | 0.06203437 | 0.08487909 | 0.01524038 | 0.03632014 |
| | 10/18/20 | 0.11362207 | 0.00849104 | -0.0193095 | 0.05668567 |
| | 10/19/20 | 0.02322575 | 0.04783887 | -0.0197297 | 0.09630996 |
| | 10/20/20 | 0.05145983 | 0.01117645 | -0.0572149 | 0.11249094 |
| | 10/21/20 | -0.0157626 | 0.01290985 | -0.0223108 | 0.01876476 |
| | 10/22/20 | 0.08936708 | -0.0086527 | -0.0410364 | 0.05247996 |
| | 10/23/20 | 0.05295146 | 0.03098444 | 0.08147199 | 0.07840578 |
| | 10/24/20 | 0.15765366 | -0.1093133 | -0.0744994 | -0.0947516 |
| | 10/25/20 | 0.03406258 | 0.04810617 | 0.01382765 | 0.07615975 |
| | 10/26/20 | 0.01639158 | -0.0879296 | -0.0681836 | -0.0593095 |
| | 10/27/20 | 0.07681573 | -0.0741408 | -0.0610925 | 0.03111256 |
| | 10/28/20 | 0.00340163 | 0.06931681 | 0.10540402 | 0.06617329 |
| | 10/29/20 | 0.03588036 | -0.0224102 | 0.06904365 | 0.01707054 |
| | 10/30/20 | 0.13164961 | 0.09018614 | 0.06770451 | 0.11011408 |
| | 10/31/20 | 0.1192321 | 0.06912498 | -0.0089304 | 0.11881551 |
| | 11/1/20 | 0.18044096 | -0.0147651 | -0.0393168 | -0.0353636 |
| | 11/2/20 | 0.05234501 | -0.0539291 | 0.03292136 | 0.08675104 |
| | 11/3/20 | 0.03294804 | -0.0640073 | -0.0416712 | 0.0179613 |
| | 11/4/20 | 0.01797248 | 0.02587942 | -0.0761838 | -0.0350457 |
| | 11/5/20 | 0.04986078 | -0.11638 | -0.1063716 | -0.0936633 |
| | 11/6/20 | -0.007678 | 0.02138656 | 0.00654381 | 0.03143106 |
| | 11/7/20 | 0.10419223 | -0.0725146 | 0.02701446 | 0.03552507 |
| | 11/8/20 | 0.03719689 | -0.0324361 | -0.0097108 | -0.0036236 |
| | 11/9/20 | 0.03135437 | 0.13086885 | 0.18720034 | 0.14907286 |
| | 11/10/20 | 0.19106624 | 0.03678167 | -0.0475367 | 0.00183725 |
| | 11/11/20 | 0.08993081 | 0.06927527 | 0.00383627 | 0.19988334 |

Figure 5: Visualization of joint spatiotemporal attention in Local-LSTM model for an example patient.

When comparing the clinical model, LSTM models with and without attention and Local-LSTM model with attention, we can see that the Local-LSTM model achieves higher performance (AUC=0.87) than the rest models. The regular LSTM model entangles the learning of the short-term/ordering and long-term dependencies together. In contrast, the Local-LSTM model enables the separate learning of short-term and long-term dependency in longitudinal data. The joint spatiotemporal attention receives latent features processed by the Local-LSTM and can only learn the remaining information about long-term dependency. Since attention is unable to model the sequential patterns in local structure, leaving the short-term/ordering dependency learning to LocalLSTM avoided attention's drawback. This enhanced the whole model (Local-LSTM with spatiotemporal attention)'s capacity in pattern recognition of longitudinal data. The results suggest separate learning of short-term/ordering and long-term dependency could improve the outcome prediction. In addition, the clinical model does not perform as good as the LSTM model or the Local-LSTM model for many reasons. First of all, it may have to do with its linear addition of scores, which excluded nonlinear interactions among variables. Secondly, some of the variables for Apache II score's calculation is not available in our dataset and could also limit Apache II system's performance. In addition, the clinical model is incapable of modeling longitudinal data.

### 5.2 Clinical impact

Our study provides a comprehensive assessment of the prognostic values for Long COVID in multiple data modalities. Patients with Long COVID exhibit different phenotype. Identifying the prognostic variables has great significance. Our study shows that the vital signs, laboratory tests, demographics, medical history, and medical images all have their unique prognostic values for Long COVID outcome prediction. Combining them all together achieved higher AUC for mortality prediction than other combinations. Previous studies on Long COVID separately investigated the non-imaging electronic medical records and medical imaging data. Our study reveals the slightly complementary role of medical imaging to electronic health record in Long COVID mortality prediction.

Our study reveals key prognostic biomarkers for the mortality prediction of Long COVID. The comparison of critical time points highlighted by joint spatiotemporal attention and by Apache II system enabled us to identify three physiologic variables, i.e., respiratory rate, heart rate, and creatinine level. The identification of these variables at specific time points is clinically meaningful. They could not only help us better understand the progression of Long COVID but also serve as prognostic biomarkers to monitor the condition of patients with Long COVID.

## 6 Conclusion

In conclusion, we proposed a joint spatiotemporal attention mechanism for Long COVID mortality prediction from longitudinal medical data. We integrated the proposed attention mechanism to deep learning frameworks including a LSTM model and a Local-LSTM model. Our experiments on the mortality prediction of patients with Long COVID show the effectiveness of the proposed method. We also demonstrated the explainability of the joint spatiotemporal attention by identifying three prognostic biomarkers including respiratory rate, heart rate, and creatinine level. In this post-pandemic era, our method not only provided an in-time clinical tool for mortality prediction of Long COVID but also advanced the medical knowledge about the disease progression of Long COVID.

**Limitations & Future work** The proposed joint spatiotemporal attention has a few limitations. It requires the same set of features at each time point. Though we performed data imputation with forward filling, it may still exclude some patients with sparse data. To accommodate the sparse data issue, one way is to impute the missing data from patients with similar conditions, which needs unsupervised learning in future work. In addition, to better represent imaging data, replacing the RALE scores with CNN features merits further investigation. Finally, we plan to evaluate the method on additional longitudinal prediction tasks and other diseases.

## REFERENCES

Fatih Altay, Guillermo Ramón Sánchez, Yanli James, Stephen V Faraone, Senem Velipasalar, and Asif Salekin. Preclinical stage alzheimer's disease detection using magnetic resonance image scans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 15088–15097, 2021.

Bjørn Blomberg, Kristin Greve-Isdahl Mohn, Karl Albert Brokstad, Fan Zhou, Dagrun Waag Linchausen, Bent-Are Hansen, Sarah Lartey, Therese Bredholt Onyango, Kanika Kuwelker, Marianne Sævik, et al. Long covid in a prospective cohort of home-isolated patients. *Nature medicine*, 27(9):1607–1613, 2021.

Marialuisa Bocchino, Roberta Lieto, Federica Romano, Giacomo Sica, Giorgio Bocchini, Emanuele Muto, Ludovica Capitelli, Davide Sequino, Tullio Valente, Giuseppe Fiorentino, et al. Chest ct-based assessment of 1-year outcomes after moderate covid-19 pneumonia. *Radiology*, pp. 220019, 2022.

Jianhong Cheng, John Sollee, Celina Hsieh, Hailin Yue, Nicholas Vandal, Justin Shanahan, Ji Whae Choi, Thi My Linh Tran, Kasey Halsey, Franklin Iheanacho, et al. Covid-19 mortality prediction in the intensive care unit with deep learning based on longitudinal chest x-rays and clinical data. *European radiology*, pp. 1–11, 2022.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pp. 301–318. PMLR, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alireza Ganjdanesh, Jipeng Zhang, Emily Y Chew, Ying Ding, Heng Huang, and Wei Chen. Longlnet: temporal correlation structure guided deep learning model to predict longitudinal age-related macular degeneration severity. *PNAS nexus*, 1(1):pgab003, 2022.

Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02): 107–116, 1998.

Seong Tae Kim, Leili Goli, Magdalini Paschali, Ashkan Khakzar, Matthias Keicher, Tobias Czempiel, Egon Burian, Rickmer Braren, Nassir Navab, and Thomas Wendler. Longitudinal quantitative assessment of covid-19 infection progression from chest cts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 273–282. Springer, 2021.

William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

Changhee Lee, Jinsung Yoon, and Mihaela Van Der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2019.

Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 369–378, 2018.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

Olivier Morin, Martin Vallières, Steve Braunstein, Jorge Barrios Ginart, Taman Upadhaya, Henry C Woodruff, Alex Zwanenburg, Avishek Chatterjee, Javier E Villanueva-Meyer, Gilmer Valdes, et al. An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. *Nature Cancer*, 2(7):709–722, 2021.

Minh Nguyen, Tong He, Lijun An, Daniel C Alexander, Jiashi Feng, BT Thomas Yeo, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, 222:117203, 2020.

Osvald Nitski, Amirhossein Azhie, Fakhar Ali Qazi-Arisar, Xueqi Wang, Shihao Ma, Leslie Lilly, Kymberly D Watt, Josh Levitsky, Sumeet K Asrani, Douglas S Lee, et al. Long-term mortality risk stratification of liver transplant recipients: real-time application of deep learning algorithms on longitudinal data. *The Lancet Digital Health*, 3(5):e295–e305, 2021.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Emily Pfaff, Charisse Madlock-Brown, John M Baratta, Abhishek Bhatia, Hannah Davis, Andrew T Girvin, Elaine Hill, Liz Kelly, Kristin Kostka, Johanna Loomba, et al. Coding long covid: Characterizing a new disease through an icd-10 lens. *medRxiv*, 2022a.

Emily R Pfaff, Andrew T Girvin, Tellen D Bennett, Abhishek Bhatia, Ian M Brooks, Rachel R Deer, Jonathan P Dekermanjian, Sarah Elizabeth Jolley, Michael G Kahn, Kristin Kostka, et al. Identifying who has long covid in the usa: a machine learning approach using n3c data. *The Lancet Digital Health*, 2022b.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Baskaran Sankaran, Haitao Mi, Yaser Al-Onaizan, and Abe Ittycheriah. Temporal attention model for neural machine translation. *arXiv preprint arXiv:1608.02927*, 2016.

Michael C Sneller, C Jason Liang, Adriana R Marques, Joyce Y Chung, Sujata M Shanbhag, Joseph R Fontana, Haniya Raza, Onyi Okeke, Robin L Dewar, Bryan P Higgins, et al. A longitudinal study of covid-19 sequelae and immunity: baseline findings. *Annals of Internal Medicine*, 2022.

Joan B Soriano, Srinivas Murthy, John C Marshall, Pryanka Relan, Janet V Diaz, WHO Clinical Case Definition Working Group, et al. A clinical case definition of post-covid-19 condition by a delphi consensus. *The Lancet Infectious Diseases*, 2021.

Carole H Sudre, Benjamin Murray, Thomas Varsavsky, Mark S Graham, Rose S Penfold, Ruth C Bowyer, Joan Capdevila Pujol, Kerstin Klaser, Michela Antonelli, Liane S Canas, et al. Attributes and predictors of long covid. *Nature medicine*, 27(4):626–631, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tingyan Wang, Yuanxin Tian, and Robin G Qiu. Long short-term memory recurrent neural networks for multiple diseases risk prediction by leveraging longitudinal medical records. *IEEE journal of biomedical and health informatics*, 24(8):2337–2346, 2019a.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

Zhiwei Wang, Yao Ma, Zitao Liu, and Jiliang Tang. R-transformer: Recurrent neural network enhanced transformer. *arXiv preprint arXiv:1907.05572*, 2019b.

Melissa A Warren, Zhiguou Zhao, Tatsuki Koyama, Julie A Bastarache, Ciara M Shaver, Matthew W Semler, Todd W Rice, Michael A Matthay, Carolyn S Calfee, and Lorraine B Ware. Severity scoring of lung oedema on the chest radiograph is associated with clinical outcomes in ards. *Thorax*, 73(9):840–846, 2018.

Xiaojing Zou, Shusheng Li, Minghao Fang, Ming Hu, Yi Bian, Jianmin Ling, Shanshan Yu, Liang Jing, Donghui Li, and Jiao Huang. Acute physiology and chronic health evaluation ii score as a predictor of hospital mortality in patients of coronavirus disease 2019. *Critical care medicine*, 48 (8):e657, 2020.

# A APPENDIX

Table 3: Characteristics of the cohort on admission to the hospital.

| Characteristics | Value |
|---|---|
| Total number of patients | 365 |
| Deceased after 60 days | 65 |
| Mean age, years | 58±13 |
| Sex | |
| Male | 217 |
| Female | 217 |
| Mean body mass index, kg/m$^2$ | 28±13 |
| Medical history | |
| Diabetes mellitus | 69 |
| Hyperlipidemia | 70 |
| Obesity | 113 |
| Hypertension | 192 |
| Vital signs | |
| Systolic Blood Pressure, millimeters of mercury | 124±17 |
| Diastolic Blood Pressure, millimeters of mercury | 72±10 |
| Pulse Rate, beats per minute | 76±16 |
| Respiratory Rate, breaths per minute | 23 |
| Laboratory test | See Table 4 |

Table 4: Laboratory test variables of an example patient on day one.

| Laboratory test variable name | Value | Units |
|---|---|---|
| Albumin | 3.2 | g/dL |
| Alkaline Phosphatase | 80 | U/L |
| Alanine Aminotransferase | 26 | U/L |
| Activated Partial Thromboplastin Time | 36 | sec |
| Aspartate Aminotransferase | 21 | U/L |
| Basophils | 0.1 | $10^3/uL$ |
| Bilirubin | 0.7 | mg/dL |
| N-Terminal ProB-type Natriuretic Peptide | 171.9 | pg/mL |
| Blood Urea Nitrogen | 15 | mg/dL |
| Calcium | 8.5 | mg/dL |
| Chloride | 100 | mEq/L |
| Carbon Dioxide | 28 | mEq/L |
| Creatinine | 0.6 | mg/dL |
| C Reactive Protein | 86.1 | mg/L |
| D-Dimer | 0.35 | ug/mL FEU |
| Eosinophils | 0 | $10^3/uL$ |
| Ferritin | 1815 | ng/mL |
| Fibrinogen | 662 | mg/dL |
| Glucose | 66 | mg/dL |
| Hematocrit | 40.1 | % |
| Hemoglobin | 13.9 | g/dL |
| Prothrombin International Normalized Ratio | 1.3 | RATIO |
| Potassium | 3.6 | mEq/L |
| Lactate Dehydrogenase | 229 | U/L |
| Lymphocytes | 1.2 | $10^3/uL$ |
| Monocytes | 0.6 | $10^3/uL$ |
| Neutrophils | 6.5 | $10^3/uL$ |
| Phosphate | 2.3 | mg/dL |
| Platelets | 314 | $10^3/uL$ |
| Protein | 6.4 | g/dL |
| Prothrombin Time | 16 | sec |
| Erythrocytes | 4.43 | $10^6/uL$ |
| Soluble Interleukin 1 Receptor-Like 1 | 67700 | pg/mL |
| Sodium | 139 | mEq/L |
| Troponin T | 0.01 | ng/mL |
| Unspecified Cells | 0 | $10^3/uL$ |
| Urate | 1.9 | mg/dL |
| Leukocytes | 8.3 | $10^3/uL$ |