Learning Deeply Enriched Representations of Temporal Data of COVID-19 Patients for Improved Mortality Prediction

Anonymous Authors

Abstract

An influx of Coronavirus Disease 19 (COVID-19) infections puts pressure on healthcare systems and disrupts their general care routine, and efficient allocation of finite resources becomes a crucial problem. To aid logistical decision-making, we propose a novel semi-supervised learning framework to predict clinical outcomes of COVID-19 patients. Because the clinical measurements of COVID-19 patients can be collected over time, our model aims to predict the clinical outcomes from the multivariate time series (MTS) which may contain missing data. We leverage Long Short Term Memory (LSTM) architecture to learn the vectorial representation of MTS with missing data and uneven time intervals between records. Armed with the vectorial representation of MTS, conventional machine learning models can be used to predict the clinical outcomes. In our experiments, the proposed model predicts mortality of COVID-19 patient with the blood samples collected from the 358 patients infected with COVID-19 in Wuhan, China. Our embedding framework shows 88% to 94% prediction accuracy, even if very few samples are labeled. In addition, we identify the mortality relevant biomarkers from the proposed method.[1]

1 Introduction

Sudden increases in COVID-19 cases, such as during seasonal waves, quickly deplete the limited resources of health care systems, forcing clinicians to set criteria for distribution of scarce treatments [for Disease Control *et al.*,]. In an earlier study, Yan et al. advocated for a machine learning mortality-prediction model to inform logistical planning [Yan *et al.*, 2020]. The model [Yan *et al.*, 2020] proposed in the study identifies the most predictive biomarkers for the patient's mortality using a XGBoost classifier [Chen and Guestrin, 2016] trained on a publicly available MTS dataset collected from COVID-19-positive patients admitted to Tongji Hospital in China. Acknowledging the limitations of random XGBoost classifier in synthesizing longitudinal data, only the final record was used to train and test the model. While the model accomplishes the task of determining the most predictive biomarkers, it hardly captures the temporal trends, which are crucial to capturing the progression of this disease. In addition, their model is not able to predict the mortality when those principle biomarkers are not measured.

Conventional classification methods in MTS analysis usually require sequences to be of consistent length [Lu *et al.*, 2018], which is rarely the case when data is collected outside of a carefully controlled environment. Our samples were collected during an outbreak of a pandemic, a time when hospitals function out of routine and blood draws are not performed for a controlled study. As a result, the number of blood draws is different for every patient and time intervals between samples are also often irregular. The data for this specific dataset was also collected from CRFs (case report forms), or surveys filled out by individuals rather than being transcribed directly from a test result. The resultant inputs are uneven sequences of vectors with missing data that introduce several new difficulties, particularly when the missing data includes patient outcome, the target variable of predictive model. One might suggest imputation as a possible solution, at least to tackle the missing data problem. Modern imputation methods rely on a framework of generative adversarial networks (GANs) [Yoon *et al.*, 2018] to successfully impute sizeable portions of high-dimensional data. Although they are efficient, the imputation may introduce undesirable bias in prediction.

In addition, data collected in the field usually consists of labeled and unlabeled samples, and the supervised [Yan *et al.*, 2020] or unsupervised [Längkvist *et al.*, 2014; Srivastava *et al.*, 2015] learning models cannot fully utilize both samples. Therefore we propose an alternate solution in the form of a semi-supervised enrichment method to transform the incomplete MTS into the vectorial representation which can be readily fed into the conventional machine learning models. We outline the proposed model in Fig 1.

2 Methods

In this section, we describe the structure of proposed autoencoder (AE) which is a composite model of three components. The LSTM encoder ϕ_E encodes the longitudinal record of each patient into the enriched representation. The multilayer perceptron (MLP) decoder ϕ_D decodes the enriched repre-

[1]The codes used in our experiments will be made publicly accessible on GitHub upon paper acceptance.

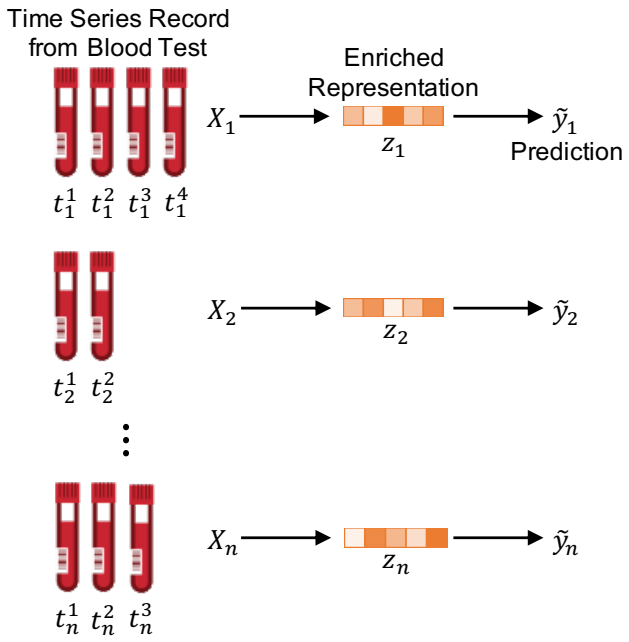

Figure 1: The prediction via enrichment. MTS with uneven time intervals and missing data is represented by the sparse matrix $\mathbf{X}_i$ of varying size. We first compress the matrix of MTS into the enriched representation in a format of fixed length vector, and then predict target labels from the enriched vectorial representation.

sentation into the original record. The MLP predictor ϕ_P predicts the target labels from the enriched representation. The input and output of the proposed model is described in Fig. 2.

2.1 Dataset

We obtain the blood sample records, demographic information, and associated mortality outcomes of 375 patients collected throughout their stay in Tongji hospital between 10th January and 24th February 2020 following the previous research [Yan *et al.*, 2020]. We discard 17 samples whose no time stamp is recorded. Among the remaining 358 samples, 192 patients survived (labeled as 1) and 166 patients died (labeled as 0). The order of samples is randomly shuffled to prevent the bias from the order. Each feature of dataset is normalized by min-max scaling to the range [0, 1].

2.2 Notations

In this paper, we denote a vector as a bold lower case letter, and matrix as a bold upper case letter. We use i and j to index i-th participant and j-th record respectively. We describe the records of i-th participant as $\mathcal{X}_i = \{\mathbf{x}_i^b, \mathbf{X}_i, \mathbf{M}_i, \mathbf{t}_i\}$ as follows:

- $\mathbf{x}_i^b \in \Re^{D_b}$ is a vector of basic demographic information, such as age and gender.
- $\mathbf{X}_i = [\mathbf{x}_i^1; \mathbf{x}_i^2; \cdots; \mathbf{x}_i^{n_i}] \in \Re^{n_i \times D_l}$ are the longitudinal records collected from the blood tests across the n_i time points.
- $\mathbf{M}_i = [\mathbf{m}_i^1; \mathbf{m}_i^2; \cdots; \mathbf{m}_i^{n_i}] \in \{1, 0\}^{n_i \times D_l}$ are the binary masks of observabilities of longitudinal records $\mathbf{X}_i$, where 1 and 0 indicates the observed and unobserved entry respectively.
- $\mathbf{t}_i = [t_i^1; t_i^2; \cdots; t_i^{n_i}] \in \Re^{n_i}$ are the time stamps of n_i records.

The missing entries in $\mathbf{X}_i$ are initialized with the constant -1, and this value does not affect the result. The target label $\mathbf{y}_i \in \{0, 1\}$ is the mortality of i-th participant, which is provided in the training process if that participant is in training set, such that $i \in \Omega$.

2.3 Encoder

We leverage LSTM encoder $\phi_E : \Re^{n_i \times (2D_l+1)} \mapsto \Re^{d_z}$ to summarize the longitudinal records and learn the temporal relation between records. The time stamp of each record is crucial in learning the temporal relation between records (e.g. temporal locality), while the the missingness pattern of the entries may represent the participants' status. Thus we provide the concatenation of longitudinal records, masks, and time stamps, $[\mathbf{X}_i, \mathbf{M}_i, \mathbf{t}_i] = [\hat{\mathbf{x}}_i^1; \hat{\mathbf{x}}_i^2; \cdots; \hat{\mathbf{x}}_i^{n_i}] = \hat{\mathbf{X}}_i \in \Re^{n_i \times (2D_l+1)}$, as an input of LSTM encoder such that $\phi_E(\mathbf{X}_i, \mathbf{M}_i, \mathbf{t}_i;\ \theta_E) = \mathbf{z}_i$. Here, θ_E denotes the set of trainable parameters of LSTM.

For each time step ($1 \le j \le n_i$), the input record $\hat{\mathbf{x}}_i^j$ of i-th patient is processed by following the LSTM architecture [Yu *et al.*, 2019]:

$$\mathbf{k}_i^j = \sigma(\hat{\mathbf{x}}_i^j \mathbf{W}_{xk} + \mathbf{h}_i^{j-1} \mathbf{W}_{hk} + \mathbf{c}_i^{j-1} \mathbf{W}_{ck} + \mathbf{b}_k), \quad (1)$$

$$\mathbf{f}_i^j = \sigma(\hat{\mathbf{x}}_i^j \mathbf{W}_{xf} + \mathbf{h}_i^{j-1} \mathbf{W}_{hf} + \mathbf{c}_i^{j-1} \mathbf{W}_{cf} + \mathbf{b}_f), \quad (2)$$

$$\mathbf{c}_i^j = \mathbf{f}_i^j \odot \mathbf{c}_i^{j-1} + \mathbf{k}_i^j \odot \tanh(\hat{\mathbf{x}}_i^j \mathbf{W}_{xc} + \mathbf{h}_i^{j-1} \mathbf{W}_{hc} + \mathbf{b}_c), \quad (3)$$

$$\mathbf{o}_i^j = \sigma(\hat{\mathbf{x}}_i^j \mathbf{W}_{xo} + \mathbf{h}_i^{j-1} \mathbf{W}_{ho} + \mathbf{c}_i^j \mathbf{W}_{co} + \mathbf{b}_o), \quad (4)$$

$$\mathbf{h}_i^j = \mathbf{o}_i^j \odot \tanh(\mathbf{c}_i^j), \quad (5)$$

where σ and tanh is the logistic sigmoid and hyperbolic tangent activation function respectively, and $\mathbf{k}_i^j$, $\mathbf{o}_i^j$, $\mathbf{f}_i^j$ are input, output, forget gate of j-th time step respectively. $\{\mathbf{W}_{xk}, \mathbf{W}_{hk}, \mathbf{W}_{ck}, \mathbf{W}_{xf}, \mathbf{W}_{hf}, \mathbf{W}_{cf}, \mathbf{W}_{xc}, \mathbf{W}_{hc}, \mathbf{W}_{xo}, \mathbf{W}_{ho}, \mathbf{W}_{co}\} \subset \theta_E$ are trainable weight matrices and $\{\mathbf{b}_k, \mathbf{b}_f, \mathbf{b}_c, \mathbf{b}_o\} \subset \theta_E$ are trainable bias vectors. $\mathbf{c}_i^j$ and $\mathbf{h}_i^j$ denote the cell state and hidden representation at j-th time step. The hidden representation $\mathbf{h}_i^{n_i}$ at the last time step n_i is our enriched representation of the longitudinal records $\hat{\mathbf{X}}_i$, such that $\mathbf{h}_i^{n_i} = \mathbf{z}_i \in \Re^{d_z}$.

$$\mathbf{h}_i^{n_i} = \mathbf{z}_i = \phi_E(\mathbf{X}_i, \mathbf{M}_i, \mathbf{t}_i; \theta_E). \quad (6)$$

Since the hidden representation at j-th time point aims to summarize the records from first time step to j-th time step, the LSTM cell needs to refer to the cell state $\mathbf{c}_i^j$ and reflect past records to $\mathbf{h}_i^j$. Since the cell state $\mathbf{c}_i^j$ is guided by the

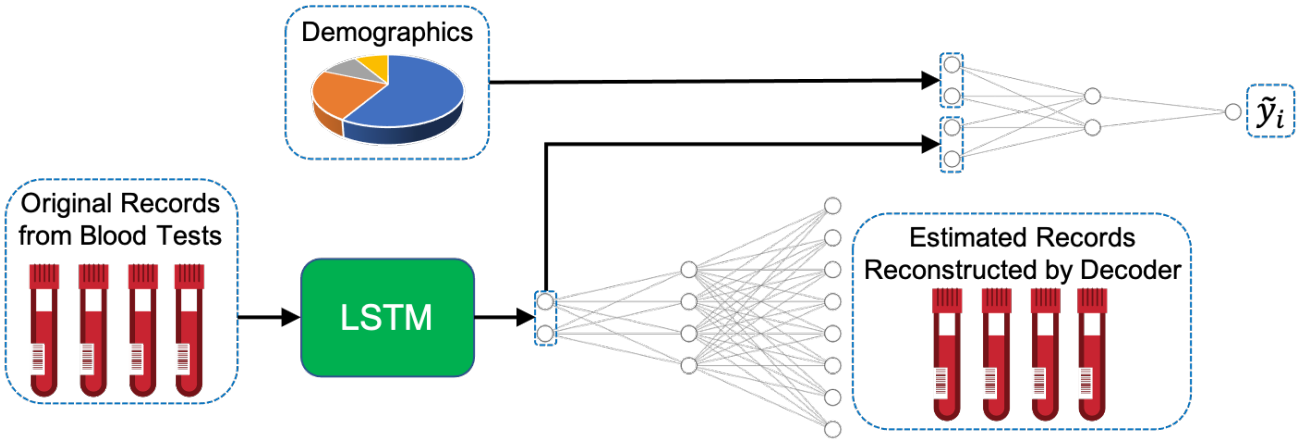

Figure 2: The workflows of the enrichment learning task. The enriched representation at the bottleneck of the LSTM autoencoder is used to reconstruct the longitudinal records and predict the mortality.

input gate $\mathbf{k}_i^j$ and forget gate $\mathbf{f}_i^j$, which control how much information came from previous step should be preserved, the cell state $\mathbf{c}_i^j$ enables the hidden representation $\mathbf{h}_i^j$ to learn long term dependencies. For example, LSTM encoder can capture the temporal trends of patient's status from the consecutive records.

2.4 Decoder and Predictor

From the enriched representation $\mathbf{z}_i$ of longitudinal records, the decoder reconstruct the original record, and predictor predicts the target label. The decoder and predictor both are the multilayer perceptron (MLP), instead of LSTM. A previous study [Srivastava *et al.*, 2015] that attempted to enrich longitudinal records with a recurrent neural network (RNN) [Medsker and Jain, 2001], did so by using RNNs for both the encoder and decoder, where the output (reconstructed record) of the decoder at each time step depends on the output at the previous time step. However, since no additional information is provided to the decoder other than a learned representation that is no longer longitudinal, there should not be dependency between the outputs of the decoder. Because the enriched representation $\mathbf{z}_i$ summarizes *whole* longitudinal records, the decoder $\phi_D : \Re^{d_z+1} \mapsto \Re^{D_l}$ should be able to reconstruct the j-th record $\mathbf{x}_i^j$ given time stamp t_i^j without any additional information, such that $\phi_D(\mathbf{z}_i, t_i^j;\ \theta_D) = \tilde{\mathbf{x}}_i^j \approx \mathbf{x}_i^j$, where θ_D is a set of weight matrices and bias vectors of the decoder. This autoencoder architecture, to the best of our knowledge, has not yet been proposed.

MLP consists of the consecutive hidden layers as follows:

$$\mathbf{h}_k = \sigma(\mathbf{W}_k \mathbf{h}_{k-1} + \mathbf{b}_k), \tag{7}$$

where $\mathbf{h}_k$ is the output of k-th hidden layer, and σ is the activation function, and $\mathbf{W}_k$, $\mathbf{b}_k$ are the trainable weights matrix and bias vector of k-th hidden layer. The output is earned by forwarding the input vector $\mathbf{h}_0$ to the last hidden layer. To recover the original record at the specific time point, the decoder needs to know that time point. Thus the input vector of decoder is the concatenation of enriched representation $\mathbf{z}_i$ and time stamp t_i^j, which is $[\mathbf{z}_i, t_i^j] \in \Re^{d_z+1}$. For predictor, the demographic information such as age may be crucial to predict target label. Therefore we provide the demographic information to the predictor with the enriched representation, which is $[\mathbf{z}_i, \mathbf{x}_i^b] \in \Re^{d_z+D_b}$. By forwarding each input of decoder and predictor to their own MLP, we earn the reconstructed record $\tilde{\mathbf{x}}_i^j$ and prediction $\tilde{y}_i$ on target label:

$$\phi_D(\mathbf{z}_i, t_i^j; \theta_D) = \tilde{\mathbf{x}}_i^j, \tag{8}$$

$$\phi_P(\mathbf{z}_i, \mathbf{x}_i^b; \theta_P) = \tilde{\mathbf{y}}_i, \tag{9}$$

and we have the stack of reconstructed records of i-th patient:

$$\tilde{\mathbf{X}}_i = [\tilde{\mathbf{x}}_i^1; \tilde{\mathbf{x}}_i^2; \cdots; \tilde{\mathbf{x}}_i^{n_i}]. \tag{10}$$

2.5 Loss Functions

AE is asked to accomplish the two tasks - reconstructing original records and predicting target label by minimizing:

$$\min_{\theta_E, \theta_D, \theta_P} \mathcal{L}_{total} = \min_{\theta_E, \theta_D, \theta_P} (\gamma_1 \mathcal{L}_{reconstruct} + \gamma_2 \mathcal{L}_{predict}), \tag{11}$$

where γ_1 and γ_2 are the hyperparameters to adjust the impact of each loss. The reconstruction loss is defined as the scaled Mean Squared Error (MSE):

$$\mathcal{L}_{reconstruct} = \frac{\left\| (\tilde{\mathbf{X}}_i - \mathbf{X}_i) \odot \mathbf{M}_i \right\|_F^2}{|\mathbf{M}_i|}, \tag{12}$$

where squared Frobenious norm $\|\cdot\|_F^2$ is defined as the summation of all the entries squared.

The the prediction loss is defined respect to the labeled $i \in \Omega$ and unlabeled $i \notin \Omega$ data separately:

$$\mathcal{L}_{predict} = \left\{ \begin{array}{ll} \|\tilde{y}_i - y_i\|_F^2, & \text{for } i \in \Omega \\ 0, & \text{for } i \notin \Omega \end{array} \right\}. \tag{13}$$

The high capacity unsupervised AE may suffer from the tendency to learn the trivial identity mapping and memorize the input [Srivastava *et al.*, 2015]. The addition of prediction loss can prevent this memorization problem, because the memorization is not useful to predict the target label.

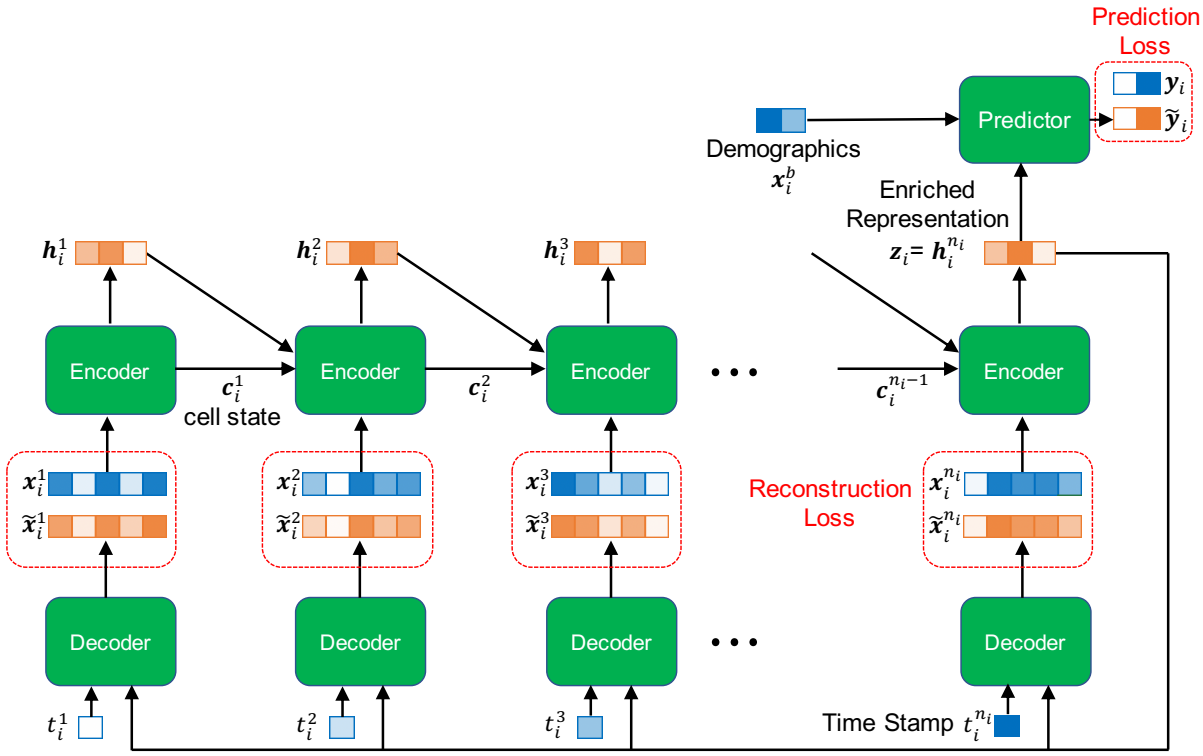

Figure 3: An illustration for the loss functions. The encoder consists of LSTM cell, which processes input record $\mathbf{x}_i^j$, and then generates the hidden representation $\mathbf{h}_i^j$ while conveying the cell state $\mathbf{c}_i^j$ to the next step. The enriched representation of the whole time series is defined as the hidden representation $\mathbf{h}_i^{n_i}$ at the last step.

3 Experiment

Our experiments consist of two parts; (1) we evaluate the prediction performance of the proposed model, and (2) we identify the biomarkers which are most predictive for mortality.

3.1 Hyperparameters and Competing Models

We use the following hyperparameters found by the grid search. For our semi-supervised autoencoder (SAE), the decoder ϕ_D has 3 fully connected layers with 200, 140, 100 nodes and hyperbolic tangent activation function except for the input layer and leaky Rectified Linear Unit (leaky ReLU, alpha = 0.1) for the input layer. Here we found that the leaky ReLU largely improves the reconstruction performance of decoder, as we presume that time stamp is the most important input feature for the decoder and this can be emphasized more with leaky ReLU activation function in a range $(-\infty, \infty)$, than the other activation functions in a smaller range. The predictor ϕ_P has 3 fully connected layers with 120, 60, 20 nodes and hyperbolic tangent activation function. The encoder ϕ_E has LSTM network with 60 units and hyperbolic tangent activation function. γ_1 and γ_2 in Eq. (11) are set to 0.005 and 0.1. To minimize the loss function in Eq. (11), we use Adam optimizer [Kingma and Ba, 2014] with learning rate of 0.0003 and the other parameters kept at their default values. We do not use any regularization or dropout technique, as they have not improved the performance. Considering our loss function is defined differently depending on the existence of label of sample, we train our model with the sample by sample instead of batch of samples. We build our model with Python 3.7 and Keras [Chollet and others, 2015] framework. We use MacOS with 3.4 GHz Quad-Core Intel Core i5 CPU and 16 GB DDR4 Ram and it took 4 hours to train our model with 286 samples and 200 iterations.

For an ablation study to observe the effectiveness of our semi-supervised enrichment learning, we introduce supervised baseline LSTM (BLSTM) model as a competing model by removing decoders ϕ_D from our model SAE but with the same hyperparameters as our model. In addition to the longitudinal model, we use the following competitive models:

- MLP with 5 fully connected layers of 200, 140, 100, 140, 200 nodes.
- Random Forest [Ho, 1995] (RF) with 34 max depth.
- Ridge Regression (RR) with regularization parameter of 1.0.
- Support Vector Machine (SVM) with regularization parameter of 1.0 and kernel function of radial basis function.

Since these competitive models are not longitudinal models, we provide the concatenation of most recent record $[\mathbf{x}_i^b, \mathbf{x}_i^{n_i}]$ to them. The training and test set both are provided to train SAE in a semi-supervised manner, while only the training set is provided to train the other competing models. Although the order of participants is randomly shuffled to avoid the bias,

we use the same training and test data across all the competing methods for the fair comparison.

3.2 Classification Performance

We conduct the classification task to predict the mortality of COVID-19 patients more than 10 days in advance and evaluate the performance of the predictive models. The outputs of the models are rounded up to the binary. We evaluate the predictive models with the following metrics:

$$\begin{aligned}
\text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN}, \\
\text{Precision} &= \frac{TP}{TP+FP}, \\
\text{Recall} &= \frac{TP}{TP+FN}, \\
\text{F}_1\text{-score} &= 2 \times \frac{Precision \times Recall}{Precision + Recall}.
\end{aligned} \tag{14}$$

In table 1, the average and standard deviation of the metrics across k subgroups are reported following k-fold cross validation scheme. k is set to 4 (the size of training set is 25% or 75% of the studied cohort) or 5 (the size of training set is 20% or 80% of the studied cohort).

The experimental results in Table 1 show that our model outperforms the other competitive models for all the different size of test sets. Interestingly, the performance gap between SAE and the other competitive models increases as the size of test set increases. We suppose that this is because our semi-supervised learning approach can learn from the unlabeled samples in test set, while the other competitive models cannot. This finding shows robustness of our model against the large proportion of unlabeled samples, as well as its promise in early prediction of mortality of COVID-19 patient.

Classification Performance with Subset of Biomarkers

The previous study [Yan *et al.*, 2020] has achieved the promising prediction performance on the mortality of COVID-19 patients with dataset same as ours. However, their model requires the following biomarkers are measured: lactic dehydrogenase, lymphocyte and hight-sensitivity C-reactive protein. These three biomarkers have been identified as the mortality relevant biomarkers in their study, and the inclusion of these biomarkers may overly simplify the classification task. To further validate the usefulness of our novel model, we perform the classification task on the dataset whose those three biomarkers are excluded and inspect whether our model can still predict the mortality successfully. From the experimental results listed in Table 2, we have found that the proposed model shows satisfiable prediction performance even if the three principle biomarkers are not given.

3.3 Biomarkers Identification

It is vital to identify the mortality relevant biomarkers to classify the patients in serious condition. We identifies the risk factors from the blood sample measures of COVID-19 patients. Despite of the high performance of deep learning models, their outputs are notoriously difficult to be interpreted. To identify which biomarkers (features) largely affect to the predictions, we add the perturbation to the input data and observe the changes in prediction.

For each q-th biomarker ($1 \leq q \leq D_l$) and i-th patient, we sample the column vector of perturbation $\mathbf{p}_{i,q} \in \Re^{n_i}$ from the normal distribution $\mathcal{N}(0, \sigma_q^2)$ with zero mean and same standard deviation σ_q as the observed distribution of q-th biomarker across all n_i time points and n patients, and then perturb the measurement of q-th biomarker as follows:

$$\begin{aligned}
N &= \sum_{i=1}^{n}\sum_{j=1}^{n_i} m_{i,q}^j, \; \mu_q = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i} m_{i,q}^j * x_{i,q}^j, \\
\sigma_q^2 &= \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i} m_{i,q}^j (x_{i,q}^j - \mu_q)^2,
\end{aligned} \tag{15}$$

where $x_{i,q}^j$ and $m_{i,q}^j$ denote an measurement of j-th time step and q-th biomarker of $\mathbf{X}_i$ and $\mathbf{M}_i$. Then the records whose m-th biomarker is perturbed and it's prediction change is:

$$\begin{aligned}
\mathbf{X}_i' &= [\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \cdots, \mathbf{x}_{i,q} + \mathbf{p}_{i,q}, \cdots, \mathbf{x}_{i,D_l}], \\
\Delta\tilde{\mathbf{y}}_i &= \|\phi_P(\phi_E(\mathbf{X}_i, \mathbf{M}_i, \mathbf{t}_i; \theta_E), \mathbf{x}_i^b; \theta_P) \\
&\quad - \phi_P(\phi_E(\mathbf{X}_i', \mathbf{M}_i, \mathbf{t}_i; \theta_E), \mathbf{x}_i^b; \theta_P)\|,
\end{aligned} \tag{16}$$

where $\mathbf{x}_{i,q} \in \Re^{n_i}$ is the column vector of biomarker measurements of i-th patient collected across all the time points. Then the relative importance of q-th biomarker is defined as the average of prediction changes over the all samples: $\frac{1}{n}\sum_{i=1}^{n} \Delta\tilde{\mathbf{y}}_i$. We plot top 15 most important biomarkers in Fig. 4.

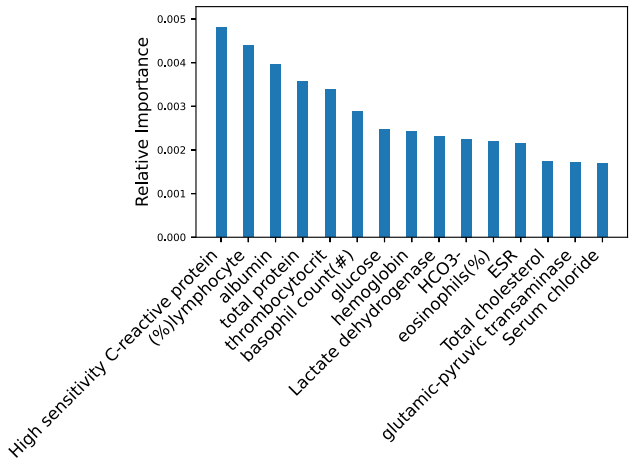

Figure 4: Top 15 important biomarkers identified by the proposed method.

The identified biomarkers have been shown in the literature to be related to the mortality of COVID-19 patient. For example, lactic dehydrogenase (LDH), lymphocyte and high-sensitivity C-reactive protein (hs-CRP) are the top 3 biomarkers relevant with the mortality of COVID-19 patients, identified by the XGBoost model [Yan *et al.*, 2020] and previous medical researches [Kishaba *et al.*, 2014; Ridker *et al.*, 2008;

Table 1: The prediction performance of SAE and the competitive models from k-fold cross validation. The best prediction is denoted as bold font.

Test Set Size	Model	Accuracy	Precision	Recall	F_1-score
20%	SAE	**94.14±2.31**	**93.35±2.45**	**93.15±2.84**	**93.48±1.21**
	BLSTM	92.56±3.24	91.11±1.45	92.45±8.12	91.43±5.40
	MLP	80.63±11.52	76.58±10.94	83.80±11.97	80.02±11.14
	RF	83.25±11.89	78.03±11.15	88.77±12.68	83.05±11.86
	RR	79.32±11.33	76.76±10.97	79.54±11.36	78.12±11.16
	SVM	79.32±11.33	77.55±11.08	78.12±11.16	77.83±11.11
25%	SAE	**92.47±3.43**	**91.84±8.41**	**91.97±2.34**	**91.63±3.94**
	BLSTM	91.91±3.69	91.38±8.94	91.62±1.89	91.17±3.98
	MLP	81.90±11.70	79.25±11.32	82.92±11.85	81.05±11.58
	RF	83.30±11.9	78.94±11.28	87.45±12.49	82.98±11.85
	RR	80.50±11.50	81.13±11.16	76.14±10.88	78.56±11.22
	SVM	80.50±8.12	80.11±11.45	77.65±11.09	78.86±11.27
75%	SAE	**89.48±1.59**	**88.59±2.27**	88.72±6.09	**88.46±2.1**
	BLSTM	87.91±1.07	87.41±3.22	86.07±4.59	86.57±1.4
	MLP	80.09±11.44	74.94±10.71	86.86±12.41	80.46±11.50
	RF	88.52±12.65	85.49±12.21	**91.32±13.0**	88.31±12.62
	RR	79.03±11.29	75.54±10.79	82.41±11.77	78.82±11.26
	SVM	77.98±11.11	75.06±10.7	80.18±11.45	77.54±11.08
80%	SAE	**88.06±1.07**	**86.98±1.82**	**87.03±3.29**	**86.95±1.44**
	BLSTM	84.07±5.55	86.55±2.91	77.60±16.00	80.66±9.94
	MLP	79.46±11.35	72.80±10.40	82.19±11.17	77.21±11.03
	RF	51.15±4.09	64.18±14.00	30.00±25.83	59.81±16.37
	RR	78.14±11.16	70.78±10.11	82.19±11.17	76.06±10.87
	SVM	74.16±10.79	66.22±9.46	79.03±11.3	72.06±10.29

Table 2: The prediction performance of SAE when the subset biomarkers is given.

Test Set Size	Accuracy	Precision	Recall	F_1-score
20%	91.07±2.04	91.27±1.63	89.18±4.65	90.09±1.95
25%	89.37±3.06	89.26±6.55	88.07±2.99	88.48±3.24
75%	87.81±1.94	87.30±5.06	86.83±6.72	86.68±2.24
80%	86.80±1.81	84.36±4.91	87.70±2.17	85.65±2.27

Wang *et al.*, 2020]. To be specific, the increase of LDH indicates tissue or cell destruction and this is the strong sign of tissue or cell damage [Yan *et al.*, 2020]. The activity of idiopathic pulmonary fibrosis can be detected by Serum LDH [Kishaba *et al.*, 2014]. The hs-CRP is the risk factor for the continuous inflammation [Bajwa *et al.*, 2009] and poor prognosis in acute respiratory distress syndrome [Kishaba *et al.*, 2014; Sharma *et al.*, 2016]. The lymphocyte is the common risk factor of COVID-19 patients [Chan *et al.*, 2020], and lymphocyte has relation with the decrease in CD4 and CD8 T cells [Liu *et al.*, 2020]. Albumin have been found to be independently associated with mortality, at the Cox regression analysis [Violi *et al.*, 2020]. The basophil count is known to be the risk factor of mortality and organ injury in COVID-19 patients [Li *et al.*, 2020]. The biomarkers identified by our model are in nice accordance with the previous studies, and provide the substantial evidence that our approach can identify the features associated with the prediction target.

4 Conclusion

We propose semi-supervised enrichment methods based on the novel LSTM autoencoder which fits to the clinical applicability and can be used in a real-time automatic mortality prediction. The enriched representation in the vectorial format summarizes the uneven and incomplete MTS data. Armed with the enriched representation, one can fully utilize the prediction capability of the MTS dataset. In our experiments, the proposed model shows the state-of-the-art prediction performance on mortality prediction task. In addition, combined with the perturbation based feature identification method, our model discovers the most predictive biomarkers, further adding value to our approach in the interpretability aspect. Our model is purely data-driven, which means the prediction performance can be further improved with additional data. Since there is no assumption or limitation in the dataset property, the other models stemming from our enrichment approach are able to perform the different prediction tasks with the various MTS dataset.

References

[Bajwa *et al.*, 2009] Ednan K Bajwa, Uzma A Khan, James L Januzzi, Michelle N Gong, B Taylor Thompson, and David C Christiani. Plasma c-reactive protein levels are associated with improved outcome in ards. *Chest*, 136(2):471–480, 2009.

[Chan *et al.*, 2020] Jasper Fuk-Woo Chan, Shuofeng Yuan, Kin-Hang Kok, Kelvin Kai-Wang To, Hin Chu, Jin Yang, Fanfan Xing, Jieling Liu, Cyril Chik-Yan Yip, Rosana Wing-Shan Poon, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, 395(10223):514–523, 2020.

[Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[Chollet and others, 2015] François Chollet et al. Keras. https://keras.io, 2015.

[for Disease Control *et al.*,] Centers for Disease Control, Prevention (CDC), et al. Strategies for optimizing the supply of n95 respirators. 2020 apri 3 https://www. cdc. gov/coronavirus/2019-ncov/hcp/respiratorsstrategy/index. html. *CDC_AA_refVal= https% 3A% 2F% 2Fwww. cdc. gov% 2Fcoro navirus% 2F2019-ncov% 2Fhcp% 2Frespiratorsstrategy% 2Fcrisis-alternate-strategies*.

[Ho, 1995] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kishaba *et al.*, 2014] Tomoo Kishaba, Hitoshi Tamaki, Yousuke Shimaoka, Hajime Fukuyama, and Shin Yamashiro. Staging of acute exacerbation in patients with idiopathic pulmonary fibrosis. *Lung*, 192(1):141–149, 2014.

[Längkvist *et al.*, 2014] Martin Längkvist, Lars Karlsson, and Amy Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42:11–24, 2014.

[Li *et al.*, 2020] Dongze Li, You Chen, Hong Liu, Yu Jia, Fanghui Li, Wei Wang, Jiang Wu, Zhi Wan, Yu Cao, and Rui Zeng. Immune dysfunction leads to mortality and organ injury in patients with covid-19 in china: insights from ers-covid-19 study. *Signal Transduction and Targeted Therapy*, 5(1):1–3, 2020.

[Liu *et al.*, 2020] Jing Liu, Sumeng Li, Jia Liu, Boyun Liang, Xiaobei Wang, Hua Wang, Wei Li, Qiaoxia Tong, Jianhua Yi, Lei Zhao, et al. Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of sars-cov-2 infected patients. *EBioMedicine*, page 102763, 2020.

[Lu *et al.*, 2018] Lyujian Lu, Hua Wang, Xiaohui Yao, Shannon Risacher, Andrew Saykin, and Li Shen. Predicting progressions of cognitive outcomes via high-order multimodal multi-task feature learning. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 545–548. IEEE, 2018.

[Medsker and Jain, 2001] Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5, 2001.

[Ridker *et al.*, 2008] Paul M Ridker, Eleanor Danielson, Francisco AH Fonseca, Jacques Genest, Antonio M Gotto Jr, John JP Kastelein, Wolfgang Koenig, Peter Libby, Alberto J Lorenzatti, Jean G MacFadyen, et al. Rosuvastatin to prevent vascular events in men and women with elevated c-reactive protein. *New England journal of medicine*, 359(21):2195–2207, 2008.

[Sharma *et al.*, 2016] Surendra K Sharma, Anunay Gupta, Ashutosh Biswas, Abhishek Sharma, Atul Malhotra, KT Prasad, Sreenivas Vishnubhatla, Sajal Ajmani, Hridesh Mishra, Manish Soneja, et al. Aetiology, outcomes & predictors of mortality in acute respiratory distress syndrome from a tertiary care centre in north india. *The Indian journal of medical research*, 143(6):782, 2016.

[Srivastava *et al.*, 2015] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.

[Violi *et al.*, 2020] Francesco Violi, Roberto Cangemi, Giulio Francesco Romiti, Giancarlo Ceccarelli, Alessandra Oliva, Francesco Alessandri, Matteo Pirro, Pasquale Pignatelli, Miriam Lichtner, Anna Carraro, et al. Is albumin predictor of mortality in covid-19? *Antioxidants and Redox Signaling*, (ja), 2020.

[Wang *et al.*, 2020] Dawei Wang, Bo Hu, Chang Hu, Fangfang Zhu, Xing Liu, Jing Zhang, Binbin Wang, Hui Xiang, Zhenshun Cheng, Yong Xiong, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china. *Jama*, 323(11):1061–1069, 2020.

[Yan *et al.*, 2020] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, et al. An interpretable mortality prediction model for covid-19 patients. *Nature Machine Intelligence*, pages 1–6, 2020.

[Yoon *et al.*, 2018] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Gain: Missing data imputation using generative adversarial nets. *arXiv preprint arXiv:1806.02920*, 2018.

[Yu *et al.*, 2019] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.