Clinical Intervention Prediction and Understanding using Deep Networks

Harini Suresh   HSURESH@MIT.EDU

Nathan Hunt   NHUNT@MIT.EDU

Alistair Johnson   AEWJ@MIT.EDU

Leo Anthony Celi   LCELI@MIT.EDU

Peter Szolovits   PSZ@MIT.EDU

Marzyeh Ghassemi   MGHASSEM@MIT.EDU

Computer Science and Artificial Intelligence Lab, MIT
Cambridge, MA

Abstract

Real-time prediction of clinical interventions remains a challenge within intensive care units (ICUs). This task is complicated by data sources that are noisy, sparse, heterogeneous and outcomes that are imbalanced. In this paper, we integrate data from all available ICU sources (vitals, labs, notes, demographics) and focus on learning rich representations of this data to predict onset and weaning of multiple invasive interventions. In particular, we compare both long short-term memory networks (LSTM) and convolutional neural networks (CNN) for prediction of five intervention tasks: invasive ventilation, non-invasive ventilation, vasopressors, colloid boluses, and crystalloid boluses. Our predictions are done in a forward-facing manner to enable "real-time" performance, and predictions are made with a six hour gap time to support clinically actionable planning. We achieve state-of-the-art results on our predictive tasks using deep architectures. We explore the use of feature occlusion to interpret LSTM models, and compare this to the interpretability gained from examining inputs that maximally activate CNN outputs. We show that our models are able to significantly outperform baselines in intervention prediction, and provide insight into model learning, which is crucial for the adoption of such models in practice.

1. Introduction

As Intensive Care Units (ICUs) play an increasing role in acute healthcare delivery (Vincent, 2013), clinicians must anticipate patient care needs in a fast-paced, data-overloaded setting. The secondary analysis of healthcare data is a critical step toward improving modern healthcare, as it affords the study of care in the real care settings and patient populations. The widespread availability of electronic healthcare data (Charles et al., 2013; Jamoom E and E, 2016) allows new investigations into evidence-based decision support, where we can learn when patients need a given intervention. Continuous, forward-facing event prediction is particularly applicable in the ICU setting where we want to account for evolving clinical needs and information throughout the patient's stay.

In this work, we focus on predicting the onset and weaning of interventions. The efficacy of interventions can vary drastically from patient to patient, and unnecessarily administering an in-

tervention can be harmful and expensive. Any treatments come with inherent risks, and we target interventions that span a wide severity of needs in critical care— specifically, invasive ventilation, non-invasive ventilation, vasopressors, colloid boluses, and crystalloid boluses. Mechanical ventilation is commonly used for breathing assistance, but has many potential complications (Yang and Tobin) and small changes in ventilation settings can have large impact in patient outcomes (Tobin, 2006). Vasopressors are a common ICU medication, but there is no robust evidence of improved outcomes from their use (Müllner et al., 2004), and some evidence they may be harmful (DAragon et al., 2015). Fluid boluses are used to improve cardiovascular function and organ perfusion. There are two bolus types: crystalloid and colloid. Both are often considered as less aggressive alternatives to vasopressors, but there are no multi-center trials studying whether fluid bolus therapy should be given to critically ill patients, only studies trying to distinguish which type of fluid should be given (Malbrain et al., 2014).

Capturing complex relationships across many disparate data types is key for predictive performance in our tasks. To this end, we take advantage of the success of deep learning models to capture rich representations of data with little hand-engineering by domain experts. We use long short-term memory networks (LSTM) (Hochreiter and Schmidhuber, 1997), which have been shown to effectively model complicated dependencies in timeseries data (Bengio et al., 1994). Previously, LSTMs have achieved state-of-the-art results in many different applications, such as machine translation (Hermann et al., 2015), dialogue systems (Chorowski et al., 2015) and image captioning (Xu et al., 2015). They are well-suited to our modeling tasks because clinical conditions may be spread apart over several hours. We compare the LSTM models to a convolutional neural network (CNN) architecture that has previously been explored for longitudinal laboratory data (Razavian et al., 2016). All models predict outcomes in a continuous manner given any patient record over vitals, labs, demographic, and notes. In doing so, we:

1. Achieve state-of-the-art prediction results in our forward-facing, hourly prediction of clinical interventions (onset, weaning, and continuity) that could be used at the time of care.
2. Demonstrate that different data modalities and features are most important for different types of predictive tasks in our LSTM using feature occlusion. This is an important step in making models more interpretable by physicians.
3. Highlight patient trajectories that lead to the most and least confident predictions in our CNN across outcomes and features, also aiding in interpretability.

2. Background and Related Work

Clinical decision-making often happens in settings of limited knowledge and high uncertainty; for example, only 10 of the 72 ICU interventions evaluated in randomized controlled trials (RCTs) are not associated with improved outcomes (Ospina-Tascón et al., 2008). Our goal is to gain insight from healthcare data previously collected for the primary purpose of facilitating patient care.

Recent studies have applied recurrent neural networks (RNNs) to modeling sequential EHR data to tag ICU signals with billing code labels (Che et al., 2016; Lipton et al., 2015; Choi et al., 2015), to identify the impact of different drugs for diabetes (Krishnan et al., 2015). Razavian et al. (2016) compared CNNs to LSTMs for longitudinal outcome prediction on billing codes using lab tests. With regard to interpretability, Choi et al. (2016) used temporal attention to identify important features in early diagnostic prediction of chronic diseases from time-ordered billing codes. Others have focused on using representations of clinical notes (Ghassemi et al., 2014) or patient physiological signals to predict mortality (Ghassemi et al., 2015).

Previous work on interventions in ICU populations have often either focused on a single outcome or used data from specialized cohorts. Such models with vasopressors as a predictive target have achieved AUCs of 0.79 in patients receiving fluid resuscitation (Fialho et al., 2013), 0.85 in septic shock patients (Salgado et al., 2016), and 0.88 for onset after a 4 hour gap and 0.71 for weaning, only trained on patients who did receive a vasopressor (Wu et al., 2016). However, we train our models on general ICU populations in order to make them more applicable. In the most recent prior work on interventions, also on a general ICU population, the best AUC performances were 0.67 (ventilation), 0.78 (vasopressor) for vasopressor onset prediction after a 4 hour gap (Ghassemi et al., 2017). These were lowered to 0.66 and 0.74 with a longer gap time of 8 hours.

3. Data and Preprocessing

See Figure 1 for an overall description of data flow.

3.1 Data Source

We use data from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III v1.4) database (Johnson et al., 2016). MIMIC is publicly available, and contains over 58,000 hospital admissions from approximately 38,600 adults. We consider patients 15 and older who had ICU stays from 12 to 240 hours and consider each patient's first ICU stay only. This yields 34,148 unique ICU stays.

3.2 Data Extraction and Preprocessing

For each patient, we extract:
1. 5 static variables such as gender and age
2. 29 time-varying vitals and labs such as oxygen saturation and blood urea nitrogen
3. All available, de-identified clinical notes for each patient as timeseries across their entire stay
(See Table 3 for a complete listing of variables)

Static variables were replicated across all timesteps for each patient. Vital and lab measurements are given timestamps that are rounded to the nearest hour. If an hour has multiple measurements for a signal, those measurements are averaged.

3.3 Representation of Notes and Vitals

Clinical narrative notes were processed to create a 50-dimensional vector of topic proportions for each note using Latent Dirichlet Allocation (Blei et al., 2003; Griffiths and Steyvers, 2004). These vectors are replicated forward and aggregated through time (Ghassemi et al., 2014). For example, if a patient had a note A recorded at hour 3 and a note B at hour 7, hours 3–6 would contain the topic distribution from A, while hours 7 onward would contain the aggregated topic distribution from A and B combined.

We compare raw physiological data to *physiological words*, where we categorize the vitals data by first converting each value into a z-score based on the population mean and standard deviation for that variable, and then rounding this score to the nearest integer and capping it to be between -4 and 4. Each z-score value then becomes its own column, which explicitly allows for a representation of missingness (e.g., all columns for a particular variable zeroed) that does not require imputation (Figure 7 in Appendix B) (Wu et al., 2016).

The physiological variables, topic distribution, and static variables for each patient are concatenated into a single feature vector per patient per hour (Esteban et al., 2016). The intervention state
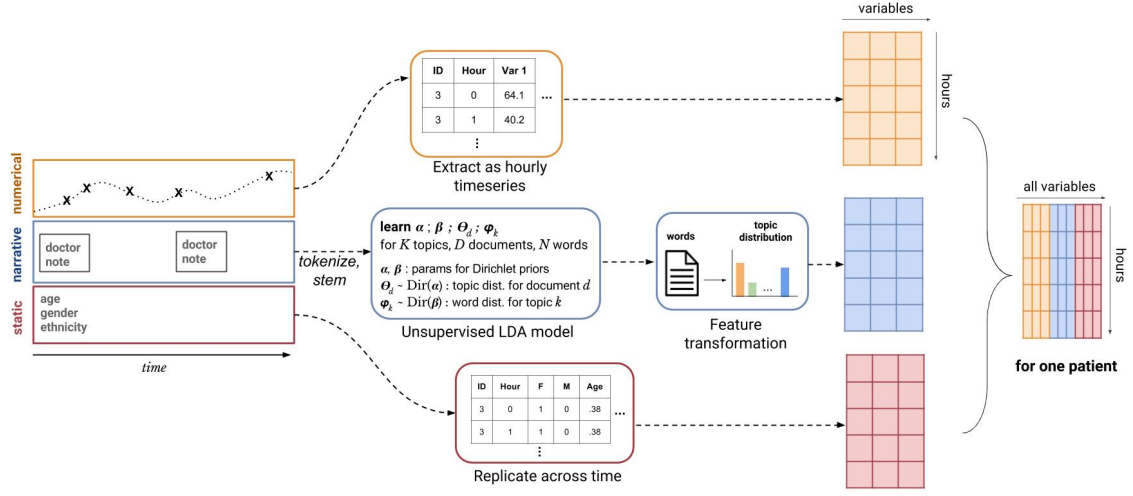
Figure 1: Data preprocessing and feature extraction with numerical measurements and lab values, clinical notes and static demographics.
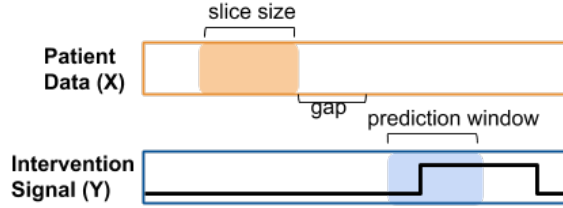

Figure 2: Given data from a fixed-length (6 hour) sliding window, models predict the status of intervention in a prediction window (4 hours) after a gap time (6 hours). Windows slide along the entire patient record, creating multiple examples from each record.

of each patient (a binary value indicating whether or not they are on the intervention of interest at each timestep) and the time of day for each timestep (an integer from 0 to 23 representing the hour) are also added to this feature vector. Using the time of day as a feature makes it easier for the model to capture circadian rhythms that may be present in, e.g., the vitals data.

3.4 Prediction Task

We split each patients record into 6 hour chunks using a sliding window and make a prediction for a window of 4 hours after a gap time of 6 hours (Figure 2). When predicting ventilation, non-invasive ventilation, or vasopressors, the model classifies the prediction window as one of four possible outcomes: 1) Onset, 2) Wean, 3) Staying on intervention, 4) Staying off intervention. A prediction window is an onset if there is a transition from a label of 0 to 1 for the patient during that window; weaning is the opposite: a transition from 1 to 0. A window is classified as "stay on" if the label for the entire window is 1 or "stay off" if 0. When predicting colloid or crystalloid boluses, we classify the prediction window into one of two classes: 1) Onset, or 2) No Onset, since these interventions are not administered for on-going durations of time. After splitting the patient records into fixed-length chunks, we end up with 1,154,101 examples. Table 1 lists the proportions of each class for each intervention.

4

	Onset	Weaning	Stay Off	Stay On
Ventilation	0.005	0.017	0.798	0.18
Vasopressor	0.008	0.016	0.862	0.114
NI-Ventilation	0.024	0.035	0.695	0.246
Colloid Bolus	0.003	-	-	-
Crystalloid Bol	0.022	-	-	-

Table 1: The proportion of each intervention class. Note that colloid and crystalloid boluses are not administered for specific durations, and thus have only a single class (onset). NI = non-invasive.

4. Methods

4.1 Long Short-Term Memory Network (LSTM)

We use long short-term memory networks (LSTM) as our first model. Having seen the input sequence $x_1 \ldots x_t$ of a given example, we predict $\hat{y}_t$, a probability distribution over the outcomes, with target outcome y_t:

$$h_1 \ldots h_t = LSTM(x_1 \ldots x_t) \tag{1}$$

$$\hat{y}_t = softmax(W_y h_t + b_y) \tag{2}$$

where $x_i \in \mathbb{R}^V, W_y \in \mathbb{R}^{N_C \times L_2}, h_t \in \mathbb{R}^{L_2}, b_y \in \mathbb{R}^{N_C}$ where V is the dimensionality of the input (number of variables), N_C is the number of classes we predict, and L_2 is the second hidden layer size. For a model schematic, see Figure 3a, and for more details on model implementation, see the Appendix.

4.2 Convolution Neural Network (CNN)

We employ a similar CNN architecture to Razavian et al. (2016), except that we do not initially convolve the features into an intermediate representation. We represent features as channels and perform 1D temporal convolutions, rather than treating the input as a 2D image. Our architecture consists of temporal convolutions at three different temporal granularities with 64 filters each. The dimensions of the filters are $1 \times i$, where $i \in \{3, 4, 5\}$.

We pad the inputs such that the outputs from the convolutional layers are the same size, and we use a stride of 1. Each convolution is followed by a max pooling layer with a pooling size of 3. The outputs from all three temporal granularities are concatenated and flattened, and followed by 2 fully connected layers with dropout in between and a softmax over the output (Figure 3b).

4.3 Experimental Settings

We use a train/validation/test split of 70/10/20 and stratify the splits based on outcome. For the LSTM, we use dropout with a keep probability of 0.8 during training (only on stacked layers), and L2 regularization with lambda = 0.0001. We use 2 hidden LSTM layers of 512 nodes each. For the CNN, we use dropout between fully-connected layers with a keep probability of 0.5. We use a weighted loss function during optimization to account for class imbalances. All parameters were determined using cross-validation with the validation set. We implemented all models in TensorFlow version 1.0.1 using the Adam optimizer on mini-batches of 128 examples. We determine when to stop training with early stopping based on the AUC on the validation set.
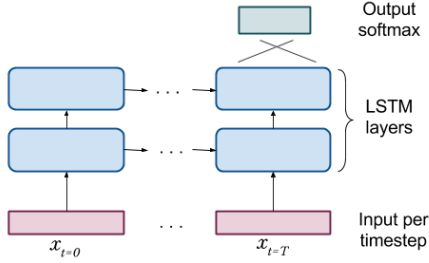
(a) The LSTM consists of two hidden layers with 512 nodes each. We sequentially feed in each hour's data. At the end of the example window, we use the final hidden state to predict the output.
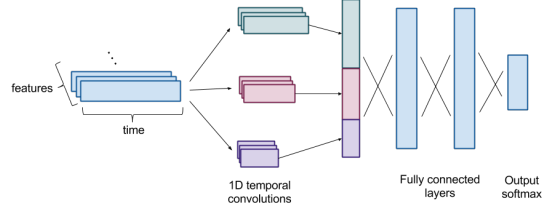
(b) The CNN architecture performs temporal convolutions at 3 different granularities (3, 4, and 5 hours), max-pools and combines the outputs, and runs this through 2 fully connected layers to arrive at the prediction.

Figure 3: Schematics of LSTM and CNN model architectures.

4.4 Evaluation

We evaluate our results based on per-class AUCs as well as aggregated macro AUCs. If there are K classes each with a per-class AUC of AUC_k then the macro AUC is defined as the average of the per-class AUCS, $AUC_{macro} = \frac{1}{K} \sum_k AUC_k$. We use the macro AUC as an aggregate score because it weights the AUCs of all classes equally, regardless of class size (Manning et al., 2008). This is important because of the large class imbalance present in the data.

We use L2 regularized logistic regression (LR) as a baseline for comparison with the neural networks (Pedregosa et al., 2011). The same input is used as for the numerical LSTM and CNN (imputed 6 hour chunks of data).

4.5 Interpretibility

4.5.1 LSTM FEATURE-LEVEL OCCLUSIONS

Because of the additional time dependencies of recurrent neural networks, getting feature-level interpretability from LSTMs is notoriously difficult. To achieve this, we borrow an idea from image recognition to help understand how the LSTM uses different features of the patients. Zeiler and Fergus (2013) use occlusion to understand how models process images: they remove a region of the image (by setting all values in that region to 0) and compare the model's prediction of this occluded image with the original prediction. A large shift in the prediction implies that the occluded region contains important information for the correct prediction. With our LSTM model, we remove features one by one from the patients (by replacing the given feature with noise drawn from a uniform distribution in [0,1)). We then compare the predictive ability of the model with and without each feature; when this difference is large, then the model was relying heavily on that feature to make the prediction.

4.5.2 CNN FILTER/ACTIVATION VISUALIZATION

We get interpretability from the CNN models in two ways. First, in order to understand how the CNN is using the patient data to predict certain tasks, we find and compare the top 10 real examples that our model predicts are most and least likely to have a specific outcome. As our gap time is 6 hours, this means that the model predicts high probability of onset of the given task 6 hours after the end of the identified trajectories.

Second, we generate "hallucinations" from the model which maximize the predicted probability for a given task (Erhan et al., 2009). This is done by creating an objective function that maximizes the activation of a specific output node, and backpropagating gradients back to the input image, adjusting the image so that it maximally activates the output node.

5. Results

We found deep architectures achieved state-of-the-art prediction results for our intervention tasks. The AUCs for each of our five intervention types and 4 prediction tasks are shown for all models in Table 2. All models use 6 hour chucks of "raw" data which have either been transformed to a 0-1 range (normalized and mean imputed), or discretized into physiological words (Section 3.3).

5.1 Physiological Words Improve Predictive Task Performance With High Class Imbalance

We observed a significantly increased AUC for some interventions when we used physiological words — specifically for ventilation onset (from 0.61 to 0.75) and colloid bolus onset (from 0.52 to 0.72), which have the lowest proportion of onset examples (Table 1). This may be because physiological words have a smoothing effect. Since we round the z-score for each value to the nearest integer, if a patient has a heart rate of 87 at one hour and then 89 at the next, those will probably be represented as the same word. This may make the model invariant to small fluctuations in the patient's data and more resilient to overfitting small classes. In addition, the physiological word representation has an explicit encoding for missing data. This is in contrast to the raw data that has been forward-filled and mean-imputed, introducing noise and making it difficult for the model to know how confident to be in the measurements it is given (Che et al., 2016).

5.2 Feature-Level Occlusions Identify Important Per-Class Features

We are able to interpret the LSTM's predictions using feature occlusion (Section 4.5.1). We note that vitals, labs, topics and static data are important for different interventions (Figure 4). Table 5 has a complete listing of the most probable words for each topic mentioned.

For mechanical ventilation, the top five important features are consistent for weaning and onset (pH, sodium, lactate, hemoglobin, and potassium). This is sensible, because all are important lab values used to assess a patient's physiological stability, and ventilation is an aggressive intervention. However, ventilation onset additionally places importance on a patient's Glasgow Coma Score (GCS) and Topic 4 (assessing patient consciousness), likely because patient sedation is a critical part of mechanical ventilation. We also note that the scale of AUC difference between ventilation onset and weaning is the largest observed (up to 0.30 for weaning and 0.12 for onset).

In vasopressor onset prediction, physiological variables such as potassium and hematocrit are consistently important, which agrees with clinical assessment of cardiovascular state (Bassi et al., 2013). Similarly, Topic 3 (noting many physiological values) is also important for both onset and weaning. Note that the overall difference in AUC for onset ranges up to 0.16, but there is no significant decrease in AUC for weaning (< 0.02). This is consistent with previous work that demonstrated weaning to be a more difficult task in general for vasopressors (Wu et al., 2016). We also note that weaning prediction places importance on time of day. As noted by Wu et al. (2016), this could be a side-effect of patients being left on interventions longer than necessary.

For non-invasive ventilation onset and weaning the learned topics are more important than physiological variables. This may mean that the need for less severe interventions can only be detected

Task	Model	Intervention Type				
		VENT	NI-VENT	VASO	COL BOL	CRYS BOL
Onset AUC	Baseline	0.60	0.66	0.43	0.65	0.67
	LSTM Raw	0.61	0.75	**0.77**	0.52	0.70
	LSTM Words	**0.75**	**0.76**	0.76	**0.72**	**0.71**
	CNN	0.62	0.73	**0.77**	0.70	0.69
Wean AUC	Baseline	0.83	0.71	0.74	-	-
	LSTM Raw	0.90	0.80	**0.91**	-	-
	LSTM Words	0.90	**0.81**	**0.91**	-	-
	CNN	**0.91**	0.80	**0.91**	-	-
Stay On AUC	Baseline	0.50	0.79	0.55	-	-
	LSTM Raw	0.96	**0.86**	**0.96**	-	-
	LSTM Words	**0.97**	**0.86**	0.95	-	-
	CNN	0.96	**0.86**	**0.96**	-	-
Stay Off AUC	Baseline	0.94	0.71	0.93	-	-
	LSTM Raw	0.95	**0.86**	**0.96**	-	-
	LSTM Words	**0.97**	**0.86**	0.95	-	-
	CNN	0.95	**0.86**	**0.96**	-	-
Macro AUC	Baseline	0.72	0.72	0.66	-	-
	LSTM Raw	0.86	**0.82**	**0.90**	-	-
	LSTM Words	**0.90**	**0.82**	0.89	-	-
	CNN	0.86	0.81	**0.90**	-	-

Table 2: Comparison of model performance on five targeted interventions. Models that perform best for a given (intervention, task) pair are bolded.

from clinical insights derived in notes. Similarly to vasopressors, we note that onset AUCs vary more than weaning AUCs (0.14 vs 0.01), and that time of day is important for weaning.

For crystalloid and colloid bolus onsets, topics are all but one of the five most important features for detection. Colloid boluses in general have more AUC variance for the topic features (0.14 vs. 0.05), which is likely due to the larger class imbalance compared to crystalloids.

5.3 Convolutional Filters Target Short-term Trajectories

We are able to understand the CNN by examining maximally activating patient trajectories (Section 4.5.2). Figure 5 shows the mean with standard deviation error bars for four of the most differentiated features of the 10 real patient trajectories that are the highest and lowest activating for each task. The trends suggest that patients who will require ventilation in the future have higher diastolic blood pressure, respiratory rate, and heart rate, and lower oxygen saturation — possibly corresponding to patients who are experiencing hyperventilation. For vasopressor onsets, we see a decreased systolic blood pressure, heart rate and oxygen saturation rate. These could either indicate altered peripheral perfusion or stress hyperglycemia. Topic 3, which was important for vasopressor onset using occlusion 4, is also increased.

In the less invasive tasks, we saw decreased creatinine, phosphate, oxygen saturation and blood urea nitrogen for non-invasive ventilation, potentially indicating neuromuscular respiratory failure.
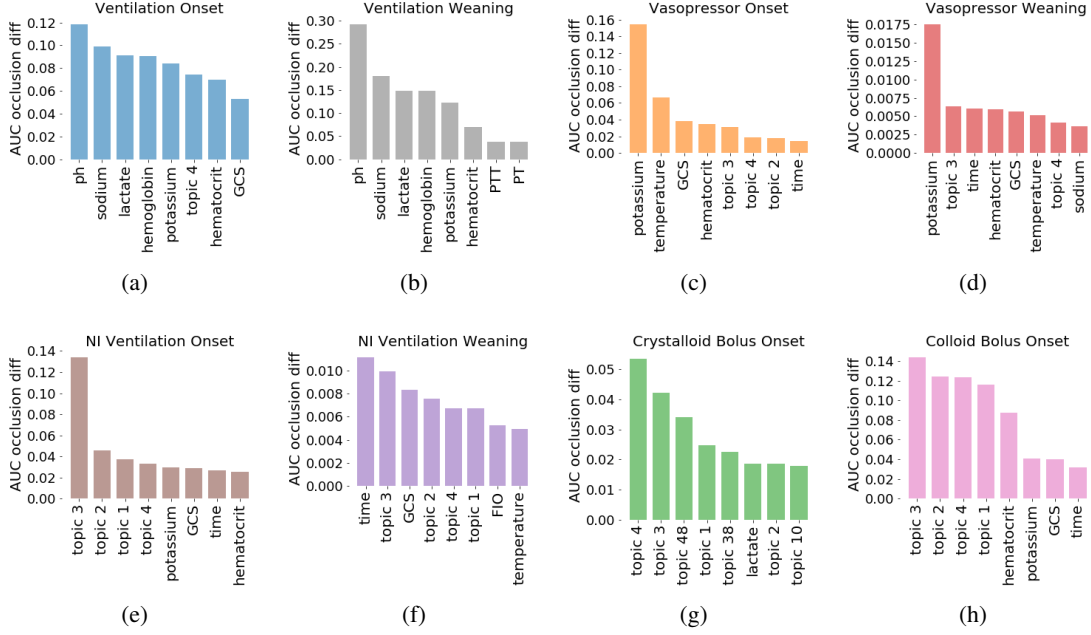
Figure 4: We are able to make interpretable predictions using the LSTM and occluding specific features. The top eight features that cause a decrease in prediction AUC for each intervention task. In general, physiological data were more important for the more invasive interventions — mechanical ventilation (4a, 4b) and vasopressors (4c, 4d) — while clinical note topics were more important for less invasive tasks — non-invasive ventilation (4e, 4f) and fluid boluses (4g, 4h). Note that all weaning tasks except for ventilation have significantly less AUC variance.

For colloid and crystalloid boluses we note general indicators of physiological decline, as boluses are given for a wide range of conditions.

"Hallucinations" for vasopressor and ventilation onset are shown in Figure 6. While our model was not trained with any physiological knowledge or priors, we note that it identifies blood pressure drops as being maximally activating for vasopressor onset, and respiratory rate decreasing for ventilation onset. This suggests that it is still able to independently learn physiological factors that are important for intervention prediction. We note that these hallucinations give us more insight into underlying properties of the network and what it is looking for. However, since these trajectories are made to maximize the output of the model, they do not necessarily correspond to physiologically plausible trajectories.

6. Conclusion

In this work, we targeted forward-facing prediction of ICU interventions covering multiple physiological organ systems. To our knowledge, our model is the first to use deep neural networks to predict both onset and weaning of interventions using all available modalities of ICU data. In our tasks, deep learning methods beat state-of-the-art AUCs reported in prior work for intervention prediction tasks — this is sensible given that prior works have focused on single targets with smaller datasets (Wu et al., 2016) or unsupervised representations prior to supervised training (Ghassemi et al., 2017). We also note that the LSTM over physiological words significantly improved perfor-
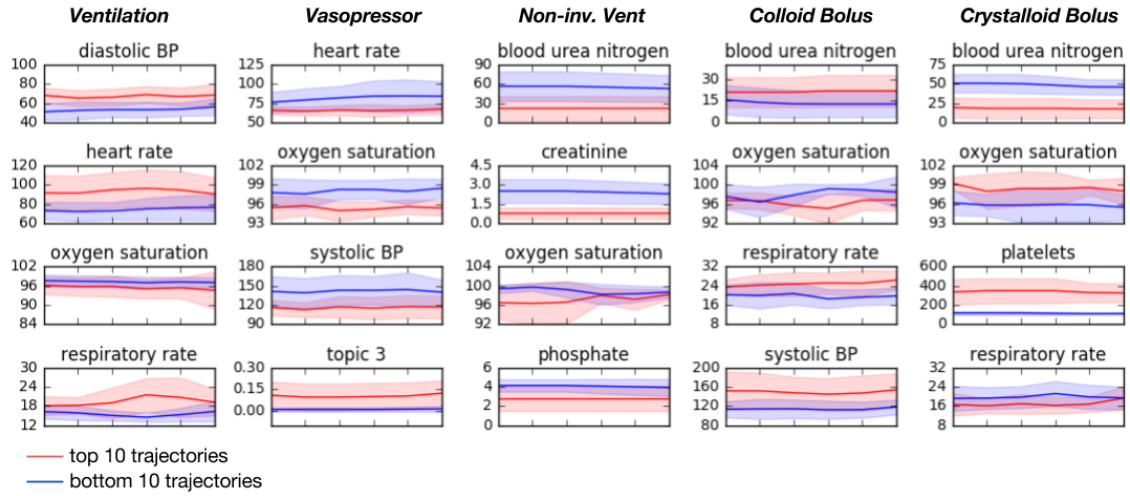
Figure 5: Trajectories of the 10 maximally and minimally activating examples for onset of each of the interventions.
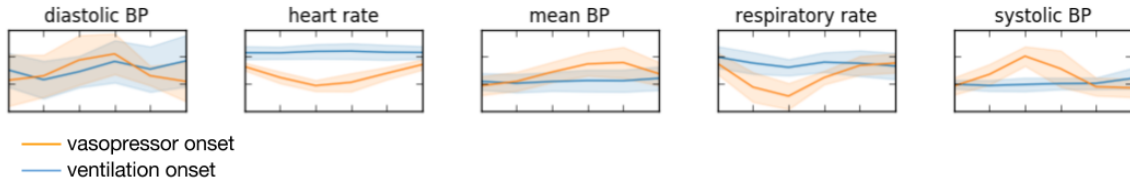

Figure 6: Trajectories generated by adjusting inputs to maximally activate a specific output node of the CNN.

mance in the two intervention tasks with the lowest incidence rate — possibly because this representation encodes important information about what is "normal" for each physiological value, or is more robust to missingness in the physiological data.

Importantly, we were able to gain interpretability in both models. In the LSTMs, we examined feature importance using occlusion, and found that physiological data were important in more invasive tasks, while clinical note topics were more important for less invasive interventions. This could indicate that there is more clinical discretion at play for less invasive tasks. We also found that all weaning tasks save ventilation had less AUC variance, which could indicate that these decisions are also made with a large amount of clinical judgment.

The temporal convolutions in our CNN filters over the multi-channel input learnt interesting and clinically-relevant trends in real patient trajectories, and these were further mimicked in the hallucinations generated by the network. As in prior work, we found that RNNs often have similar or improved performance as compared to CNNs Razavian et al. (2016). However, it is possible that more complex models would perform better as they uncover more long and short-term dependencies.

Our results are an interesting start to extracting interpretability from neural networks on patient data, and future work to expand this will enable these models to be adopted in real clinical settings.

Acknowledgments

References

Estevão Bassi, Marcelo Park, and Luciano Cesar Pontes Azevedo. Therapeutic strategies for high-dose vasopressor-dependent shock. *Critical care research and practice*, 2013, 2013.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3(5):993–1022, 2003.

Dustin Charles, Meghan Gabriel, and Michael F Furukawa. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2012. *ONC data brief*, 9:1–9, 2013.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *arXiv preprint arXiv:1606.01865*, 2016.

Edward Choi, Mohammad Taha Bahadori, and Jimeng Sun. Doctor AI: predicting clinical events via recurrent neural networks. *CoRR*, abs/1511.05942, 2015. URL http://arxiv.org/abs/1511.05942.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, pages 577–585, 2015.

Frederick DAragon, Emilie P Belley-Cote, Maureen O Meade, François Lauzier, Neill KJ Adhikari, Matthias Briel, Manoj Lalu, Salmaan Kanji, Pierre Asfar, Alexis F Turgeon, et al. Blood pressure targets for vasopressor therapy: A systematic review. *Shock*, 43(6):530–539, 2015.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. Technical report, University of Montreal, 2009.

Cristóbal Esteban, Oliver Staeck, Stephan Baier, Yinchong Yang, and Volker Tresp. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*, pages 93–101. IEEE, 2016.

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

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive

care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84. ACM, 2014.

Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *Proc. Twenty-Ninth AAAI Conf. on Artificial Intelligence*, 2015.

Marzyeh Ghassemi, Mike Wu, Michael Hughes, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. In *Proceedings of the AMIA Summit on Clinical Research Informatics (CRI)*, volume 2017. American Medical Informatics Association, 2017.

T. Griffiths and M. Steyvers. Finding scientific topics. In *PNAS*, volume 101, pages 5228–5235, 2004.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Yang N Jamoom E and Hing E. 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.

Rahul G Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

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

ML Malbrain, Paul E Marik, Ine Witters, Colin Cordemans, Andrew W Kirkpatrick, Derek J Roberts, and Niels Van Regenmortel. Fluid overload, de-resuscitation, and outcomes in critically ill or injured patients: a systematic review with suggestions for clinical practice. *Anaesthesiol Intensive Ther*, 46(5):361–80, 2014.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Marcus Müllner, Bernhard Urbanek, Christof Havel, Heidrun Losert, Gunnar Gamper, and Harald Herkner. Vasopressors for shock. *The Cochrane Library*, 2004.

Gustavo A Ospina-Tascón, Gustavo Luiz Büchele, and Jean-Louis Vincent. Multicenter, randomized, controlled trials evaluating mortality in intensive care: Doomed to fail? *Critical care medicine*, 36(4):1311–1322, 2008.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Pretten-hofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Narges Razavian, Jake Marcus, and David Sontag. Multi-task prediction of disease onsets from longitudinal lab tests. In *JMLR (Journal of Machine Learning Research): MLHC Conference Proceedings*, 2016.

Cátia M Salgado, Susana M Vieira, Luís F Mendonça, Stan Finkelstein, and João MC Sousa. Ensemble fuzzy models in personalized medicine: Application to vasopressors administration. *Engineering Applications of Artificial Intelligence*, 49:141–148, 2016.

Martin J Tobin. Principles and practice of mechanical ventilation, 2006.

Jean-Louis Vincent. Critical care-where have we been and where are we going? *Critical Care*, 17 (Suppl 1):S2, 2013.

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.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81, 2015.

Karl L Yang and Martin J Tobin. A prospective study of indexes predicting the outcome of trials of weaning from mechanical ventilation. *New England Journal of Medicine*, 324.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. URL http://arxiv.org/abs/1311.2901.

Appendix

A. Dataset Statistics

Table 3: Variables

Static Variables	Gender	Age	Ethnicity
	ICU	Admission Type	
Vitals and Labs	Anion gap	Bicarbonate	blood pH
	Blood urea nitrogen	Chloride	Creatinine
	Diastolic blood pressure	Fraction inspired oxygen	Glascow coma scale total
	Glucose	Heart rate	Hematocrit
	Hemoglobin	INR[*]	Lactate
	Magnesium	Mean blood pressure	Oxygen saturation
	Partial thromboplastin time	Phosphate	Platelets
	Potassium	Prothrombin time	Respiratory rate
	Sodium	Systolic blood pressure	Temperature
	Weight	White blood cell count	

[*]International normalized ratio of the prothrombin time

Table 4: Dataset Statistics

	Train	Test	Total
Patients	27,318	6,830	34,148
Notes	564,652	140,089	703,877
Elective Admission	4,536	1,158	5,694
Urgent Admission	746	188	934
Emergency Admission	22,036	5,484	27,520
Mean Age	63.9	64.1	63.9
Black/African American	1,921	512	2,433
Hispanic/Latino	702	166	868
White	19,424	4,786	24,210
CCU (coronary care unit)	4,156	993	5,149
CSRU (cardiac surgery recovery)	5,625	1,408	7,033
MICU (medical ICU)	9,580	2,494	12,074
SICU (surgical ICU)	4,384	1,074	5,458
TSICU (trauma SICU)	3,573	861	4,434
Female	11,918	2,924	14,842
Male	15,400	3,906	19,306
ICU Mortalities	1,741	439	2,180
In-hospital Mortalities	2,569	642	3,211
30 Day Mortalities	2,605	656	3,216
90 Day Mortalities	2,835	722	3,557
Vasopressor Usage	8,347	2,069	10,416
Ventilator Usage	11,096	2,732	13,828

B. Physiological Word Generation

See Figure 7.

Numerical

patient	hours in	glucose
3	1	NaN
3	2	NaN
3	3	101.2344
⋮	⋮	⋮

Physiological Words

patient	hours in	glucose_-2	glucose_-1	glucose_0	glucose_1	glucose_2
3	1	0	0	0	0	0
3	2	0	0	0	0	0
3	3	0	1	0	0	0
	⋮	⋮	⋮	⋮ ⋮	⋮	⋮

Figure 7: Converting data from continuous timeseries format to discrete "physiological words." The numeric values are first z-scored and rounded, and then each z-score is made into its own category. On the right, glucose_-2 indicates the presence of a glucose value that was 2 standard deviations below the mean. A row containing all zeros for a given variable indicates that the value for that variable was missing at the timestep.

C. LSTM Model Details

LSTM performs the following update equations for a single layer, given its previous hidden state and the new input:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{3}$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{4}$$
$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{5}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{6}$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{7}$$
$$h_t = o_t \odot tanh(c_t) \tag{8}$$

where $W_f, W_i, W_c, W_o \in \mathbb{R}^{L_1 \times (L_1 + V)}$, $b_f, b_i, b_c, b_o \in \mathbb{R}^{L_1}$ are learned parameters, and $f_t, i_t, \tilde{c}_t, c_t, o_t, h_t \in \mathbb{R}^{L_1}$. In these equations, σ stands for an element-wise application of the sigmoid (logistic) function, and $\odot$ is an element-wise product. This is generalized to multiple layers by providing h_t from the previous layer in place of the input.

We calculate classification loss using categorical cross-entropy, which sets the loss for predictions for N examples over M classes as:

$$\mathcal{L}(\hat{y}_1 \ldots \hat{y}_N) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \log \hat{y}_{ij}$$

where $\hat{y}_{ij}$ is the probability our model predicts for example i being in class j, and y_{ij} is the true value.

D. Generated Topics

Table 5: Most probable words in the topics most important for intervention predictions.

Topic	Top Ten Words	Possible Topic
Topic 1	pt care resp vent respiratory secretions remains intubated abg plan psv bs support settings cont placed changes note wean rsbi coarse cpap continue peep suctioned clear extubated rr mask weaned	Respiratory failure/infection
Topic 2	family pt ni care patient dnr stitle dr home daughter support team meeting wife son comfort note social doctor sw dni known time status hospital contact pt's work plan lastname	Discussion of end-of-life care
Topic 3	hr resp gi pt cont gu neuro bs cv id note abd soft bp today stool social noted progress clear remains nursing skin urine sats foley npn yellow stable ls	Multiple physiological changes
Topic 4	pain pt assessment response action plan control continue given dilaudid monitor chronic acute morphine iv po prn patient pca hr meds bp drain cont nausea ordered relief sbp pericardial assess	Assessments of patient responsiveness
Topic 10	pt intubated vent propofol sedation sedated fentanyl peep tube versed secretions abg wean remains continue ett suctioned plan ps increased extubation settings ac sounds min cpap sputum respiratory hr ogt	Continued need for ventilation
Topic 38	ml dl mg pm meq assessed icu ul total medications systems review pulse labs balance comments code hour rr min respiratory rhythm prophylaxis admission allergies blood urine mmhg status dose	Many labs tested
Topic 48	ed pt patient transferred hospital pain admitted denies admission days nausea received ago presented micu showed vomiting past reports history given blood bp old year arrival known osh diarrhea unit	Emergency admission/transfer patient

16