Generalizable Neural Network Representations of Patient State in the Intensive Care Unit

by Maryann M. Gong

S.B., C.S. M.I.T., 2017

Submitted to the
Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

February 2018

Author: _____
Department of Electrical Engineering and Computer Science
February 21, 2018

Certified by: _____
John Guttag, Professor, Thesis Supervisor
February 21, 2018

Certified by: _____
Jen Gong, PhD Candidate, Thesis Co-Supervisor
February 21, 2018

Accepted by: _____
Christopher Terman, Chairman, Masters of Engineering Thesis Committee

Generalizable Neural Network Representations of Patient State in the Intensive Care Unit
by Maryann M. Gong
Submitted to the Department of Electrical Engineering and Computer Science

February 2, 2018
In Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science

Abstract

Understanding changes in physiology in patients in the Intensive Care Unit (ICU) is important in determining care decisions. Machine learning algorithms have been used to model patient physiology to predict patient outcomes and administration of interventions. These predictions can be made directly on the raw patient data extracted from electronic health records. However, this data can be high dimensional with extraneous information. Neural networks, and in particular, autoencoders and sequence-to-sequence models, can be used to extract the important attributes of this time-series data without manual feature selection. In this work, we explore how learned encoded representations of physiological time-series and events time-series can be used to effectively predict outcomes on a variety of tasks. We compare the representations extracted from sequence-to-sequence models with representations extracted from autoencoders. We evaluate these representations on the task of predicting patient mortality and first onset of ventilator and vasopressor interventions. Our best representations achieve AUCs of 0.83, 0.91, and 0.91 on the tasks of mortality prediction, vasopressor first onset prediction, and ventilator first onset prediction, which is comparable to the performances using the raw features.

Contents

List of Figures

List of Tables

1. Problem Statement

The increased use of health monitoring tools and electronic health records (EHR) at hospitals enables the collection of a wealth of patient data (Jamoom et al. (2016)). Specifically, the extensive monitoring of patients and administration of lab tests and measurements in the intensive care unit (ICU) yields a large amount of data that could be used to gain further insight into a patient's physiological state. Earlier identification and treatment of ICU patients at risk or in need of medical interventions has been shown to improve ICU patient outcomes (Buist et al. (2002)). Applications of machine learning to these scenarios can be useful in the clinical decision-making process, such as when to administer vasopressors or ventilate a patient (Wu et al. (2016)). These interventions are used to treat patients in the ICU, and there is need for improvements in the timing, dosages, and setting of these interventions to avoid health complications (Gupta et al. (2017), Craven et al. (1986))

In this work, we learn unsupervised latent representations of patient state that can be used to predict patient mortality and the administration of medical interventions. These representations are not designed to predict one specific outcome, but are general representations that can be used as features for a variety of prediction tasks. These representations are learned from physiological and clinical events data for patients in the ICU.

In this document, we summarize the results of our work. First, we describe the state of the field and previous research that we build upon. Next, we explain the details of our methods for building these representations, the data we used, and the experiments we ran to test these representations. After, we show the results of our experiments and describe the conclusions that we drew from those results. Finally, we discuss the significance of our work and further research that can build off of our work.

2. Background and Related Works

Previous studies used data from patient monitors and medical records to predict patient outcomes and risk. In Henry et al. (2015), researchers developed a system called TREWScore to predict which

patients are at risk for septic shock before the septic shock sets in. Their system was able to identify patients who would eventually go into septic shock with an AUC of 0.83. A recent study focused on predicting long-term 5-year mortality in patients post Coronary Artery Bypass Grafting procedures, using traditional machine learning methods on a variety of physiological features, including blood sample values taken during the patient's ICU stay (Forte et al. (2017)). Other research focused on directly predicting medical interventions in patients in the ICU. In Fialho et al. (2013), researchers focused on predicting vasopressor administration in a subset of the population of ICU patients that receive fluid resuscitation. These previous studies focused on supervised prediction tasks on specific subsets of the ICU population, rather than the entire general ICU cohort. They display the potential of using collected patient electronic health data to predict meaningful patient outcomes, but are not widely generalizable.

In Suresh et al. (2017), researchers focus on a general ICU patient cohort. They predict various medical intervention onsets and weanings by applying supervised deep neural networks to a combination of different sources of patient ICU data, including physiological patient vitals and labs, clinical notes, and static patient features. Their methods achieve vasopressor onset prediction AUC of 0.77 with a 6 hour prediction gap. This work shows the potential power of supervised neural network architectures to capture critical information on patient state for predicting interventions.

In our work, we constructed general unsupervised representations of patient state that are not intervention or outcome specific. The predictive power of unsupervised representations were demonstrated by previous work in Ghassemi et al. (2017). They used an unsupervised switching state autoregressive model to build latent representations of patient states. In our work, we also built unsupervised latent representations of patient states, but we utilized neural network models for our encoding method.

These representations can be learned using a variety of machine learning techniques. We focused on neural network autoencoders and sequence-to-sequence models. Using neural networks to automatically extract features allows for additional insight beyond traditional user-defined, manually extracted patient features. Autoencoders use neural networks to encode the original input into a compressed representation, which is then decoded to output a reconstruction of the original in-

put with the objective of minimal distortion and loss of information. In Tan and Eswaran (2011), researchers demonstrated the ability of autoencoders to build compressed, accurate representation of images that still capture fundamental features of the original uncompressed image. These features can be leveraged to preprocess medical images in prediction tasks. Autoencoders have also shown utility in language modeling and machine translation tasks (Chandar AP et al. (2014)). These compressed representations can make applications of machine learning algorithms manageable in terms of both training time and space. Since we utilized autoencoders to encode and reconstruct the original patient data without optimizing for one outcome, our learned representations are robust and generalizable to a variety of prediction tasks.

We also explored sequence-to-sequence models. Similarly to autoencoders, the sequence-to-sequence model utilizes a neural network architecture to encode a representation in the hidden layer, but instead of outputting the reconstructed input, sequence-to-sequence models output the predicted sequence in a future timestep. Sequence-to-sequence models have achieved notable success in language translation tasks and also video captioning (Sutskever et al. (2014), Venugopalan et al. (2015)). In our framework, input sequences are time windows of patient clinical ICU data and the output sequences are future time windows of patient data. Like the autoencoder representations, the representations learned from sequence-to-sequence modeling are generalizable to multiple prediction tasks. Since they are trained to predict future patient physiological state and data, these learned representations can have strong predictive abilities for which interventions should be administered in the near future for a patient.

In summary, our work builds upon and combines various neural network architectures for autoencoding and sequence-to-sequence modeling to learn latent representations of underlying physiological patient state that are generalizable and predictive of a variety of patient outcomes.

3. Data

3.1 Cohort

We utilized the Medical Information Mart for Intensive Care (MIMIC-III), a publicly available EHR dataset collected from the intensive care units at Beth Israel Deaconess Medical Center (Johnson

Table 1: The percent of patients in our cohort who died in hospital or within 30 days of discharge, who were ventilated at least once during their ICU stay, and who were administered vasopressors during their ICU stay.

Outcome	% of Total Cohort
Mortality	9.53%
Ventilator	40.49%
Vasopressor	30.50%

et al. (2016)). For this study, we used two main cohorts of 34,148 and 14,301 ICU patients from MIMIC-III. These two cohorts are used to study two different data modalities (physiological and clinical events data) that we will describe in the next section. In Figure 1, we display a breakdown of our cohort by mortality outcome and whether they receive ventilator or vasopressor interventions. Roughly 10% of our cohort passed away in the hospital or within 30 days of being discharged, and a large percentage of our cohort were ventilated or received vasopressors during their ICU stay.

3.2 Features

We used two main data modalities from the MIMIC-III database: physiological time-series data and clinical events data from patients in the intensive care unit. For the physiological data, we extracted 29 physiological time-series features per hour, including vital signs and lab test values. These features are continuous-value measurements such as blood pressure or blood cell count. For the clinical events data, we extracted features on 5595 different events, which specify whether a specific procedure was performed and the time it occurred. These procedures include events such as specific lab tests, administered inputs, or patient measurements.

3.3 Preprocessing

3.3.1 Physiological Data

The specific physiological features we utilized are detailed in Table 2. The time-series were extracted on an hourly basis and indexed to the time of ICU admission. We explored two methods for

processing the physiological data. Our first method left the features as continuous-valued features and our second method discretized the data.

For our first method, we directly utilized the continuous physiological time-series data when present. Multiple values in a given hour were reconciled by taking the median value. Missing values were imputed by forward-filling in time for each patient, setting the missing value to be the most recent recorded value for that patient. If the patient did not have any previous values for a measurement, we filled the missing values with the population mean. Afterwards, all data measurements were normalized.

Our second method did not impute the missing values. We found that imputing missing values lost some information that the missingness itself encoded. So that we could represent this missingness, we instead discretized the data. First, for each feature, we broke the range of continuous values into 8 bins with one more bin to represent missingnesss, totaling 9 bins for each of the original 29 features. Next, we dropped any categories that only occurred less than 0.01% of the time. Instead of a continuous value for each of the physiological features, we had a positive indicator in one of the nine appropriate bins for each feature. After dropping features below the 0.01% threshold, we had 252 discrete features per hour of a patient's ICU stay.

3.3.2 Events Data

Our second data set consisted of the clinical events data. The chart times for each unique event were time-discretized to the hour. For each event for each hour, we used a binary indicator to indicate whether the event occurred within the hour. This yielded a time sequence consisting of 5595 binary features per hour.

3.4 Train, Test, and Validation Splits

We split our datasets into training, validation and test sets. We stratified by the first type of ICU care unit to which the patient was admitted to account for any variations in underlying patient diseases in different care units. We trained our neural networks on the training set and selected the best hyperparameter settings by evaluating on the validation set. For our prediction tasks, we trained on

Table 2: Features included in the risk models.

	Features
Vitals	diastolic blood pressure, systolic blood pressure, mean blood pressure, Glascow Coma Scale total, heart rate, respiratory rate, temperature, weight, white blood cell count, pH
Labs	anion gap, bicarbonate, blood urea nitrogen, chloride, creatinine, fraction inspired oxygen, glucose, hematocrit, hemoglobin, INR, lactate, magnesium, oxygen saturation, partial thromboplastin time, phosphate, platelets, potassium, prothrombin time, sodium

the training set and tested on our test set. All results performance metrics are computed on the test set.

3.5 Windowing Data

For each patient stay, we broke down the entire stay into overlapping windows. We extracted windows of a length of 6 hours with a 3 hour shift, meaning that consecutive windows overlap by 3 hours. We believe that extracting encoded representations at these window lengths is more meaningful than longer time durations, particularly in the intensive care environment. While in the ICU, patients are constantly being intervened upon. Because of this, patient physiological "state" is constantly changing (Ghassemi et al. (2017)). We used this windowed data as the input to our neural networks.

4. Methods

4.1 Learning Hidden Representations of Patient State

In this section, we describe our methods for building condensed latent representations of patient state from the physiological and events data. We trained various neural network architectures to take a 6-hour window of patient data as input and attempt to reconstruct either the same original input (for autoencoding networks) or predict the following time window (for the sequence to sequence model). These networks were optimized to minimize the mean squared error between the generated

output signal and the true signal, as seen in (1).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{1}$$

Using these trained networks, we extracted the bottleneck hidden layer embeddings as our latent representations. They capture holistic features not included in standard manual feature extraction or obvious to the human clinician. The unsupervised training allows our representation to be unrestricted by a manually-extracted feature space and generalizable to multiple prediction tasks.

For each of these neural network models, we trained separate networks for each data modality and varied the number of hidden units used for our networks. The number of hidden units is the dimensionality of the bottleneck hidden layer of our neural network architectures and consequently the dimensionality of our learned patient representations. The clinical events data has 5595 dimensions per hour, which yields 33570 features for one 6-hour window. We test a range of hidden dimension sizes from 32 to 256 hidden units. For the the physiological data, which has 1512 features for one 6-hour window, we tried a range of 4 to 16 hidden units. Our final extracted hidden states were used as inputs to binary classification tasks that will be described in a following section.

To choose the best number of hidden units for each type of neural network and each data modality, we evaluated each of our models on the validation set. We chose the model with the lowest mean squared error on the reconstructed signal to build our encoded representations.

4.1.1 Fully Connected Autoencoder

We trained separate simple fully connected autoencoders for the physiological data and the clinical events data. This architecture consisted of an input layer, one hidden layer, and an output layer. The input and output layers were fully connected to the hidden layer. The hidden dimension sizes that yielded the lowest mean squared error were 256 and 16 for the events and physiological data respectively. For our output activation, we used the sigmoid function to constrain our outputs between zero and one. Since the original windowed time series data for each input was a matrix with size of window length (6) by the number of features (252 or 5595), we first flattened the matrix into a

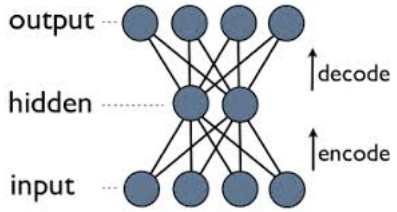

Figure 1: Fully Connected Autoencoder Architecture

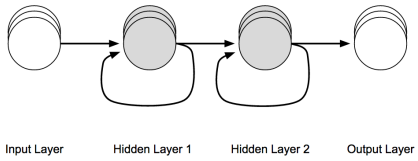

Figure 2: Long Short Term Memory Autoencoder Architecture (Gibson and Patterson (2017))

vector of size $6 \cdot numFeaturesPerHour$. The architecture of this simple fully connected network is shown in Figure 1.

4.1.2 Long Short Term Memory Autoencoder

LSTMs are often used to encode sequences, such as word sentences, because they are able to maintain and combine information from earlier in the sequence (Sundermeyer et al. (2012)). This is an effect of their input, output, and forget gates, which allow them to selectively remember information from throughout the sequence. Our time series physiological and events data also form a type of sequence. For our LSTM autoencoder, we used a simple recurrent neural network structure consisting of two LSTM layers, one for encoding and one for decoding. For these LSTM units, the hidden dimension sizes that yielded the lowest mean square error were 32 and 16 for the events and physiological data respectively. For our output activation, we used the sigmoid function. We extracted the encoded output of the encoding LSTM layer as the encoded patient state representation. See Figure 2 for a basic architecture diagram.

4.1.3 SEQUENCE-TO-SEQUENCE MODEL

Our sequence-to-sequence model, unlike the previously described autoencoders, does not try to reconstruct its input. Instead, the sequence-to-sequence model architecture takes in a sequence and outputs another sequence. Here, our input sequence was a 6-hour time series window of patient data and our output was the next 6-hour window after a 6 hour gap. Our simple sequence-to-sequence model consisted of an LSTM encoder and LSTM decoder with a final dense layer. We initialized the cell and hidden states of the decoder with the final states of our encoder. We extracted the last hidden state of our encoder as our latent representation for patient state. The hidden dimension sizes that yielded the lowest mean squared error were 128 and 16 for the events and physiological data respectively. In Figure 3, we show the general architecture of our model.

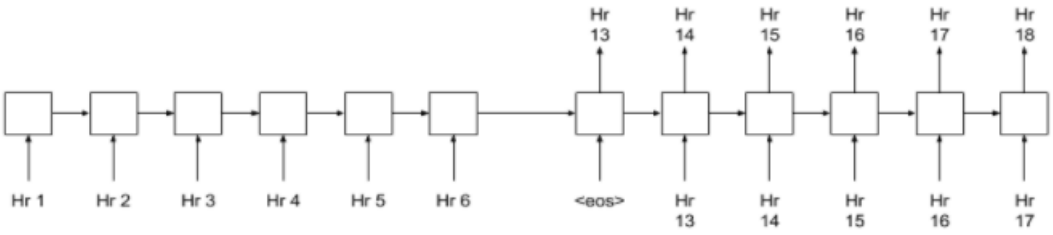

Figure 3: Sequence-to-Sequence model architecture

4.1.4 PRINCIPAL COMPONENT ANALYSIS

We also performed a Principal Component Analysis (PCA) for each of the two data modalities. We used the first n most significant components as an n-dimensional representation of a patient window. These PCA representations were learned in an unsupervised manner, meaning that they are not optimized for one specific prediction outcome. We used these representations as a baseline of comparison against our neural network representations. In order to make it a fair comparison, we selected the numbers of components n that were equal to the maximum hidden dimension sizes used by our neural network representations, which were 256 and 16 for the events and physiological data respectively.

4.2 Outcome Prediction

After building condensed representations of patient state, we used these representations as input features for a variety of prediction tasks. We compared the performance using our latent representations as features to the performance using the original raw data as features. We also explored the effect of combining the raw data with our condensed representations. We evaluated the predictive power of our representations on three different binary prediction tasks: first onset of ventilator intervention prediction, first onset of vasopressor intervention prediction, and patient mortality prediction. For these prediction tasks, we trained a simple L2-regularized logistic regression model with bootstrapping to construct 95% confidence intervals. We used L2 regularization because we did not want to introduce extra sparsity to our representations. For all of these tasks, there was a large class imbalance. About 10% of the patients died in hospital or within 30 days of being discharged, meaning that our samples for mortality prediction were largely negative. As we will describe in the following section on intervention prediction, we also had a severe class imbalance for vasopressor and ventilator first onset prediction. To mitigate the effect of class imbalance, we used balanced class weights when training our logistic regression models.

4.2.1 Intervention Prediction

For both ventilator and vasopressor interventions, we predicted only the first onset. For each 6-hour window of a patient's stay, we predicted whether that patient would first receive the specified intervention anytime during a 4-hour window after a 6-hour gap. This set up is comparable to the prediction windows and gaps used in previous studies on intervention prediction (Ghassemi et al. (2017), Suresh et al. (2017)). Since we were predicting first onset, we only predicted on patient windows preceding the first onset of the intervention and discarded all following windows. For patients who never received the intervention, we included all 6-hour windows of their ICU stay. This made the vast majority of our prediction labels negative.

We used two different patient cohorts. For both cohorts, we first excluded patients who received the specified intervention during the first window (first 6 hours) of their ICU stay. The remaining population made up our first cohort, which we will call Cohort A. For our second cohort, which we

will call Cohort B, we also excluded patients who never received the intervention during his or her ICU stay. This means that every patient in Cohort B had one positively labeled 6-hour window of their stay. Note that for all these filtered cohort prediction tasks, the latent representations were still trained and built on the original unfiltered cohort.

4.2.2 MORTALITY PREDICTION

We predicted patient mortality in the hospital or within 30 days of leaving the hospital. We used data from the first 24 hours of a patient's ICU stay as our input. We concatenated the representations for each of the 6-hour windows to create one large feature vector representing patient state for their first 24 hours. We also used a 24-hour prediction gap, meaning that we excluded patients who died within the first 48 hours of their ICU stay. We believe this increased the difficulty of our prediction task by excluding patients who passed away quickly upon entering the ICU and by excluding the patient ICU measurements and other data directly preceding their death. We compared the performance of our latent representations to the raw features themselves.

5. Results

5.1 Intervention Prediction

We first show our prediction results for first onset intervention prediction. In Table 3, we present a summary of the performance of our various representations on first onset ventilator and vasopressor prediction on Cohorts A and B. We bootstrapped to construct 95% confidence intervals and to calculate two-tailed P-values using an unpaired T-test. Our null hypothesis was that the difference between the mean AUC for our raw representations and the mean AUC for our neural network representations was equal to zero for each data modality. We display the resulting two-tailed P-values for a 0.05 level of significance in Table 4. Neural network representations with AUCs significantly lower than the mean raw AUC are shown in red while those with AUCs significantly higher are shown in blue.

For first onset ventilator prediction on Cohort B, our sequence-to-sequence representations achieved the highest AUC for both the physiological data and events data. In Figure 4, we plot

Table 3: AUC for First Onset Intervention Prediction

Data Type	**Prediction Task**	Fully Connected Autoencoder	LSTM Autoencoder	Seq-to-Seq	Raw	PCA
Physiological	Ventilator - Cohort A	0.57	0.61	**0.63**	**0.63**	0.60
	Vasopressor - Cohort A	0.72	0.68	0.71	**0.76**	0.71
	Ventilator - Cohort B	0.54	0.56	**0.58**	0.57	0.57
	Vasopressor - Cohort B	0.63	0.62	0.63	**0.64**	0.62
Events	Ventilator - Cohort A	0.90	0.83	0.89	**0.91**	0.80
	Vasopressor - Cohort A	**0.91**	0.85	0.89	0.89	0.87
	Ventilator - Cohort B	0.59	0.53	**0.65**	0.60	0.55
	Vasopressor - Cohort B	0.64	0.50	0.63	**0.67**	0.61

Table 4: Two-tailed P-values for testing the null hypothesis that the difference between the mean AUC for the raw representations and mean AUC for each neural network representations is not equal to zero. We test using a 0.05 significance level. P-values for neural network AUCs that were significantly lower than the raw are highlighted in red, while those that are significantly higher are highlighted in blue.

Data Type	**Prediction Task**	Fully Connected Autoencoder	LSTM Autoencoder	Seq-to-Seq
Physiological	Ventilator - Cohort A	0.000967	0.000149	0.667
	Vasopressor - Cohort A	0.72	0.68	0.71
	Ventilator - Cohort B	<0.00001	<0.00001	<0.00001
	Vasopressor - Cohort B	0.635	0.261	0.491
Events	Ventilator - Cohort A	0.485	<0.00001	0.03176
	Vasopressor - Cohort A	0.0318	<0.00001	0.962
	Ventilator - Cohort B	0.476	0.00006	0.0275
	Vasopressor - Cohort B	0.158	<0.00001	0.0912

the AUC for ventilator prediction on Cohort B with 95% confidence intervals. For the events data, our sequence-to-sequence model performed slightly better than the raw representations with a P-value of 0.0275. However, the AUC confidence intervals of our models, including the raw features and PCA representation, all overlap. We believe that all representations performed comparably.

In Figure 5, we show the AUC for vasopressor prediction on Cohort B. The raw features achieved the highest AUCs for both physiological and events data. All representations, except the LSTM autoencoder events data representation, have overlapping confidence intervals. The LSTM events data representation performed significantly worse than the rest of the representations for the events data with a P-value less than $1e-5$. Notably, the dimensionality of the events data LSTM autoencoder representations was only 32 compared to 128 or 256 for the other learned events representations. We believe this might explain its lower performance.

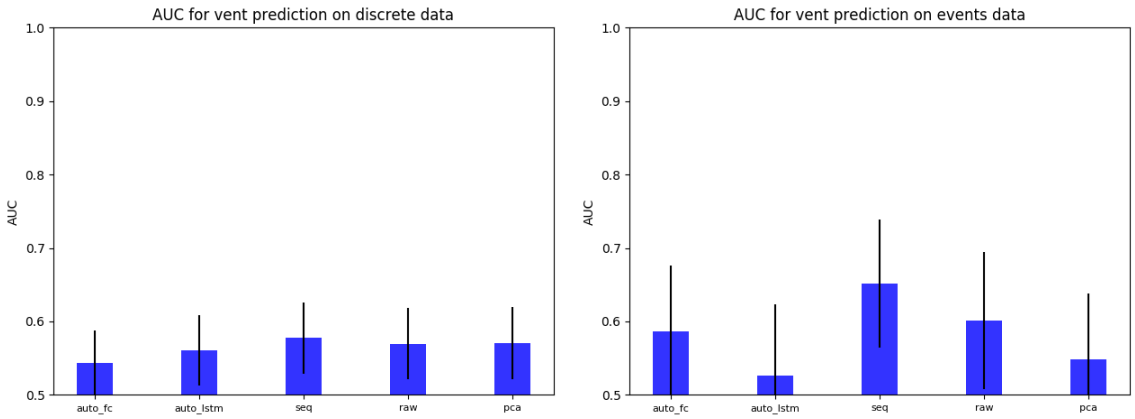

Figure 4: AUC for first onset of ventilator prediction on Cohort B patients who eventually are ventilated. On the left and right we show representations from the physiological and events data respectively.

For Cohort A, the performance increased across the board. Our best neural network representations still performed comparably to the raw features for both data modalities. In Figure 6, we plot the AUC for ventilator prediction on Cohort A. For both events and physiological data, the raw features achieved the highest AUCs of 0.91 and 0.67 respectively. The confidence intervals on Cohort A are smaller and the overall AUCs better when compared to Cohort B. We believe this is caused by the

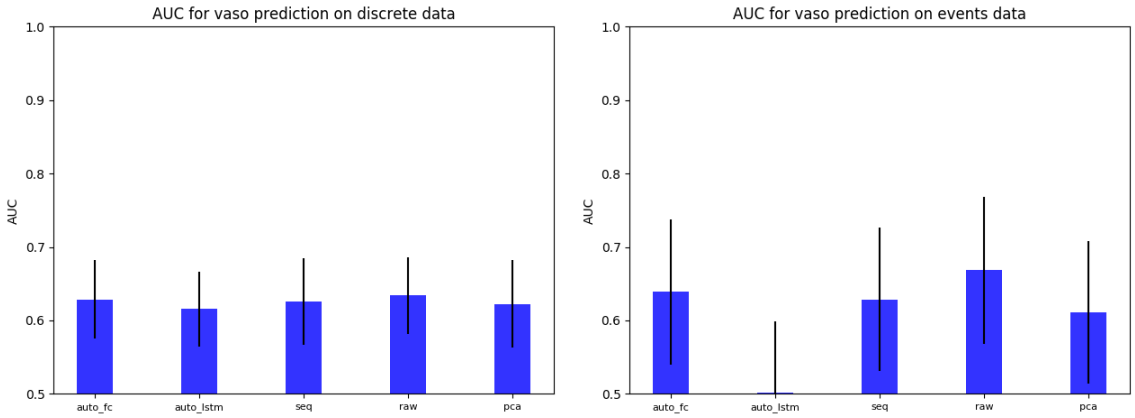

Figure 5: AUC for first onset of vasopressor prediction on Cohort B. On the left and right we show representations from the physiological and events data respectively.

increase in both sample size and the number of negative samples in Cohort A. By including patients who were never ventilated during their ICU stay, we included all negatively labeled windows for the duration of their stay. This increase in class imbalance could have inflated the AUCs. For the events data, the PCA-based representation performed worse than our best neural network representations and the raw features. We believe that our networks that were trained to encode patient state have an advantage over the representations learned using PCA. Notably, all the events data representations outperformed all the physiological data representations. We believe this could be attributed to the fact that the original events feature space is larger than the physiological feature space.

We see similar results for vasopressor prediction on Cohort A in Figure 7. For the physiological data representations, all neural network representations performed comparably to the raw features with P-values all greater than 0.6. For the events data, the fully connected autoencoder representation had the highest AUC, but the confidence intervals for all representations still overlap.

Ghassemi et al. (2017) constructed latent representations of patient state that were used to predict first onset of ventilator and vasopressor. When combined with other features such as the raw representations, their latent representations outperformed all other representations. However, our representations saw no significant performance gain when combined with the raw representations. In Figure 8, we show the AUC of our representations when combined with the raw features.

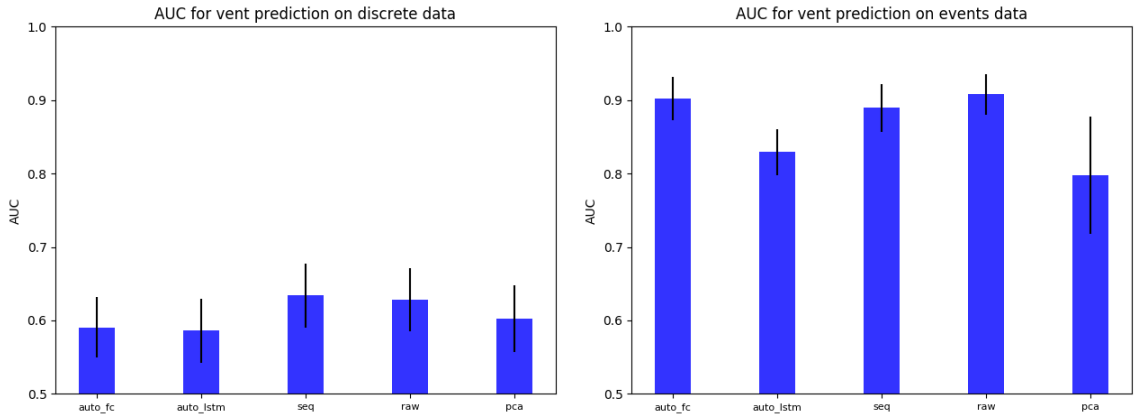

Figure 6: AUC for first onset of ventilator prediction on Cohort A. On the left and right we show representations from the physiological and events data respectively

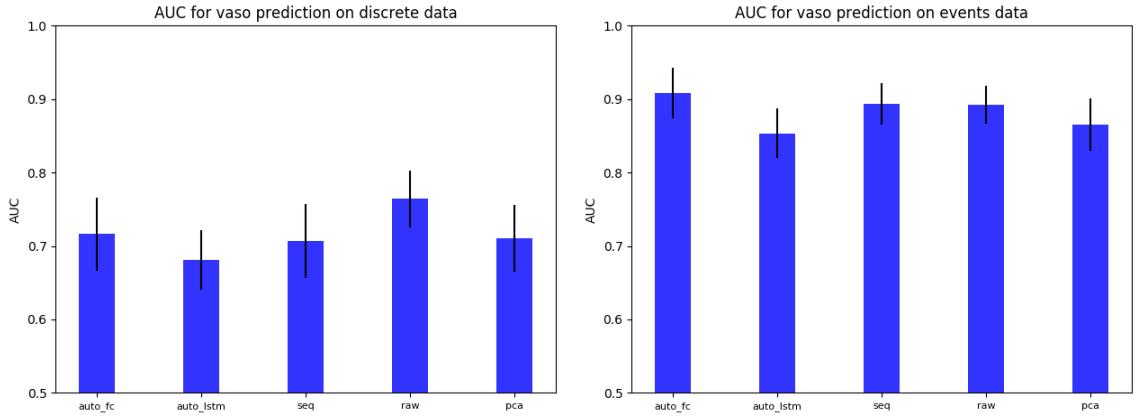

Figure 7: AUC for first onset of vasopressor prediction on Cohort A. On the left and right we show representations from the physiological and events data respectively.

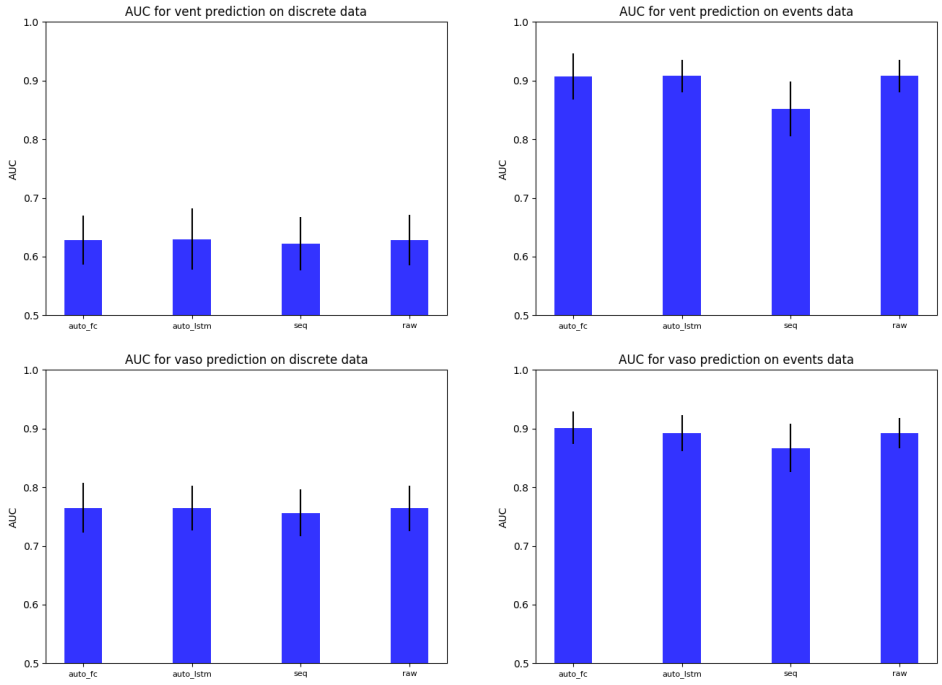

Figure 8: AUC for first onset of intervention prediction using the combined learned representations and raw features on Cohort B. Top left we have ventilator prediction using physiological data. Top right is ventilator prediction using events data. Bottom left is vasopressor prediction using physiological data. Bottom right is vasopressor prediction using events data.

Table 5: Two-tailed P-values for mortality prediction for testing the null hypothesis that the difference between the mean AUC for the raw representations and mean AUC for each neural network representations is not equal to zero. We test using a 0.05 significance level. P-values for neural network AUCs that were significantly lower than the raw are highlighted in red, while those that are significantly higer are highlighted in blue.

Data Type	Fully Connected Autoencoder	LSTM Autoencoder	Seq-to-Seq
Physiological	<0.00001	<0.00001	<0.00001
Events	0.238	<0.00001	0.882

5.2 Mortality Prediction

In Figure 9, we show the mortality prediction performance of our representations using the physiological and events data. We also display the two-tailed P-values comparing our mean neural network representation AUCs to the raw features AUCs in Table 5. For the physiological data representations, the raw data significantly outperformed the other representations. Our neural network representations also performed better than the PCA representation with an AUC of 0.74 compared to 0.54 for PCA. For the events data, our fully connected autoencoder and sequence-to-sequence representations performed comparably to the raw events features while the LSTM autoencoder representations performed significantly worse with a P-value less than $1e-4$. The events PCA features also outperformed the physiological PCA features. We believe that this disparity in performance between the two sets of PCA features could be caused by the smaller dimensionality of the physiological PCA features compared to the events PCA features. For the events data, the PCA representations used the 256 most significant components while the physiological PCA representations only used 16. Even though the physiological neural network representations also only have 16 dimensions, the neural network representations were trained to encode patient state to reconstruct the raw features or predict the following patient window in the original feature space. This seems to give the neural network representations an advantage over the PCA representations, which becomes apparent when limited to only 16 features.

In Table 6, we compare the confusion matrices for three representations for mortality prediction on the physiological data. We see that the raw features yielded better sensitivity and specificity

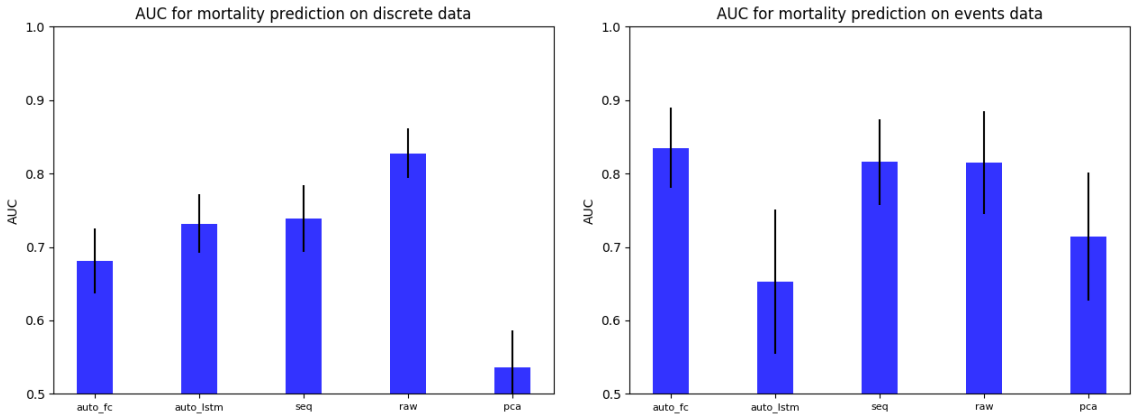

Figure 9: AUC mortality prediction using the first 24 hours of ICU stay data. On the left we have mortality prediction results using the physiological data. On the right we have mortality prediction results using events data.

Table 6: Confusion Matrices for Mortality Prediction on Fully Connected Autoencoder, PCA, and raw physiological feature representations.

	Fully Connected Autoencoder		**PCA**		**Raw**		
	Pred 0	Pred 1	Pred 0	Pred 1	Pred 0	Pred 1	
True 0	**3375**	**1867**	**2690**	**2552**	**4189**	**1053**	5242
True 1	**185**	**288**	**217**	**256**	**159**	**314**	473
Totals	3560	2155	2907	2808	4348	1367	

when compared to the 16-dimensional fully connected autoencoder and PCA representations. The sensitivity and specificity of our fully connected autoencoder representations were both better than that of the PCA representations. We believe that the lower dimensional PCA representations and autoencoder representations were less robust to the severe class imbalance towards negative samples and so had worse specificity.

5.3 Effective Dimensionality of Raw Features during Prediction

Although the dimensionality of the raw data, both events and physiological, was high, their effective dimensionality was lower. Not all features were actually used when predicting the specified

Table 7: Number features with nonzero weight

Data Type	Prediction Task	Number Features != 0	Total Number Features
Physiological	Ventilator Vasopressor Mortality	144 252 1464	1512
Events	Ventilator Vasopressor Mortality	264 330 1212	33570

outcome. To help quantify the effective dimensionality of the raw data on these specific prediction tasks, we analyzed the weights assigned to each feature by the L1-regularized logistic regression model. We used the L1 penalty to encourage sparsity. In Table 7, we show the count of the number of features that were assigned nonzero weight. For the events data, the majority of the 33570 features were forced to zero for all three prediction tasks. The number of raw features that contributed to the model is on the same order of magnitude as the dimensionality of our learned neural network representations dimensionalities. This means that the effective reduction in dimensionality from the raw features to the learned representations was smaller than it seemed. For the physiological data, most of the raw features for intervention prediction were also forced to zero. However, for mortality prediction, only 8 of the 1512 raw features were forced to zero. We believe that this high effective raw feature dimensionality for the physiological data helps explain why our learned representations performed relatively poorly on mortality prediction. We believe that our 16-dimensional representations were too small to accurately encode the necessary information for mortality prediction.

6. Discussion

In general, the learned representations built using neural network autoencoders and sequence-to-sequence models performed comparably to the raw features on all three prediction tasks. For mortality prediction on the physiological data, the PCA representations performed significantly worse than both the neural network and raw representations. We believe that our neural networks were able to capture patient state better than simple PCA because the networks were trained to minimize information loss in the original feature space. Originally, we had believed that the sequence-to-sequence representations would perform better than other learned representations because they were trained

to predict future patient state rather than just recreate the original signal. However, sequence-to-sequence representations did not consistently perform better than our other representations. We also expected the raw features to overfit on our prediction tasks. But from examining the effective dimensionality and the actual performance of the raw features, we believe that many features are not used in these prediction tasks and so the raw features did not overfit.

Performance on intervention and mortality prediction using representations learned from the physiological features was usually lower than the performance on the events data representations. Specifically for mortality prediction, performance using the learned physiological representations was worse than raw physiological data and also the events data representations. We believe that the approximate 90-times reduction in dimensionality from the raw to learned physiological representations caused information loss. When selecting the best number of hidden units when training our neural networks, all networks yielded lowest mean squared error with the largest number of hidden units that we tried (16), We believe we might achieve lower loss with a larger hidden dimensionality than 16.

When we expanded the intervention prediction patient cohort to include patients who were never administered the intervention in question (Cohort B), the performance improved. We believe these performances were inflated by the increase in negative samples. Vasopressor prediction saw a boost in performance larger than the improvement for ventilator prediction. We believe this bigger boost was caused by the larger number of patients who were ventilated during their ICU stay. Since more patients did not receive vasopressors, including these patients added a relatively larger number of samples than for ventilator prediction.

Even though our learned representations have smaller dimensionality than the raw features for both data modalities, our representations achieved comparable performance. All our learned representations were over 90 times smaller than the raw features. However, as we saw from examining the feature weights assigned to the raw features in our L1-regularized logistic regression, the effective dimensionality of the raw features was on the same order of magnitude as our learned representations.

When comparing our results with a similar study, we find our performance was comparable. In the study Ghassemi et al. (2017), researchers also built unsupervised latent-space representations of patient state from ICU physiological data and used these representations as features to predict first onset of various interventions, including ventilator and vasopressor. In our study, we used a 6-hour prediction gap, while in Ghassemi et al. (2017) they tested 1-hour, 2-hour, 4-hour, and 8-hour prediction gaps. They found that increasing the prediction gap decreases performance, so to be conservative we compare our results to their results using the 4-hour prediction gap. Our learned physiological representations achieved similar performance. For ventilator prediction, their representations achieved AUC of 0.64 whereas our best representationw achieved 0.63. For vasopressor, our representation had an AUC of 0.58 and theirs 0.56.

7. Conclusions

By using neural network autoencoders and sequence-to-sequence models, we can learn unsupervised latent representations of patient state in the ICU. These representations are trained in an unsupervised manner and are therefore generalizable to a variety of prediction tasks relevant for patient care in the ICU. These representations are able to perform comparably to the original raw data on first onset vasopressor, ventilator, and mortality prediction even though they are smaller in dimensionality. When compared to PCA representations, our representations are more robust to dimensionality reduction. While our representations do not improve on the performance of the raw features, they are more concise.

8. Future Work

We believe that there is potential for performance improvement using our learned representations. In Ghassemi et al. (2017), researchers found that their unsupervised latent representation also performed worse than the raw features themselves. But when they combined the latent representations, raw features, and static features of the patient (e.g. age, weight, BMI) their model outperformed the raw features and even the raw features combined with the static features. we believe that in conjunction the latent representations can boost predictive power. Furthermore, we might achieve

improvements by combining the latent representations learned from the two different data modalities, events and physiological data. While both data modalities describe patient state in the ICU, they contain different types of measurements and data, which may have stronger predictive power when combined. These data modalities could be combined by learning separate latent representations from each data modality (as in this study) and then concatenating the two learned representations to create a combined representation. Another option would be to concatenate the raw features first and then learn one latent representation from the combination of the two raw data sources.

Further improvements might be achieved by refinement of our current neural network architectures. As mentioned earlier, the dimensionality of the hidden representations learned on the physiological data might have been too small and led to information loss. Increasing the number of hidden units might boost our performance. Also, for all our neural network architectures, we opted for simple models and architectures. Usually each encoder and decoder was composed of one single layer. Higher-level features might be learned by adding more layers to make our networks deeper and increasing the complexity.

As with many neural network models, the question of interpretability arises. In clinical decision making, it is important to try to understand *why* a prediction is made. For the raw features, we can understand the relative importance of certain features by looking at the weights assigned to the features by the trained linear classifier. For our latent representations, if we looked at the weights assigned by the linear classifier to each dimension, we would not know how this corresponds to the original feature space. One method to shed more light on interpretability is to try feature occlusion in the original feature space. We could examine which feature occlusions cause the most drastic changes in predictive performance.

Overall, with future research, generalizable latent representations for patient state could aid clinicians to improve patient outcomes and clinical care in the intensive care unit.

References

Michael D Buist, Gaye E Moore, Stephen A Bernard, Bruce P Waxman, Jeremy N Anderson, and Tuan V Nguyen. Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: preliminary study. *Bmj*, 324(7334):387–390, 2002.

Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861, 2014.

Donald E Craven, Laureen M Kunches, Vetal Kilinsky, Deborah A Lichtenberg, Barry J Make, and William R McCabe. Risk factors for pneumonia and fatality in patients receiving continuous mechanical ventilation. *The American review of respiratory disease*, 133(5):792–796, 1986.

AS Fialho, LA Celi, F Cismondi, SM Vieira, SR Reti, JMC Sousa, SN Finkelstein, et al. Disease-based modeling to predict fluid response in intensive care units. *Methods Inf Med*, 52(6):494–502, 2013.

José Castela Forte, Marco A Wiering, Hjalmar R Bouma, Fred Geus, and Anne H Epema. Predicting long-term mortality with first week post-operative data after coronary artery bypass grafting using machine learning models. In *Machine Learning for Healthcare Conference*, pages 39–58, 2017.

Marzyeh Ghassemi, Mike Wu, Michael C Hughes, Peter Szolovits, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. *AMIA Summits on Translational Science Proceedings*, 2017:82, 2017.

Adam Gibson and Josh Patterson. Deep learning. 2017.

Babita Gupta, Neha Garg, and Rashmi Ramachandran. Vasopressors: Do they have any role in hemorrhagic shock? *Journal of anaesthesiology, clinical pharmacology*, 33(1):3, 2017.

Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science Translational Medicine*, 7(299):299ra122–299ra122, 2015.

E Jamoom, N Yang, and E Hing. Office-based physician electronic health record adoption. *Office of the National Coordinator for Health Information Technology*, 2016.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In *Machine Learning for Healthcare Conference*, pages 322–337, 2017.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

Chun Chet Tan and Chikkannan Eswaran. Using autoencoders for mammogram compression. *Journal of medical systems*, 35(1):49–58, 2011.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4534–4542, 2015.

Mike Wu, Marzyeh Ghassemi, Mengling Feng, Leo A Celi, Peter Szolovits, and Finale Doshi-Velez. Understanding vasopressor intervention and weaning: Risk prediction in a public heterogeneous clinical time series database. *Journal of the American Medical Informatics Association*, page ocw138, 2016.