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

Hoon Seo, Ibrohim Nosirov, and Hua Wang

Abstract—An influx of COVID-19 infections puts extreme pressure on healthcare systems and the efficient allocation of finite resources becomes a crucial problem. Because clinical measurements of COVID-19 patients can be collected over time, many longitudinal learning models have been presented to predict the clinical outcomes from the multivariate time series (MTS) and thus aid in logistical decision making. However, they often fail to integrate static data with the MTS or utilize the available labels of samples. We propose a novel semi-supervised learning framework that leverages a Long Short-Term Memory (LSTM) autoencoder to learn the vectorial representation of MTS which can be easily integrated with the static data. Armed with vectorial representation summarizing the MTS and static data, conventional machine learning models can be used to predict clinical outcomes. In our experiments of two case studies, the proposed model shows promising prediction performance on mortality rates of 358 COVID-19 patients and 3997 ICU patients with their temporal records collected from blood tests. In addition, the proposed model identifies the risk factors of mortality.

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 [1]. In an earlier study, Yan et al. advocated for a machine learning mortality-prediction model to inform logistical planning [2]. The model [2] proposed in the study identifies the most predictive biomarkers for the patient's mortality using a XGBoost classifier [3] trained on a publicly available MTS dataset collected from COVID-19-positive patients admitted to Tongji Hospital in China. Acknowledging the limitations of the 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, the model is not able to predict the mortality when these principle biomarkers are not measured.

To learn the relationships across variables and time steps, MTS analysis models have be proposed. The samples from COVID-19 patients 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 often irregular. The resultant inputs are uneven sequences of vectors with missing data that is difficult to be integrated with the static data represented by a fixed length vector. However, conventional MTS analysis usually require sequences to be complete and of consistent length [4], [5], [6], [7], which is rarely the case when data is collected outside of a carefully controlled environment. In addition they often do not integrate static and dynamic data from different modalities to improve prediction.

To overcome the problems from missing data and an inconsistent number of records, recurrent neural networks (RNNs) [8], especially Long Short Term Memory (LSTM) [9], have been successfully applied in MTS analysis. Due to their flexibility in handling missing data and ability to learn long-term dependencies, they achieve state of the art results in supervised learning tasks involving MTS with large numbers of records. However, due to the fact that a significant portion of COVID-19 patient's records are collected in the field and are unlabeled, supervised learning models may not be efficient and practical tools in the prediction of clinical outcomes of COVID-19 patients.

Recent unsupervised learning models [10], [11], [12], [13] have leveraged RNNs to learn the representation of MTS in a lower dimensional space. A proper representation of MTS data is crucial because MTS usually contains noises and redundant information from the large number of features and records [11], [14]. However, existing RNN models encode MTS into the other longitudinal enriched representations which are difficult to be integrated with the static data, such as demographic or genetic information. Another disadvantage of existing unsupervised representation learning models is they often do not utilize labeled samples, when the target labels of samples can improve the representation learning for better predictions.

Therefore, we propose an alternate solution in the form of a novel LSTM autoencoder architecture to transform the incomplete MTS into a vectorial representation which can be readily integrated with the static data. Unlike the previous LSTM autoencoder models in [10], [11], [12], the proposed model encodes the MTS into a fixed length vec-

- *H. seo, I. Nosirove and H. Wang are with the Department of Computer Science, Colorado School of Mines, Golden, CO, 80401.*
- *This work was supported in part by the National Science Foundation (NSF) under the grants of IIS 1652943, IIS 1849359, CNS 1932482 and CCF 2029543. (Corresponding author: Hua Wang.)*
- *The codes for the proposed method are provided at: https://anonymous.4open.science/r/364af6e2-45d7-4d23-a3d5-ed5faa25ef97/*

tor which is no longer longitudinal. Considering that the data collected in the field usually consists of labeled and unlabeled samples, the proposed semi-supervised learning model utilizes both, improving predictions compared to the strictly supervised [2] or unsupervised [10], [11] learning models. In our experiments with two MTS datasets, we show that the learned representation helps improve predictions on mortality of patients, especially when only a small proportion of samples are labeled. The proposed prediction model will be useful for improving the speed of triage and resource distribution.

2 METHODS

In this section, we describe the structure of the proposed autoencoder 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 fully connected layers (FCL) of decoder ϕ_D decode the enriched representation into the original record. Finally, a conventional classifier ϕ_P predicts the target labels from the enriched representation. We outline the proposed model in Fig. 1.

2.1 Notations

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

- $\mathbf{x}_i^s \in \Re^{D_s}$ is a vector of static data.
- $\mathbf{X}_i = [\mathbf{x}_i^1; \mathbf{x}_i^2; \cdots; \mathbf{x}_i^{n_i}] \in \Re^{n_i \times D_l}$ are the longitudinal records collected from the blood tests across n_i time points. Here the number of available records n_i varies over the patients.
- $\mathbf{M}_i = [\mathbf{m}_i^1; \mathbf{m}_i^2; \cdots; \mathbf{m}_i^{n_i}] \in \{0,1\}^{n_i \times D_l}$ are binary masks of observabilities of longitudinal records $\mathbf{X}_i$, where 1 and 0 indicate 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 0. The target label $\mathbf{y}_i \in \{0,1\}^{D_y}$ is the mortality of the i-th patient, which is provided in the training process if that patient is in the training set, such that $i \in \Omega$.

2.2 Encoder

We leverage the LSTM encoder $\phi_E : \Re^{n_i \times (2D_l+1)} \mapsto \Re^{d_z}$ to summarize longitudinal records and learn the temporal relationships between records. The time stamp of each record is crucial in learning the temporal relation between records (e.g. temporal locality) especially when the time intervals between the records are uneven. The pattern of missing entries helps the encoder to correctly interpret the input data. 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 the 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 the LSTM.

For each time step ($1 \le j \le n_i$), the input record $\hat{\mathbf{x}}_i^j$ of the i-th patient is processed by following the LSTM architecture [15]:

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

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

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

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

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

where σ and tanh are the logistic sigmoid and hyperbolic tangent activation function respectively and $\mathbf{k}_i^j$, $\mathbf{o}_i^j$, $\mathbf{f}_i^j$ are the input, output, and forget gate of the 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 while $\mathbf{c}_i^j$ and $\mathbf{h}_i^j$ denote the cell state and hidden representation at the 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 $\mathbf{h}_i^{n_i} = \mathbf{z}_i \in \Re^{d_z}$.

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

Since the hidden representation at the j-th time point aims to summarize the records between the first and j-th time steps, the LSTM cell refers to the cell state $\mathbf{c}_i^j$ and reflects past records to $\mathbf{h}_i^j$. In Eq. (3) the cell state $\mathbf{c}_i^j$ is guided by the input gate $\mathbf{k}_i^j$ and forget gate $\mathbf{f}_i^j$, which control how much information should be preserved from the previous step, thus cell state $\mathbf{c}_i^j$ enables the hidden representation $\mathbf{h}_i^j$ to learn long term dependencies. For example, the input gate $\mathbf{k}_i^j$ and forget gate $\mathbf{f}_i^j$ can utilize the time stamp of each record to decide how much the hidden representation $\mathbf{h}_i^j$ should be updated. Therefore our encoder preserves the temporal locality of records with uneven time intervals while capturing the long term trends of a patient's status.

2.3 Decoder

From the enriched representation $\mathbf{z}_i$ of the longitudinal records, the decoder reconstructs the original record. We propose a decoder for dynamic data enrichment with a fully connected layers (FCL) architecture instead of another LSTM. Previous studies [10], [12], [16] that attempted to enrich longitudinal records with a recurrent neural network (RNN) [8], 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 a dependency between the outputs of the decoder. The enriched representation $\mathbf{z}_i$ summarizes *whole* longitudinal records, thus 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 the 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.

FCL consist of consecutive hidden layers as follows:

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

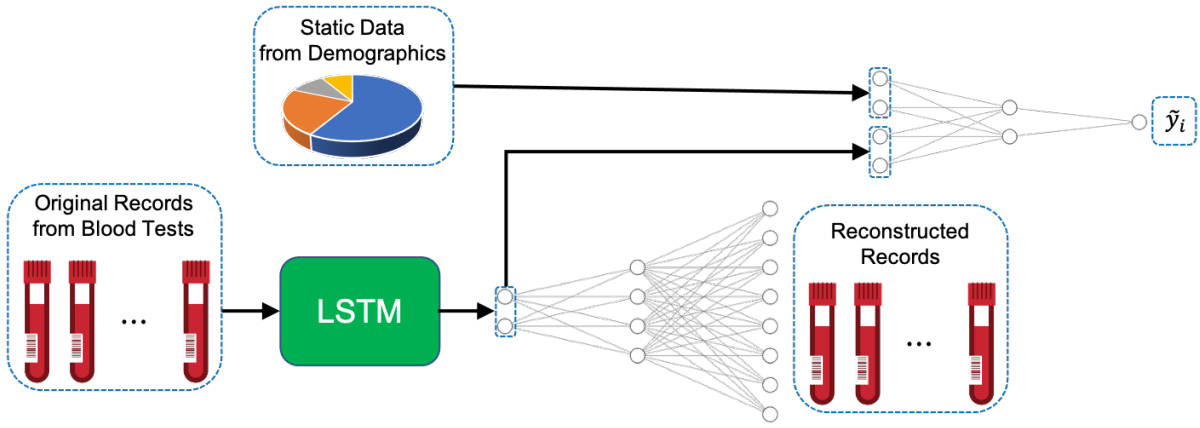

Fig. 1. The structure of the enrichment learning model. The enriched representation of dynamic records is in a fixed-length vector format which can be readily integrated with the static data.

where $\mathbf{h}_k$ is the output of the k-th hidden layer, σ is the activation function, and $\mathbf{W}_k$, $\mathbf{b}_k$ are the trainable weights matrix and bias vector of the k-th 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 the decoder is the concatenation of the enriched representation $\mathbf{z}_i$ and the time stamp t_i^j, which is $[\mathbf{z}_i, t_i^j] \in \Re^{d_z+1}$. By forwarding the input to the decoder's FCL, we can generate the reconstructed record $\phi_D(\mathbf{z}_i, t_i^j; \theta_D) = \tilde{\mathbf{x}}_i^j$, and we have the stack of reconstructed records of the i-th participant:

$$\tilde{\mathbf{X}}_i = [\tilde{\mathbf{x}}_i^1; \tilde{\mathbf{x}}_i^2; \cdots; \tilde{\mathbf{x}}_i^{n_i}]. \tag{8}$$

2.4 Prediction

Because the learned representation of the dynamic data is in a fixed-length vector format, conventional classifiers can be connected to predict the target label. Because static data such as age may be crucial to predicting the target label, we provide the static data $\mathbf{x}_i^s$ with the learned representation $\mathbf{z}_i$ as an input $\mathbf{x}_i^P = [\mathbf{z}_i, \mathbf{x}_i^s] \in \Re^{d_z+D_s}$ of classifier ϕ_P:

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

In this study we have chosen a Support Vector Machine (SVM) or FCL as a classifier. The support vector machine [17] leverages a kernel matrix to learn the nonlinear relationship between input and output. However, a medical dataset typically consists of many samples and computing the kernel matrix requires large resources. In addition, all the enriched representations must be prepared in advance to calculate the kernel matrix during the optimization process. Our semi-supervised learning model aims to optimize the prediction error and reconstruction error simultaneously for each batch of inputs. To tackle this problem, the previous study [18] proposed the random features mapping $\phi_F : \Re^{d_z+D_s} \mapsto \Re^{d_f}$ which approximates shift-invariant kernels efficiently. The SVM classifier first applies nonlinear transformation $\phi_F : \Re^{d_z+D_s} \mapsto \Re^{d_f}$ to the enriched representation and then trains the linear model on top of the transformed features:

$$\phi_F(\mathbf{x}_i^P)\mathbf{W}_P + \mathbf{b}_P = \tilde{\mathbf{y}}_i, \tag{10}$$

where $\mathbf{W}_P$, $\mathbf{b}_P \in \theta_P$ are the weights matrix and biases vector for the linear mapping from random feature to prediction.

2.5 Loss Functions

The Semi-supervised autoencoder accomplishes two tasks: reconstructing the original records and predicting the 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 prediction loss is defined with respect to the labeled $i \in \Omega$ and unlabeled $i \notin \Omega$ data separately:

$$\mathcal{L}_{predict} = \left\{ \begin{array}{ll} -\alpha_c H(\tilde{\mathbf{y}}_i, \mathbf{y}_i), & \text{for } i \in \Omega \\ 0, & \text{for } i \notin \Omega \end{array} \right\}. \tag{13}$$

For the FCL predictor, $H(\tilde{\mathbf{y}}_i, \mathbf{y}_i)$ is defined as binary cross-entropy loss $\|\mathbf{y}_i \odot \log(\tilde{\mathbf{y}}_i) + (\mathbf{1} - \mathbf{y}_i) \odot \log(\mathbf{1} - \tilde{\mathbf{y}}_i)\|_1$ and $\mathbf{1}$ is a vector of 1's and log is an element-wise logarithm function. For the SVM predictor, $H(\tilde{\mathbf{y}}_i, \mathbf{y}_i)$ is defined as squared Hinge loss $(\max(1 - \mathbf{y}_i \cdot \tilde{\mathbf{y}}_i^T), 0)^2$.

Here we introduce the weighing factor $\alpha_c \in [0, 1]$ to alleviate the unbalanced distribution of labels when c is the class of $\mathbf{y}_i$. For the unbalanced dataset, the target label has more observations in one specific class (majority class) than others (minority classes) and the prediction model may tend to classify most instances as the majority class to minimize the prediction error. To tackle this problem, we increase the prediction error for the minority classes by defining $\alpha_c = 1 - \frac{|\{j \in \mathbb{N} | y_j = c,\ j \leq N\}|}{N}$ for N samples.

The high capacity unsupervised autoencoder may suffer from the tendency to learn trivial identity mapping and memorize the input [10] which is not useful for predicting

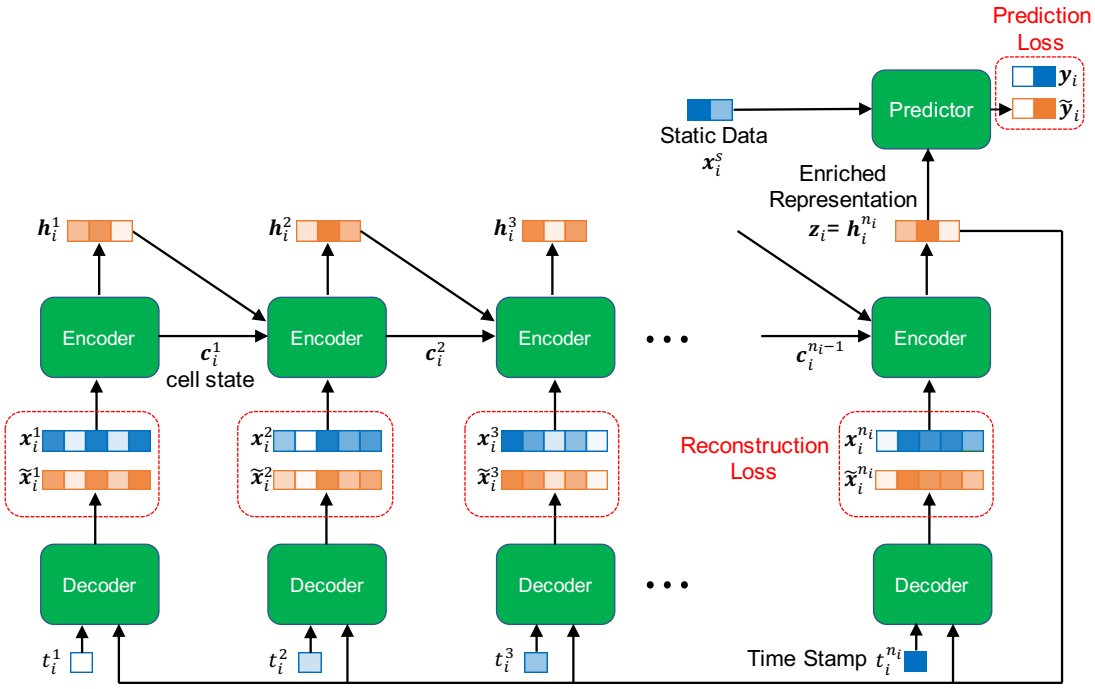

Fig. 2. An illustration for the loss functions. The encoder consists of LSTM cells, which process the input record $\mathbf{x}_i^j$, and then generate the hidden representation $\mathbf{h}_i^j$ while conveying the cell state $\mathbf{c}_i^j$ to the next step. The enriched representation of the whole time series is defined as the hidden representation $\mathbf{h}_i^{n_i}$ at the last step.

the target label. However, the addition of our prediction loss function can prevent this memorization problem.

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. We conduct the experiments on two case studies from COVID-19 patients and intensive care unit (ICU) patients.

3.1 Hyperparameters and Competitive Models

We use the following hyperparameters of our semi-supervised autoencoder with a FCL or SVM classifier (SA-FCL, SA-SVM) found by the grid search. The accuracy on the test set is used as a criterion for selecting hyperparameters. The decoder ϕ_D has 3 fully connected layers with 200, 140, 100 nodes with a leaky Rectified Linear Unit (leaky ReLU, alpha = 0.1). Here we found that the leaky ReLU largely improves the reconstruction performance of the decoder, as we presume that the 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 encoder ϕ_E has a LSTM network with 60 units (thus the dimensionality of enriched representation $d_z = 60$) and a hyperbolic tangent activation function. γ_1 and γ_2 in Eq. (11) are set to 0.005 and 0.1. The FCL classifier has 3 fully connected layers. The first two (with 120 and 60 nodes respectively) utilize the leaky ReLU activation function while the third (20 nodes) uses the soft-max activation function. The SVM classifier first transforms the enriched representations into 400 random features approximating the Gaussian kernel $k(\mathbf{x}, \mathbf{y}) \approx e^{-\frac{(\mathbf{x}-\mathbf{y})^2}{2\gamma_k^2}}$. The scale factor γ_k is set to 10.

To minimize the loss function in Eq. (11), we use the Adam optimizer [19] with a learning rate of 0.0003 and the other parameters kept at their default values. We do not use regularization or dropout techniques, as they have not shown to improve the performance. Our model is built with Python 3.7 and Keras [20] framework and was tested using MacOS with a 3.4 GHz Quad-Core Intel Core i5 CPU and 16 GB DDR4 Ram. It took 2 and 9 hours to train the proposed model with 375 COVID-19 patients (200 iterations) and 3997 ICU patients (70 iterations).

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

- Deep Neural Network (DNN) of 5 fully connected layers with 150, 125, 100, 50, 25 nodes and ReLU activation function.
- Random Forest [21] (RF) with 34 max depth and 100 trees.
- Ridge Classifier (RC) with regularization parameter of 1.0.
- Support Vector Machine (SVM) with regularization parameter of 1.0 and radial basis kernel function.

Since these competitive models are not longitudinal models, we provide the concatenation of the most recent record $[\mathbf{x}_i^s, \mathbf{x}_i^{n_i}]$ to them. The training and test set are both provided to train SA 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 a fair comparison.

3.2 Predictions on Mortality of COVID-19 patients

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. We obtain the blood sample records (74 features), demographic information (age and gender), and associated mortality outcomes of 375 patients collected throughout their stay in Tongji hospital between January 10th and February 24th, 2020 following the previous research [2]. We discard 17 samples where no time stamp was recorded. Among the remaining 358 samples, 192 patients survived and 166 patients died. The proportion of observed entries is 13.4%. The order of samples is randomly shuffled to prevent bias. Each feature of the dataset is normalized by min-max scaling to the range [0, 1]. The outputs of the models are rounded up to the binary. We evaluate the predictive models with the following metrics:

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

Since the precision and recall can be different when we predict fatality and survival respectively, we calculate and report the mean of the two cases. In table 1 and table 3, the average and standard deviation of the metrics across k subgroups are reported following a k-fold cross validation scheme. k is set to 4 (the size of test set is 25% or 75% of the studied cohort) or 5 (the size of test set is 20% or 80% of the studied cohort).

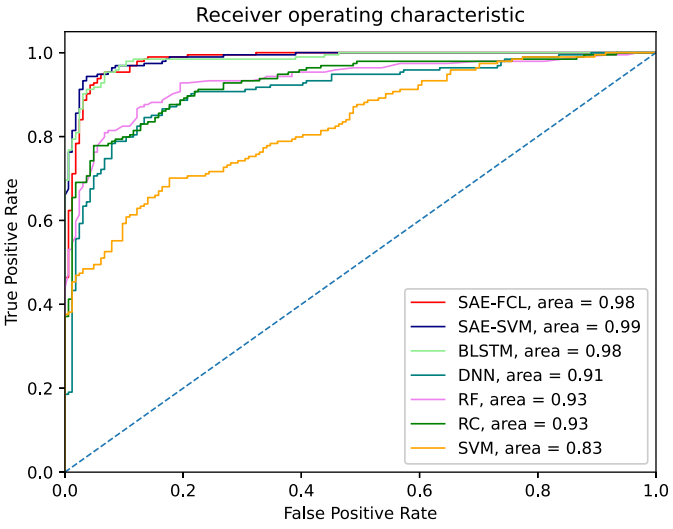

Fig. 3. Receiver operating characteristic curves (ROC) and their area under the curve (AUC) of predictions on mortality of COVID-19 patients when proportion of test set is 20%.

TABLE 1
The prediction performance on COVID-19 patients. The best prediction is highlighted bold.

Test Set	Model	Accuracy	Precision	Recall	F_1-score
20%	SA-FCL	94.14±2.31	93.35±2.45	93.15±2.84	93.48±1.21
	SA-SVM	**94.15±1.21**	**94.11±1.18**	**94.69±1.38**	**94.13±1.08**
	BLSTM	92.56±3.24	91.11±1.45	92.45±8.12	91.43±5.40
	DNN	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
	RC	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%	SA-FCL	92.47±3.43	91.84±8.41	91.97±2.34	91.63±3.94
	SA-SVM	**92.69±3.02**	**92.2±2.25**	**92.31±2.09**	**92.15±1.54**
	BLSTM	91.91±3.69	91.38±8.94	91.62±1.89	91.17±3.98
	DNN	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
	RC	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%	SA-FCL	**89.48±1.59**	88.59±2.27	88.72±6.09	88.46±2.1
	SA-SVM	89.36±1.28	**88.69±1.81**	89.04±3.15	**88.62±1.94**
	BLSTM	87.91±1.07	87.41±3.22	86.07±4.59	86.57±1.4
	DNN	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
	RC	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%	SA-FCL	**88.06±1.07**	**86.98±1.69**	**87.03±3.18**	**86.95±1.44**
	SA-SVM	87.35±1.25	85.04±1.82	86.03±3.29	85.49±1.95
	BLSTM	84.07±5.55	86.55±2.91	77.60±16.00	80.66±9.94
	DNN	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
	RC	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

In the experimental results in Table 1 and Fig. 3, our model SA-FCL and SA-SVM are compared with the five competitive models. As shown by the results, our model outperforms the others for all different sizes of test sets. The performance gap between SA and the other competitive models even increases as the size of the test set grows. We suspect this is due to our semi-supervised learning approach that allows our model to learn from unlabeled samples while still fully utilizing the benefits of labeled samples. In addition, the predictions of the proposed model are more stable (with a smaller standard deviation) when compared to competing models. This gives our model increased flexibility and a competitive advantage in performance. This finding emphasizes the robustness of our model against large proportions of unlabeled samples, as well as its promise in early prediction of mortality.

3.2.1 Classification Performance with Subset of Biomarkers

The previous study [2] has achieved the promising prediction performance on the mortality of COVID-19 patients with dataset same as ours. However, their model requires that 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 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 acceptable prediction performance even if the three principle biomarkers are not given.

TABLE 2
The prediction performance of SA-FCL when the subset biomarkers is given.

Test Set	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

3.2.2 Risk Factors of Mortality of COVID-19 patients

It is vital to identify the mortality relevant biomarkers to predict the course of disease at diagnosis. We identify the risk factors from the blood sample measures of COVID-19 patients. Despite the high performance of deep learning models, their outputs are notoriously difficult to interpret. 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 the same standard deviation σ_q as the observed distribution of the q-th biomarker across all n_i time points and n patients, and then perturb the measurement of q-th biomarker as follows:

$$N_q = \sum_{i=1}^{n}\sum_{j=1}^{n_i} m_{i,q}^j,\ \mu_q = \frac{1}{N_q}\sum_{i=1}^{n}\sum_{j=1}^{n_i} m_{i,q}^j \cdot x_{i,q}^j,$$
$$\sigma_q^2 = \frac{1}{N_q}\sum_{i=1}^{n}\sum_{j=1}^{n_i} m_{i,q}^j (x_{i,q}^j - \mu_q)^2, \quad (15)$$

where $x_{i,q}^j$ and $m_{i,q}^j$ denote a 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:

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

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 the q-th biomarker is defined as the average of prediction changes over the all samples: $\frac{1}{n}\sum_{i=1}^{n}(\Delta\tilde{\mathbf{y}}_i / \sum_{j=1}^{n_i} m_{i,q}^j)$. By dividing the changes in predictions by the number of observations $\sum_{j=1}^{n_i} m_{i,q}^j$, we can prevent feature importance from being exaggerated by the large number of observations. We plot top 15 risk factors of mortality in Fig. 4.

The identified biomarkers have been shown in the literature to be related to the mortality of COVID-19 patients. 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 [2] and previous medical researches [22], [23], [24]. To be specific, the increase of LDH indicates tissue or cell destruction and this is the strong sign of tissue or cell damage [2]. The activity of idiopathic pulmonary fibrosis can be detected by Serum LDH [22]. The hs-CRP is the risk factor for the continuous inflammation [25] and poor prognosis in acute respiratory distress syndrome [22], [26]. The lymphocyte is the common risk factor of COVID-19 patients [27], and lymphocyte has relation with the decrease in CD4 and CD8 T cells [28]. Albumin have been found to be independently associated with mortality, at the Cox regression analysis [29]. The basophil count is known to be the risk factor of mortality and organ injury in COVID-19 patients [30].

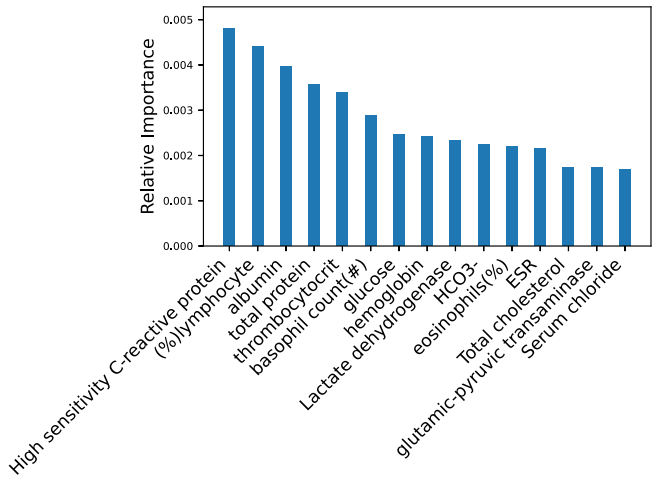

Fig. 4. Top 15 important biomarkers in blood samples of COVID-19 patients.

3.3 Prediction on Mortality of Intensive Care Unit Patients

Since the COVID-19 pandemic started less than two years ago, large-scale data of COVID-19 patients have not yet been accumulated. To evaluate the prediction performance on a larger dataset, we have downloaded time series measurements from blood tests (37 features), demographic information and in-hospital mortality of 4000 patients admitted to the Intensive Care Unit (ICU) from PhysioNet Challenge 2012 [31] and discarded 3 patients of no record. The detailed description on this dataset can be found in PhysioNet (https://physionet.org/content/challenge-2012/1.0.0/). Age, gender, height, and ICU type (Coronary Care Unit, Cardiac Surgery Recovery Unit, Medical ICU, and Surgical ICU) are provided as static data. The proportion of observed entries is 18%. The overall mortality rate is 13.9% and prediction with this unbalanced dataset can be challenging. We use the same hyperparameters used in COVID-19 patients dataset, except the followings:

- LSTM with 30 units, $\gamma_1 = 1e-7$, and $\gamma_2 = 10$ in SA-FCL, SA-SVM, and BLSTM.
- Deep Neural Network of 3 fully connected layers with 100, 50, 25 nodes and ReLU activation function.
- Random Forest with 20 max depth and 70 trees.
- Ridge Classifier with regularization parameter of 0.5.

As shown in Table 3 and Fig. 5, the overall performances are decreased in all models compared to COVID-19

TABLE 3
The prediction performance on mortality of ICU patients. The best prediction is highlighted in bold.

Test Set	Model	Accuracy	Precision	Recall	F_1-score
20%	SA-FCL	86.39±2.23	74.36±1.14	**61.69±1.37**	**68.43±1.38**
	SA-SVM	**86.87±2.68**	**77.45±1.05**	59.14±1.26	68.06±1.16
	BLSTM	78.95±4.75	74.51±3.64	54.92±2.86	61.42±2.71
	DNN	76.09±3.54	60.5±4.69	54.83±3.86	55.83±3.82
	RF	80.27±3.54	64.12±2.61	58.15±2.27	59.18±2.72
	RC	79.64±8.84	57.79±6.8	61.35±4.29	59.14±5.25
	SVM	50.17±9.33	50.22±7.64	55.11±9.64	51.76±8.24
25%	SA-FCL	85.6±3.15	73.92±4.4	**60.5±2.47**	65.29±3.65
	SA-SVM	**85.94±2.11**	**75.92±1.25**	59.5±1.35	**65.91±2.17**
	BLSTM	74.42±3.69	70.45±5.2	51.6±2.1	59.19±3.34
	DNN	72.42±5.34	57.11±5.81	51.2±3.02	53.25±5.45
	RF	77.19±5.12	61.27±4.35	56.65±4.15	58.67±3.28
	RC	76.29±7.2	55.09±4.7	59.25±6.84	56.2±5.84
	SVM	48.2±5.45	49.87±4.84	53.64±5.88	50.08±5.94
75%	SA-FCL	**83.09±2.94**	70.61±3.15	56.34±2.61	62.46±2.98
	SA-SVM	82.56±1.28	**71.11±2.48**	**56.81±1.81**	**62.81±1.73**
	BLSTM	69.13±3.25	67.19±4.85	47.15±6.61	50.38±5.59
	DNN	62.54±4.05	51.14±5.25	47.35±8.1	48.24±6.05
	RF	67.41±5.21	56.29±5.04	52.21±5.42	53.29±5.45
	RC	66.9±11.45	51.94±10.18	53.15±7.2	51.61±8.92
	SVM	46.15±4.8	45.7±5.25	49.65±3.8	46.34±4.15
80%	SA-FCL	**82.08±3.41**	69.97±3.92	**56.08±2.61**	**61.95±3.24**
	SA-SVM	81.82±2.89	**70.14±2.46**	55.91±1.93	61.13±2.19
	BLSTM	68.91±5.19	68.34±4.11	47.04±8.85	55.14±6.53
	DNN	61.31±4.8	51.01±6.24	46.13±4.15	48.7±4.11
	RF	71.14±2.19	56.25±5.95	51.9±3.71	53.14±5.89
	RC	65.95±5.62	50.15±9.45	51.48±8.1	51.14±9.42
	SVM	45.25±4.19	44.3±5.85	48.65±4.15	45.24±5.91

case study and we presume that this is because the target labels (mortality rate is 13.9%) are unbalanced. However the proposed model SA outperforms the competing models especially when the proportion of labeled samples is small, and the performance gaps are larger when compared to the experimental results from COVID-19 patients cohort. We presume that this is because the proposed model focuses on learning how to enrich the records of major cases (survived patients) to minimize the reconstruction error. As a result, the enriched representation of minor cases (died patients) can be quite different compared to major cases, therefore the classifier detects this differences and the enrichment approach can further improve the prediction. These results show that the proposed model can be successfully applied to the unbalanced and larger dataset.

3.3.1 *Risk Factors of Mortality of ICU patients*

We also identify the most predictive biomarkers in mortality of ICU patients: Heart rate (HR), Invasive diastolic arterial blood pressure (DiasABP), and Non-invasive systolic arterial blood pressure (NISysABP). Based on the previous studies, heart rate is associated with the all-cause and cardiovascular mortality [32] and blood pressure control is important issue for minimizing the risk of mortality [33]. The low diastolic blood pressure is risk factor of mortality in systolic heart failure [34]. The biomarkers identified by our model in two case studies 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.

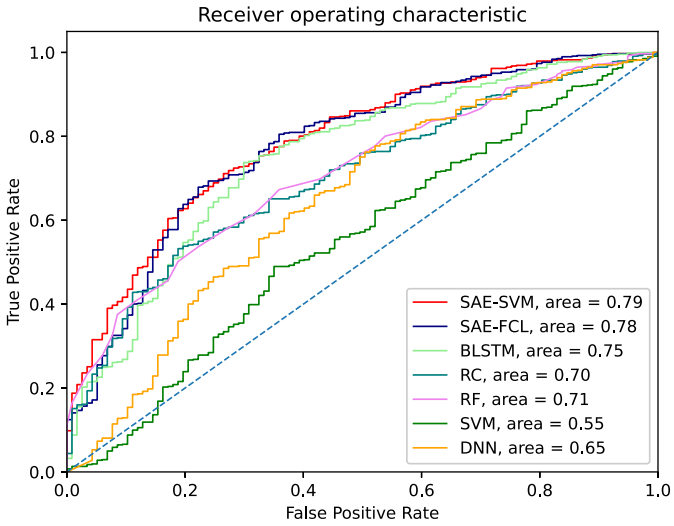

Fig. 5. ROC and AUC of predictions on mortality of ICU patient when proportion of test set is 20%.

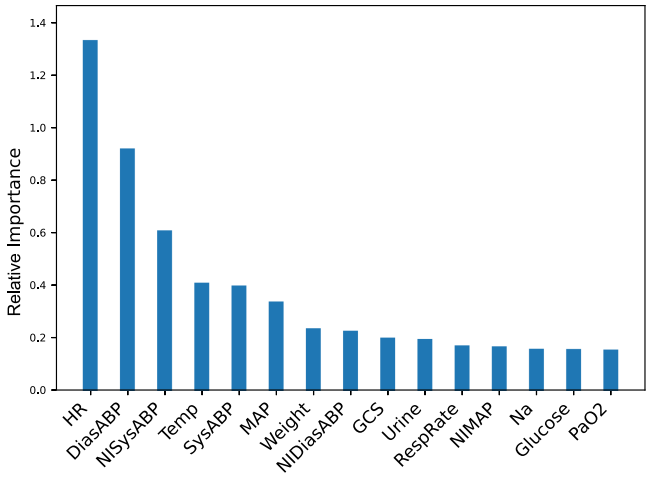

Fig. 6. Top 15 important biomarkers in blood samples records of ICU patients.

4 Conclusion

We propose a semi-supervised enrichment method based on a novel LSTM autoencoder that is clinicaly applicable and can make real-time automatic mortality predictions. The enriched representation of MTS data is in a fixed-length vector format and can be readily integrated with the static data. In our experiments and case studies, the proposed model shows state-of-the-art performance in predicting mortality as well as increased flexability in handeling labeled and unlabled data. Additionly, when combined with the perturbation based feature identification method, our model identifies the risk factors of mortality that are consistant with the findings of previous medical studies and predictive models. Since there is no assumption or limitation in the dataset property, this research proposes the general framework to fully utilize the MTS dataset, and other models stemming from our enrichment approach are able to perform different prediction tasks.

REFERENCES

[1] C. for Disease Control, P. (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% 2Fcoronavirus% 2F2019-ncov% 2Fhcp% 2Frespiratorsstrategy% 2Fcrisis-alternate-strategies*.

[2] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang *et al.*, "An interpretable mortality prediction model for covid-19 patients," *Nature Machine Intelligence*, pp. 1–6, 2020.

[3] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[4] L. Lu, H. Wang, X. Yao, S. Risacher, A. Saykin, and L. 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)*. IEEE, 2018, pp. 545–548.

[5] X. Wang, J. Yan, X. Yao, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, H. Huang *et al.*, "Longitudinal genotype-phenotype association study via temporal structure auto-learning predictive model," in *International Conference on Research in Computational Molecular Biology*. Springer, 2017, pp. 287–302.

[6] X. Wang, D. Shen, and H. Huang, "Prediction of memory impairment with mri data: a longitudinal study of alzheimer's disease," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 273–281.

[7] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, and L. Shen, "High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction," in *Advances in neural information processing systems*, 2012, pp. 1277–1285.

[8] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, 2001.

[9] J. Schmidhuber and S. Hochreiter, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International conference on machine learning*, 2015, pp. 843–852.

[11] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognition Letters*, vol. 42, pp. 11–24, 2014.

[12] S. Saumya and J. P. Singh, "Spam review detection using lstm autoencoder: an unsupervised approach," *Electronic Commerce Research*, pp. 1–21, 2020.

[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[14] K. S. Tuncel and M. G. Baydogan, "Autoregressive forests for multivariate time series modeling," *Pattern recognition*, vol. 73, pp. 202–215, 2018.

[15] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[16] A. Sagheer and M. Kotb, "Unsupervised pre-training of a deep lstm-based stacked autoencoder for multivariate time series forecasting problems," *Scientific reports*, vol. 9, no. 1, pp. 1–16, 2019.

[17] V. Vapnik, "The support vector method of function estimation," in *Nonlinear modeling*. Springer, 1998, pp. 55–85.

[18] A. Rahimi, B. Recht *et al.*, "Random features for large-scale kernel machines." in *NIPS*, vol. 3, no. 4. Citeseer, 2007, p. 5.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[21] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[22] T. Kishaba, H. Tamaki, Y. Shimaoka, H. Fukuyama, and S. Yamashiro, "Staging of acute exacerbation in patients with idiopathic pulmonary fibrosis," *Lung*, vol. 192, no. 1, pp. 141–149, 2014.

[23] P. M. Ridker, E. Danielson, F. A. Fonseca, J. Genest, A. M. Gotto Jr, J. J. Kastelein, W. Koenig, P. Libby, A. J. Lorenzatti, J. G. MacFadyen *et al.*, "Rosuvastatin to prevent vascular events in men and women with elevated c-reactive protein," *New England journal of medicine*, vol. 359, no. 21, pp. 2195–2207, 2008.

[24] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong *et al.*, "Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china," *Jama*, vol. 323, no. 11, pp. 1061–1069, 2020.

[25] E. K. Bajwa, U. A. Khan, J. L. Januzzi, M. N. Gong, B. T. Thompson, and D. C. Christiani, "Plasma c-reactive protein levels are associated with improved outcome in ards," *Chest*, vol. 136, no. 2, pp. 471–480, 2009.

[26] S. K. Sharma, A. Gupta, A. Biswas, A. Sharma, A. Malhotra, K. Prasad, S. Vishnubhatla, S. Ajmani, H. Mishra, M. 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*, vol. 143, no. 6, p. 782, 2016.

[27] J. F.-W. Chan, S. Yuan, K.-H. Kok, K. K.-W. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C.-Y. Yip, R. W.-S. 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*, vol. 395, no. 10223, pp. 514–523, 2020.

[28] J. Liu, S. Li, J. Liu, B. Liang, X. Wang, H. Wang, W. Li, Q. Tong, J. Yi, L. Zhao *et al.*, "Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of sars-cov-2 infected patients," *EBioMedicine*, p. 102763, 2020.

[29] F. Violi, R. Cangemi, G. F. Romiti, G. Ceccarelli, A. Oliva, F. Alessandri, M. Pirro, P. Pignatelli, M. Lichtner, A. Carraro *et al.*, "Is albumin predictor of mortality in covid-19?" *Antioxidants and Redox Signaling*, no. ja, 2020.

[30] D. Li, Y. Chen, H. Liu, Y. Jia, F. Li, W. Wang, J. Wu, Z. Wan, Y. Cao, and R. 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*, vol. 5, no. 1, pp. 1–3, 2020.

[31] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark, "Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012," *Computing in cardiology*, vol. 39, p. 245, 2012.

[32] D. Zhang, X. Shen, and X. Qi, "Resting heart rate and all-cause and cardiovascular mortality in the general population: a meta-analysis," *Cmaj*, vol. 188, no. 3, pp. E53–E63, 2016.

[33] M.-C. Wei, E. Kornelius, Y.-H. Chou, Y.-S. Yang, J.-Y. Huang, and C.-N. Huang, "Optimal initial blood pressure in intensive care unit patients with non-traumatic intracranial hemorrhage," *International journal of environmental research and public health*, vol. 17, no. 10, p. 3436, 2020.

[34] S. Javaheri, R. Shukla, H. Zeigler, and L. Wexler, "Central sleep apnea, right ventricular dysfunction, and low diastolic blood pressure are predictors of mortality in systolic heart failure," *Journal of the American College of Cardiology*, vol. 49, no. 20, pp. 2028–2034, 2007.