Switching State Space Modeling via Constrained Inference for Clinical Outcome Prediction

Arnold Su ARNOLDSU@MIT.EDU
Anna Wong ANNAWONG@ALUM.MIT.EDU
Ardavan Saeedi AV.SAEEDI@GMAIL.COM
Li-wei H Lehman LILEHMAN@MIT.EDU

Massachusetts Institute of Technology,
Cambridge, MA, USA

Abstract

In clinical settings, timely and accurate prediction of adverse patient outcomes can help guide treatment decisions. While deep learning models have demonstrated strong predictive performance, they often lack interpretability. To address this gap, we propose a framework that combines the predictive strength of a black-box discriminative model, such as a deep neural network, with the interpretability of a latent variable model. Specifically, we develop a constrained inference approach to train a switching state-space model—an autoregressive hidden Markov model (AR-HMM)—that learns interpretable discrete latent states from multivariate clinical time series, enabling the modeling of patient trajectories as transitions among these states while also achieving high predictive accuracy in downstream outcomes. Our method leverages knowledge distillation: a high-capacity LSTM "teacher" model is first trained to predict a target clinical outcome of interest, and its predictive behavior is then transferred to an interpretable AR-HMM "student" model through a similarity constraint during training. We use a constrained variational inference approach to estimate the parameters of the AR-HMM with a similarity preserving constraint, ensuring that input pairs with similar representation in the teacher model also have similar representation in the student model. We evaluated our approach using two real-world clinical datasets. Our approach demonstrates predictive performance comparable to state-of-the-art deep learning models, while producing interpretable latent trajectories that reflect clinically meaningful patient states.

Keywords: Switching state space models, constrained inference, knowledge distillation, deep learning, latent variable models, clinical time series, Autoregressive Hidden Markov Models (AR-HMM), interpretability.

1. Introduction

In order to accurately predict a patient's outcome in challenging settings such as during treatment of patients in critical care, one approach is to train a neural network to make these predictions, and previous approaches have demonstrated effectiveness of these deep learning models for clinical outcome prediction (Shamout et al., 2021; Rajkomar et al., 2018; Placido et al., 2023; Boussina et al., 2024; Xiong et al., 2024). While neural networks are powerful predictive models, they often lack interpretability important for many clinical applications.

Latent variable models, such as Switching Linear Dynamical Systems (SLDS) and their variants including the autoregressive Hidden Markov Models (AR-HMM) (Murphy, 1998; Ghahramani and Hinton, 2001; Fox et al., 2010, 2014), can be trained on clinical time series data to learn interpretable latent structures that capture a patient's evolving health states (Quinn and Williams, 2011; Lehman et al., 2012, 2013, 2018; Nemati et al., 2013) and provide prognostic insights for forecasting downstream outcomes (Lehman et al., 2012, 2015a). However, while the state representations learned from these models are interpretable and informative, they are typically not explicitly optimized for predicting specific outcomes.

In this paper, we present a modeling framework that combines the predictive strength of a black-box discriminative model, such as a deep neural network, with the interpretability of latent variable models to learn latent state representations from multivariate time series, enabling discovery of latent states that are both clinically meaningful and highly predictive of downstream outcomes. Our approach leverages knowledge distillation to distill the knowledge of a high-performing discriminative teacher model into a more interpretable latent variable student model. Specifically, in this work, the teacher is an LSTM trained for outcome prediction, while the student is an autoregressive hidden Markov model (AR-HMM) that learns discrete latent states from multivariate clinical time series. We train the AR-HMM using automatic differentiation variational inference (ADVI) (Kucukelbir et al., 2016), and introduce a similarity constraint to incorporate guidance from a high-performing deep learning model as the teacher model during training. This approach allows the AR-HMM to capture underlying temporal dynamics in the observed data while aligning its latent structure with the predictive representations of the neural network.

We evaluate the proposed method on two real-world datasets from patients with sepsis and respiratory failure in the MIMIC-IV database (Johnson et al., 2023). In addition to predicting mortality, we assess performance on several clinically important outcomes, including the development of pulmonary edema, initiation of dialysis, administration of diuretics, and the need for mechanical ventilation. Our results show that the proposed approach outperformed baselines without the constrained inference from knowledge distillation, and achieves predictive performance comparable to that of the deep learning approach, while preserving strong generative capabilities and interpretability through its latent state structure.

Generalizable Insights In this work, we demonstrate that constrained inference techniques through knowledge distillation can effectively bridge the gap between high-performing black-box deep learning models and interpretable latent variable models for time series data. Second, by leveraging switching state space models to capture temporally evolving physiological states, we show that interpretable and dynamic representations of patient health can be used not only for outcome prediction but also for identifying clinically meaningful states from multivariate clinical time series data.

2. Related Works

Interpretable Time Series Models A wide range of models have been developed for time series prediction, with varying approaches to interpretability. For example, deep learning models such as RETAIN (Choi et al., 2016), and various attention-based approaches (Song et al., 2018) offer feature-level interpretability by highlighting important variables or time steps that contribute to predictions. Nemati et al. (2018) proposed a machine learn-

ing model for early sepsis prediction in the ICU leveraging feature-based interpretability. In contrast, our approach combines deep learning and and switching state-space models to provide discrete state-based representations that capture evolving dynamics in patient trajectories. This structured interpretability enables insight into disease progression and patient states over time, which feature- or attention-based models may not explicitly model.

(Semi-)Supervised Learning with Probabilistic Graphical Models A common approach to learning interpretable models to predict final outcomes is a 2-stage approach. In the first stage, a graphical model is used to infer interpretable latent features for a dataset. These features are then passed into the second stage of the model, which is a discriminative model that predicts an outcome. Lehman et al. (2012, 2013, 2014, 2015a) used switching vector autoregressive (SVAR) processes—also referred to as autoregressive hidden Markov models (AR-HMMs)—to learn latent states from multivariate clinical time series; they showed that these latent states capture prognostic information relevant for predicting clinical outcomes, and that incorporating the time-averaged state probabilities per patient as input features alongside conventional acuity measures significantly improves performance for clinical outcome prediction. Subsequent studies (Lehman et al., 2015b,c, 2018; Wu et al., 2017; Ghassemi et al., 2017) also followed the same approach, using state probabilities from variants of AR-HMM for outcome prediction and reported similar findings. Another example of this 2-stage method used a graphical model to derive features that are then used as input for a linear regression to predict neuroticism and depression (Resnik et al., 2013). These types of models, while interpretable, are limited in their predictive performance in comparison to neural networks. Our proposed approach in this work can be used to improve the performance of these methods by distilling the knowledge of a pre-trained discriminative model into the generative model. Other approaches (Blei and McAuliffe, 2010; Chen et al., 2015) have been proposed with supervised objective that combining interpretable latent structure discovery with predictive modeling by integrating discriminative training into the generative framework. In (Nemati et al., 2013; Nemati and Adams, 2014, 2015), a supervised framework was proposed for gradient-based learning of outcome-discriminative dynamics in switching state space models. Stanculescu et al. (2014) present an AR-HMM for early detection of neonatal sepsis by modeling physiological event dynamics from NICU monitoring data—leveraging domain-informed inference.

Constrained Inference Another strategy to enhance model performance is constrained inference, where the posterior distribution over latent states is restricted to satisfy specific performance criteria. For instance, adding a discriminative constraint can improve predictive accuracy while preserving interpretability by incorporating a supervised loss into the objective function (Hughes et al., 2018). Saeedi et al. (2022) developed a general knowledge distillation framework in which a high-performing discriminative "teacher" model guides the learning of a more interpretable latent variable "student" model. The present work builds on Saeedi et al. (2022), but focuses on clinical time series modeling and outcome prediction using switching state space models as the "student" latent variable model.

3. Methods

We use knowledge distillation to develop models that achieve both strong predictive performance and interpretability. Building on the knowledge distillation framework proposed by (Saeedi et al., 2022), our approach leverages a high-capacity "teacher" model with strong predictive performance to guide the training of an interpretable switching state space "student" model for clinical time series modeling and outcome prediction. In this work, we first train a deep learning teacher model using LSTM (Hochreiter and Schmidhuber, 1997) to predict clinical outcomes, and then use an autoregressive hidden Markov model (AR-HMM) as the student model to learn discrete latent states from clinical time series (Lehman et al., 2015a). To learn this AR-HMM student model, we used a variational technique ADVI (Kucukelbir et al., 2016) to learn the global model parameters, including the initial state distribution, state transitions probabilities, and autoregressive emission parameters, and the similarity-based constraint was used to distill knowledge from the LSTM teacher to the student AR-HMM model.

3.1. AR-HMM

In an AR-HMM, each state is parameterized by a set of AR coefficients, a covariance matrix Σ, and a bias term b, so we can define the k-th state as $\theta_k = \{A_k, \Sigma_k, b_k\}$. In an AR-HMM process with order r, each observation depends on the r observations before it. Let $x_t^{(i)}$ be the observation vector of the i-th patient at time t, and let $z_t^{(i)}$ be the state of the corresponding Markov chain for that patient at time t. Let π_k be the transition probabilities for state k. Then, since this is a Markov chain, we know that $z_t^{(i)} \sim \pi_{z_{t-1}^{(i)}}$, for all $t > 1$. An order r AR-HMM process, denoted by VAR(r), is defined as follows

$$z_t^{(i)} \sim \pi_{z_{t-1}^{(i)}} \tag{1}$$

$$x_t^{(i)} = \sum_{l=1}^{r} A_l^{z_t^{(i)}} x_{t-l}^{(i)} + e_t^{(i)}(z_t^{(i)}) + b_{z_t^{(i)}} \tag{2}$$

$$\triangleq A_{z_t^{(i)}} \tilde{x}_t^{(i)} + e_t^{(i)}(z_t^{(i)}) + b_{z_t^{(i)}} \tag{3}$$

where $e_t^{(i)}(z_t^{(i)}) \sim \mathcal{N}(0, \Sigma_{(Z_t)})$ is the state-specific noise, $A_k = [A_1^k ... A_r^k]$ are the lag matrices, and $\tilde{x}_t^{(i)} = [x_{t-1}^{(i)\top} ... x_{t-r}^{(i)\top}]^\top$ are the observations.

Our AR-HMM also has $\mu_{init,k}$ and $\Sigma_{init,k}$ as parameters for approximating the distribution of an initial set of observations, where $x_{init} \sim \mathcal{N}(\mu_{init,k}, \Sigma_{init,k})$. And so in total, the k-th state corresponds to the parameters $\theta_k = \{A_k, \Sigma_k, b_k, \mu_{init,k}, \Sigma_{init,k}\}$.

3.2. Variational Inference

The mechanism behind many models that identify patterns in a model or make predictions is typically a computation of the posterior distribution of latent variables, given the observation, $p(z, \theta|x)$, where z is local latent variables, θ is global latent variables, and x is observations. However, it is often difficult to calculate the posterior, which leads to the use of approximate inference techniques like variational inference. A typical framework of

variational inference maximizes the Evidence Lower Bound (ELBO) with respect to the variational parameter ϕ_θ. We approximate the posterior distribution of global latent variables with a distribution parametrized by ϕ_θ, while the local latent variables z is exactly computed given θ and the observations x. We let $\Theta = T(\theta)$ be the transformed global latent variables, and $\phi_\theta = (\mu_1, ..., \mu_K, \omega_1, ..., \omega_K)$ represent the variational parameters in the unconstrained space of $\mathbb{R}^{2K}$, where the global latent variables are sampled independently such that variable i is sampled from the Gaussian distribution $\mathcal{N}(\mu_i, \exp(\omega_i)^2)$.

$$\mathcal{L}(\phi_\theta; x) \triangleq \mathbb{E}_{q_{\phi_\theta}(\theta)} \left[\log p(x, z^*, \theta) - \log q_{\phi_\theta}(\theta)\right] \tag{4}$$

$$z^* \triangleq \arg\max_z \, p(z \mid x, \theta) \tag{5}$$

For example, in an AR(1) process, consider a subset of the global latent variables, the lag matrix $A_k \in \mathbb{R}^{D \times D}$. Then the entries $(A_K)_{ij}$ are independently sampled from the distribution $\mathcal{N}(\mu_{ij}, \exp(\omega_{ij})^2)$ where μ_{ij} and ω_{ij} are the variational parameters corresponding to the variable $(A_k)_{ij}$.

The ELBO is a lower bound on the log-likelihood. To make our approach more accessible to a wider range of applications, our framework, similar to Saeedi et al. (2022), uses Automatic Differentiation Variational Inference (ADVI) to approximate our global latent variables θ. However, in contrast to Saeedi et al. (2022) which uses a recognition network ϕ_z to derive an approximate posterior q_{ϕ_z} on the local latent variables z, we derive the exact posterior $p_{z|\theta,x}$ for the local latent variables based on the estimated global parameters.

Global Latent Variables θ While there are many techniques that can be used to perform variational inference, we chose automatic-differentiation variational inference (ADVI) (Kucukelbir et al., 2016). It is a flexible black-box variational inference method that can be used for many different probabilistic models. It achieves this by transforming the K-dimensional latent variables θ such that they live in the real coordinate space, $\mathbb{R}^K : T : \text{supp}(p(\theta)) \to \mathbb{R}^K$, so ADVI can choose the variational distribution independent of the generative model. Note also the variational approximation in the original space can then be written as $q(\theta; \phi_\theta) = q(T(\theta); \phi_\theta)|\det(J_T(\theta))|$. Then, one can further assume a factorized Gaussian distribution is the variational approximation for the transformed latent variables: $q(T(\theta); \phi_{T(\theta)}) = \mathcal{N}(T(\theta); \mu, \text{diag}(\exp(\omega)^2))$, where $\phi_{T(\theta)} = (\mu_1, \ldots, \mu_K, \omega_1, \ldots, \omega_K)$ are the variational parameters in the unconstrained space. Note that these implicitly induce non-Gaussian variational distributions in the original latent variable space (Kucukelbir et al., 2016).

Local Latent Variables z Rather than using a recognition network or other proxy to approximate the local latent posterior, we directly use the observations to compute $p(z \mid x; \theta^{(t)})$ using Viterbi (Viterbi, 1967), a well-established AR-HMM training algorithm.

ELBO In AR-HMMs, the joint factorizes as

$$p(x_i, z, \theta) = p(x_i \mid z, \theta) p(z \mid \pi) p(\theta) \tag{6}$$

We let $\Theta = T(\theta)$ be the transformed global latent variable. In the ADVI transformation, this becomes

$$p(x_i, z, T^{-1}(\Theta)) = p(x_i \mid z, T^{-1}(\Theta)) p(z \mid \pi) p(T^{-1}(\Theta)) \tag{7}$$

Specifically, the global latent variational distribution transformation satisfies

$$q(\Theta, \phi_\Theta) = \mathcal{N}(\Theta; \mu, \text{diag}(\exp(\omega)^2)), \quad q(\theta; \phi_\theta) = q(T(\theta); \phi_{T(\theta)}) |\det(J_T(\theta))|$$

where $\Theta = T(\theta)$ are the transformed global latent variables, and $\phi_\theta = (\mu_1, ..., \mu_K, \omega_1, ..., \omega_K)$ are the variational parameters in the unconstrained space of $\mathbb{R}^{2K}$.

For example, in an AR order 1 process, consider a subset of the global latent variables, the lag matrix $A_k \in \mathbb{R}^{D \times D}$. Then the entries $(A_K)_{ij}$ are independently sampled from the distribution $\mathcal{N}(\mu_{ij}, \exp(\omega_{ij})^2)$ where μ_{ij} and ω_{ij} are the variational parameters corresponding to the variable $(A_k)_{ij}$.

We use ADVI to get posterior samplers of global latent variables to compute the ELBO, incorporating the global latent priors into the objective:

$$\begin{aligned}\mathcal{L}(\phi_\theta; x_i) \triangleq \mathbb{E}_{q_\phi}[&\underbrace{\log p(x_i \mid T^{-1}(\Theta), z) + \log p(z \mid \pi) + \log p(T^{-1}(\Theta))}_{\log(p(x_i, T^{-1}(\Theta), z))}] \\ &+ \log |\det(J_{T^{-1}}(\Theta))| \\ &+ \mathbb{H}(q_{\phi_\theta}(\Theta))\end{aligned}$$

using Monte Carlo approximation. Rather than using a recognition network to approximate the local latent distributions, we directly compute the local latent distributions given θ and the observations. Using the parameterization trick in Kucukelbir et al. (2016), we can rewrite the expectation in standard Gaussian density:

$$\begin{aligned}\mathcal{L}(\phi_\theta; x_i) \triangleq \mathbb{E}_{\mathcal{N}(\epsilon; I)}[&\log p(x_i \mid T^{-1}(\Theta_\epsilon), z) + \log p(z \mid \pi) + \log p(T^{-1}(\Theta_\epsilon))] \\ &+ \log |\det(J_{T^{-1}}(\Theta_\epsilon))| \\ &+ \mathbb{H}(q_{\phi_\theta}(\Theta))\end{aligned}$$

where $\Theta_\epsilon = \text{diag}(\exp(\omega)) \odot \epsilon_{1:K} + \mu$. We use ADVI to approximate the posterior over the global latent variables θ, including the state transition matrix π and the AR parameters for each of the K latent states. In ADVI, the variational distribution is defined over an unconstrained latent space $\mathbb{R}^d$, and a differentiable bijection T^{-1} is used to map these unconstrained latent variables back to the constrained space of model parameters. Specifically, the AR parameters in each state k are already unconstrained, and thus these parameters are mapped via the identity transform. Transition matrix π lies on the K-dim prob simplex. We apply softmax to unconstrained parameters associated with each row of π. Thus, the full bijection $T^{-1}(\Theta)$ combines the identity map for the AR parameters and a softmax-based transformation for the rows of the transition matrix.

Similarity Constraint We incorporated the knowledge distillation constraint by using a similarity-based constraint between the teacher and student models. Let C_t be the feature dimensionality of the teacher model, and C_s be the feature dimensionality of the student model. For a dataset of size N, we denote the feature representations of the teacher and student models as $F^t \in \mathbb{R}^{N \times C_t}$ and $F^s \in \mathbb{R}^{N \times C_s}$, respectively.

For the student model, we assume that every row of the feature representation is a function of the inferred latent variables. The knowledge distillation constraint is designed to make sure that the differences between the two feature representations is less than some tolerance level. More specifically, we compute the similarity of feature representations across patients by taking the dot product, which results in $N \times N$ matrices:

$$\tilde{F}^s = F^s \cdot F^{s\top} \text{ and } \tilde{F}^t = F^t \cdot F^{t\top} \tag{8}$$

Because our similarity loss is implemented through a pairwise similarity matrix, it offers flexibility by allowing teacher models to use feature spaces of different dimensionality than the student model. In this work, F^t denotes the hidden state representation from the teacher LSTM at the final timestep, with dimensionality C_t. The student feature representation, F^s, can in general be any function of the inferred latent variables. Motivated by prior findings in (Lehman et al., 2012, 2015a), we set F^s to be the time-averaged probabilities across the K inferred states for each patient, yielding a dimensionality of $C_s = K$.

We also apply a normalization to the matrices. After taking the dot product of the feature matrix, we normalize by dividing each column by the ℓ^2 norm of the row. Let us denote the final normalized similarity matrices as $\bar{F}^s$ and $\bar{F}^t$ for the student and teacher models, respectively. We calculate the similarity loss as

$$\text{similarity loss} = \gamma \frac{1}{N^2} ||\bar{F}^s - \bar{F}^t||^2 \tag{9}$$

where γ is a hyperparameter that specifies how much to weight the loss from this similarity constraint in the overall loss function. The final objective function is

$$\min_{\phi_\theta} -\mathcal{L}(\phi_\theta; x_i) + \text{similarity loss}, \tag{10}$$

where the variational objective is regulated by the knowledge distillation constraint so that we maximize ELBO while ensuring that the student model has similar pair-wise similarity in latent features as the teacher model. Then a gradient descent method is used to update global latent variables such that they simultaneously maximize ELBO. The inferred posterior of local latent variables are used to perform downstream predictions, and in this case, the marginal posterior of latent states in an AR-HMM model (Lehman et al., 2012, 2015a; Saeedi et al., 2022).

3.3. Data

To evaluate our approach, we used two separate datasets from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database (Johnson et al., 2023). In the sepsis cohort, we selected patients meeting the sepsis-3 criteria (Singer et al., 2016). After ensuring patients met all inclusion and exclusion criteria (see Appendix), we were left with 7,663 patients. Hospital mortality of the cohort is approximately 13%. We employed an 80% training, 10% validation, and 10% testing split for our dataset. A detailed description of our cohort and a full list of covariates are provided in Appendix E.

The second patient cohort includes individuals with respiratory failure who required mechanical ventilation (MV) for at least 24 hours in the ICU. To predict clinical outcomes,

we used data from the 48 hours immediately preceding the first MV weaning attempt. Patients whose weaning attempts occurred within the first 48 hours of ICU admission were excluded. After ensuring patients met the appropriate criteria, we were left with 4,256 patients. Hospital mortality of the cohort is approximately 26%, and 28-day mortality is approximately 31%. Again, we employed an 80% training, 10% validation, and 10% testing split for our dataset. See Appendix for the list of covariates included in this dataset.

3.4. Teacher Model

For our teacher model, we trained an LSTM on the patient data to predict mortality. On the sepsis cohort, the teacher model included a total of 46 features, which is more features than the student model. The extra features included in the LSTM but not the student model are shown in Table 12. On the MV cohort, the LSTM model included 48 features. Unlike the sepsis cohort, the student model also has 48 features. Note that the teacher model and the student model need not have the same input dimension. In our experimental setup, the teacher LSTM is fed with higher dimensional time series data as input to predict downstream outcomes. Our premise was that the knowledge learned by the teacher model could be transferred to the student model, without the student model needing as many covariates. We trained the LSTM using various seeds and hyperparameter settings, and used the model with the best validation AUROC for our teacher model.

3.5. Student Model

For our student models, we used an autoregressive hidden Markov Model (AR-HMM) of order 1, with D-dimensional Gaussian distribution, where D is the number of covariates used for each patient's input to the model. In our models, $D = 28$ for the sepsis cohort, and $D = 48$ for the mv cohort . For each patient, we have a $D \times T$ vector input, where T is the number of timesteps. There are K possible latent states learned by the AR-HMM.

The baseline model was a basic AR-HMM without knowledge distillation or supervision, run for 20 iterations. The model that incorporates knowledge distillation via a similarity constraint is denoted as KD-AR-HMM. We also implemented a model with a discriminator constraint (DISC-AR-HMM) for performance comparison with our approach. The models with constraints were also ran for 20 iterations.

3.6. Outcome Prediction

For the baseline AR-HMM and the KD-AR-HMM, we used a logistic regression model that took as input the features outputted by the AR-HMM model, and outputted the probability of an outcome such as mortality. For the models with the discriminator constraint, we used the trained discriminator network to make predictions. In addition to predicting mortality, we also used the same set of features to predict other patient outcomes such as the development of pulmonary edema, or the need for dialysis, mechanical ventilation, or diuretics. We adjusted various hyperparameters in order to tune the model. A description of our hyperparameter search can be found in Appendix D.

3.7. Discriminative model

We compare the KD-AR-HMM performance to a baseline AR-HMM and discriminative model. The discriminative AR-HMM (DISC-AR-HMM) is trained in the same way as the KD-AR-HMM except the DISC-AR-HMM objective replaces the similarity loss in the KD-AR-HMM objective with cross-entropy loss to regularize toward better classification.

4. Results

4.1. Sepsis Outcome Prediction

4.1.1. Sepsis Mortality Prediction

We use our models on the first patient cohort dataset to predict patient mortality in sepsis patients. The results presented in this paper have the number of states as $K = 5$ and autoregressive order 1 (AR(1)) for the student AR-HMM models. We also tried using 10 states, and higher order autoregression (order 2 and order 3), and the KD-AR-HMM achieved similar performance as the 5 state, AR(1) version. In Table 1, the KD-AR-HMM consistently outperformed other baselines. We ran each model with 40 different seeds and hyperparameter settings, and chose the best hyperparameter setting based on the validation AUC. Another 10 different seeds with a fixed hyperparameter setting was ran. The results presented are from the 10 seed experiments. The baseline AR-HMM had a low AUROC of 0.642. The model with the discriminator constraint performed similarly with an AUROC of 0.648. The model that incorporated knowledge distillation performed the best, with an AUROC of 0.792.

The pairwise similarity matrices for the student and teacher models along with those for the baselines are presented in Fig 1. Pairwise similarity matrices for the test dataset of the disease subtyping experiment: Each row and each column corresponds to a patient in the test dataset. Brighter colors indicate higher similarity values. The distillation constraint encourages the pairwise similarity matrix of the student model (KD-AR-HMM) to be similar to that of the teacher model from LSTM. Compared to the the basic AR-HMM which is unsupervised, and the discriminative version of AR-HMM, the matrix from the knowledge distillation via constrained inference is more similar to that of the teacher.

Table 1: MIMIC Sepsis Test Set Performance (test AUROC, AUPRC and Log Likelihood) in hospital mortality prediction over 10 seeds. For KD-AR-HMM, we use the LSTM model with the best validation AUC as the LSTM Teacher model (which has a test AUC of 0.851).

Model	**AUROC**	**AUPRC**	**Log Likelihood**
LSTM (Teacher)	0.851	0.545	N/A
LSTM (10-seeds)	0.836 ±0.026	0.497 ± 0.080	N/A
AR-HMM (2-stage)	0.642 ± 0.002	0.280 ±0.004	5.560 ± 1.809
DISC-AR-HMM	0.648 ± 0.193	0.387 ±0.124	-981.971 ± 1115.78
KD-AR-HMM	**0.792** ± **0.009**	**0.530** ±0.006	-256.290 ± 19.465

In addition to measuring the AUROC to evaluate the predictive ability of our models, we measured the log likelihood to evaluate the generative ability of the models. Compared to the baseline AR-HMM, the constrained models had comparable log likelihoods, indicating

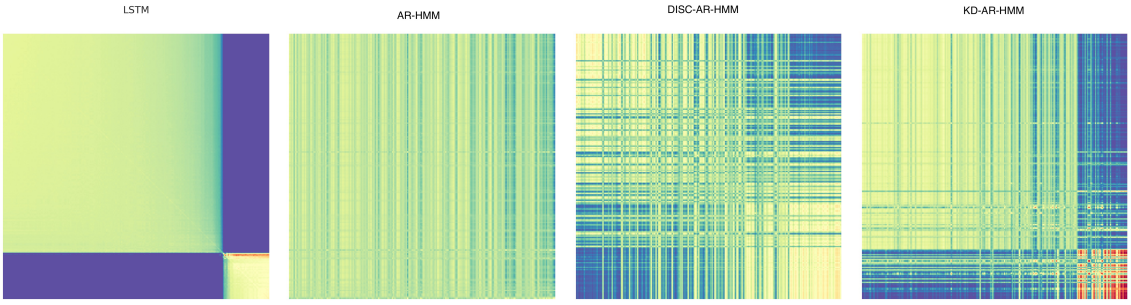


Figure 1: Generated pair-wise similarity matrices representing the feature representations of the teacher LSTM, student KD-AR-HMM model and baseline approaches for sepsis hospital mortality prediction. From left to right: LSTM, AR-HMM, DISC-AR-HMM, KD-AR-HMM.

that the constraints did not significantly decrease the model fit for the data. In particular, the log likelihood of the KD-AR-HMM was the closest to the baseline. A summary of these metrics can be found in Table 1.

Model Selection. We explored several variants of the KD-AR-HMM model by varying the number of latent states and the AR order. Full model selection results are provided in Appendix B. While all variants achieved comparable validation AUROC scores, the KD-AR-HMM with 5 latent states and an AR order 1 yielded slightly superior validation log-likelihood and Bayesian Information Criterion (BIC) scores. Consequently, this configuration was selected as the final model for reporting primary test set performance and for clinical interpretation presented in the main text.

4.1.2. Sepsis Cohort: Other Clinical Outcomes Prediction

In this section, we assess whether our knowledge distillation framework is applicable to other outcomes. We train an LSTM and KD-AR-HMM for each individual outcome. The other outcomes we tested was the onset of pulmonary edema, and the patient's need to start diuretics, mechanical ventilation (MV), or dialysis. The LSTM was ran with 40 different seeds and hyperparameter settings, and chose the best hyperparameter setting based on the validation AUC. Another 10 different seeds with a fixed hyperparameter setting was ran. As for the KD-AR-HMM, 10 different seeds were ran with top 10 hyperparameter settings from the model's results from hyperparameter tuning on mortality. Then, as per usual, another 10 different seeds with a fixed hyperparameter setting was ran. The results presented are from the 10 seed experiments.

Similar to predicting mortality, in Table 2, we find that the KD-AR-HMM consistently outperformed other baselines. Furthermore, when predicting other outcomes, compared to the baseline AR-HMM, the constrained models had comparable log likelihoods, indicating that the constraints did not significantly decrease the model fit for the data. In particular, the log likelihood of the KD-AR-HMM was the closest to the baseline. Like mortality prediction, we note that the baseline DISC-AR-HMM had one or two seeds with high

performing models, but was highly inconsistent and unstable across multiple seeds. A summary of the log-likelihood and AUPRC results can be found in Table 7 and Table8 in the Appendix.

Table 2: MIMIC Sepsis cohort: Performance (AUROCs) in Outcome Predictions. Mean and standard deviation of test AUROCs averaged over 10 seeds reported. MV = Mechanical Ventilation. Edema refers to pulmonary edema. See Appendix for AUPRC.

	Edema	**Dialysis**	**MV**	**Diuretics**
LSTM	0.898 ± 0.025	0.841 ± 0.013	0.974 ± 0.006	0.940 ± 0.053
AR-HMM	0.581 ± 0.006	0.666 ± 0.000	0.827 ± 0.001	0.641 ± 0.005
DISC-AR-HMM	0.559 ± 0.144	0.591 ± 0.189	0.626 ± 0.351	0.677 ± 0.297
KD-AR-HMM	**0.665 ± 0.006**	**0.739 ± 0.007**	**0.884 ± 0.004**	**0.951 ± 0.012**

4.2. Respiratory Failure Cohort Outcome Prediction

We applied the knowledge distillation framework to a second cohort of patients with respiratory failure, focusing on predicting 28-day mortality following hospital discharge after mechanical ventilation weaning attempts (with 31% 28-day mortality rate). As shown in Table 3, the proposed KD-AR-HMM model outperforms the baselines in AUROC, achieving an average test AUROC of 0.649 compared to 0.543 for the standard AR-HMM (trained in a two-stage fashion without distillation) and 0.605 for the DISC-AR-HMM (which applies a discriminative constraint without knowledge distillation). The LSTM model with the best validation performance (used as the teacher model) achieves a test AUROC of 0.659, with an average AUROC across all seeds of 0.630, indicating that KD-AR-HMM performs consistently close to the teacher while substantially improving over the generative baselines. KD-AR-HMM's AUPRC performance (0.475) is slightly below that of DISC-AR-HMM (0.485), though the teacher LSTM also shows a marginally lower AUPRC (0.432), likely due to class imbalance of the dataset.

Table 3: Respiratory failure dataset: performance in 28-day post-hospital discharge mortality prediction. Table shows mean and standard deviation from test set AUROC, AUPRC and log likelihood over 10 seeds.

Model	**AUC**	**AUPRC**	**Log Likelihood**
LSTM (Teacher)	0.659	0.432	N/A
LSTM (10-seeds)	0.630 ± 0.018	0.431 ± 0.001	N/A
AR-HMM	0.543 ± 0.001	0.364 ± 0.004	-600.083 ± 14.218
DISC-AR-HMM	0.605 ± 0.113	0.485 ± 0.082	-4525.252 ± 2502.313
KD-AR-HMM	0.649 ± 0.019	0.475 ± 0.018	-4291.738 ± 403.770

5. Analyses and Interpretation

5.1. Model Selection

The KD-AR-HMM model selected for our analyses and interpretation is the model with the best performing validation AUC after training the model with 10 random seed initialization. These 10 seeds all had fixed hyperparameters, which were chosen also via validation AUC in a random 40 seed grid search for hyperparameter tuning. Full results are presented in Appendix B. For clinical interpretation, we selected the KD-AR-HMM model with the best validation AUC (validation AUC of 0.812 and test AUC of 0.805, with validation log-likelihood of -272.69, and a test log-likelihood of -261.98) for outcome association analyses and state trajectory visualization.

5.2. Outcome Association Analyses of Learned States

We performed logistic regression analyses to identify KD-AR-HMM states significantly associated with clinical outcomes (see Appendix for details). To assess associations with hospital mortality, we examined odds ratios (ORs) and corresponding p-values for each state. Logistic regressions were fitted on the test set, using the state probabilities generated by KD-AR-HMM as predictors and mortality as the binary outcome. We conducted univariate logistic regressions for each state–outcome pair and applied a False Discovery Rate (FDR) adjustment to account for multiple testing across states and outcomes. For KD-AR-HMM, **state 2** was identified as a low-risk state: higher occupancy probability in state 2 was significantly associated with lower odds of mortality, edema, and dialysis. In contrast, **states 1 and 3** were identified as high-risk states, where increased state probabilities were significantly associated with higher odds of mortality. Figure 2 shows simulated trajectories for the three states most strongly associated with hospital mortality risk, defined by the smallest adjusted p-values.

Table 4: Hospital Mortality: Odds ratio (OR) of individual states of the KD-AR-HMM model, using state marginals via the latent posterior given the final learned global model parameters.

State	OR (upper/lower)	Adj p-value
1	1.036 (1.021, 1.051)	**<0.001**
2	0.952 (0.943, 0.960)	**<0.001**
3	1.044 (1.034, 1.055)	**<0.001**
4	1.034 (1.005, 1.064)	**0.024**
5	0.999 (0.975, 1.025)	**0.996**

5.3. Clinical Relevance of Learned Latent States

In this section, we evaluate whether the learned latent states correspond to clinically meaningful patterns, particularly in relation to patient outcomes. Figure 2 presents simulated trajectories for three latent states identified as significantly associated with hospital mortality in the sepsis cohort. Based on their ranked odds ratios from the outcome association analysis (see above), we label States 3, 1, and 2 as high-, medium-, and low-risk states, respectively.

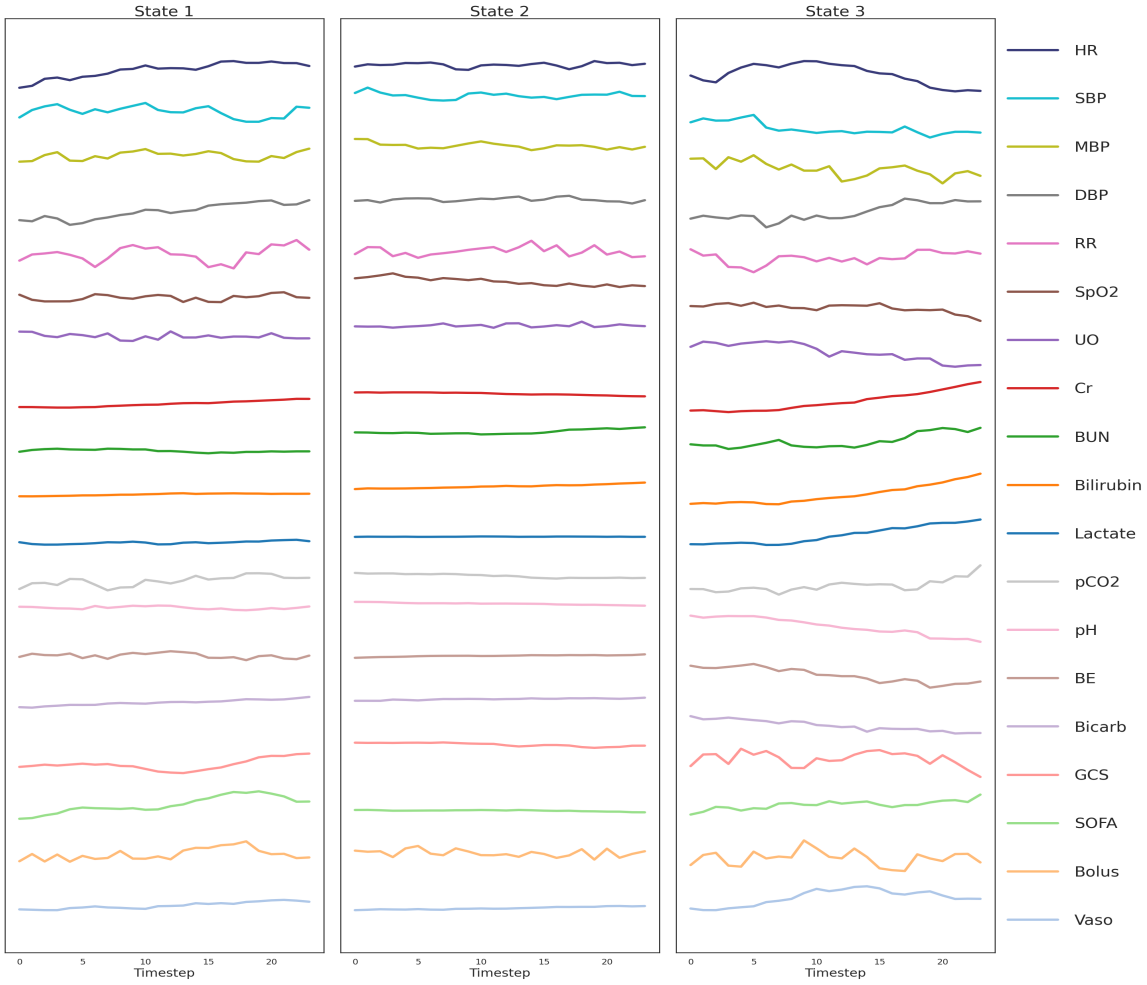


Figure 2: KD-AR-HMM simulated trajectories of selected covariates for medium-risk (State 1) and high-risk (State 3) vs lower-risk (State 2) states learned from the sepsis dataset. Each state was simulated 200 times by drawing samples from the learned AR model parameters, and the multivariate trajectory with the highest log-likelihood under the model was selected for visualization. All states were simulated and plotted with the same time duration and amplitude scale.

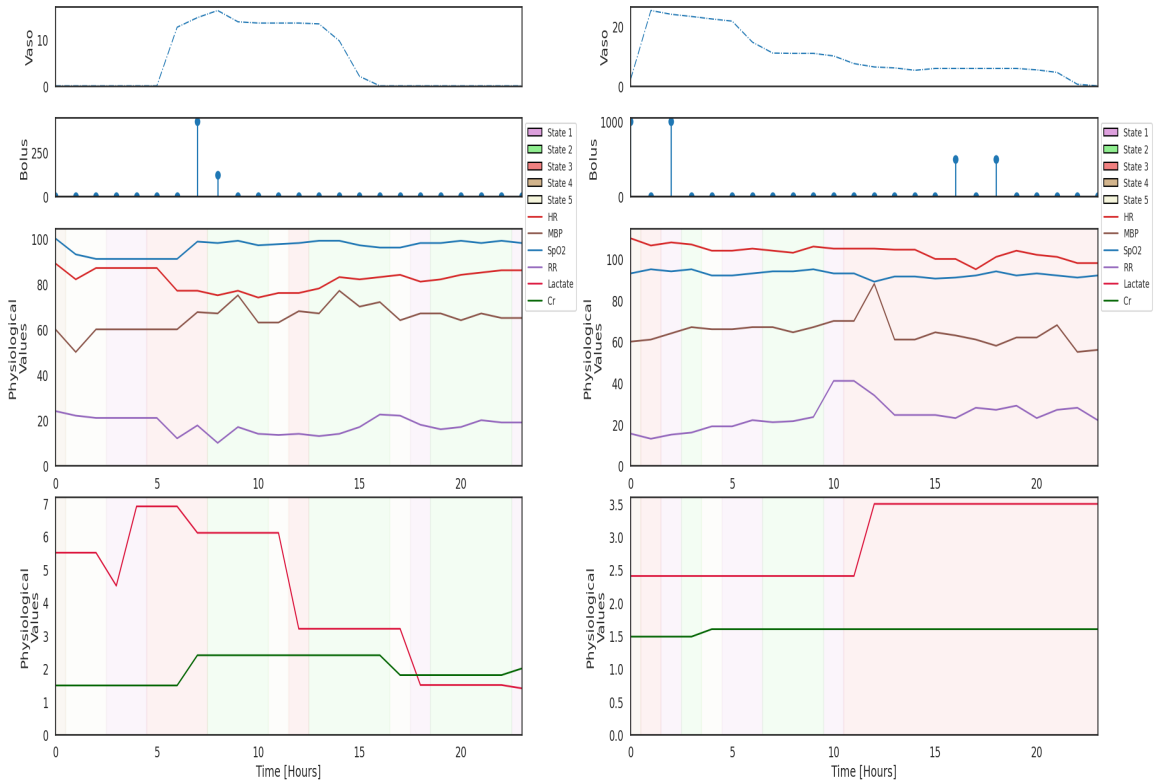
State 3 (high-risk) exhibits the highest odds ratio for hospital mortality (1.044 [1.034, 1.055]), and is characterized by declining trends in mean arterial blood pressure (MAP), SpO2, urine output (UO), pH, base excess (BE), and bicarbonate (HCO_3), along with increasing trends of creatinine, blood urea nitrogen (BUN), bilirubin, lactate, and pCO_2. Notably, we also observe a rising trend in vasopressor administration.

State 1 (medium-risk) is also associated with increased hospital mortality, though with lower odds than State 3. Its simulated trajectories show more moderate physiological deviations, lacking the pronounced deteriorations seen in State 3. Compared to the low-risk

State 2, State 1 is associated with lower levels of SpO2, pH, and Glasgow Coma Scale (GCS) scores, and higher SOFA scores, indicating a more compromised clinical profile.

State 2 (lower-risk) representing the low-risk state, generally maintains stable values across most physiological variables. Compared to the higher-risk states, it shows less variability and fewer extreme trends, suggesting more stable patient conditions. In contrast, States 1 and 3 exhibit greater variability and pronounced deviations in key variables, consistent with clinically unstable states.

The significance of latent states are also qualitatively supported by observing the generated latent state distribution across all patients in the dataset. In Figure 4 (in Appendix), we see that state 2 has a much less likely to be found in patients who died, than for patients who lived. On the contrary, states 1 and 3 were more likely in patients who died than patients who lived.



(*a*) Patient Survived (*b*) Patient Died

Figure 3: Example trajectories of a patient who (a) survived and (b) died in the hospital. Time series measurements plotted in original units before normalization. Fluid bolus (mL). Top panel: SpO2 (light blue, %), heart rate (HR, red, beats/min), mean blood pressure (MBP, brown, mmHg), respiratory rate (RR, purple, breath/min). Bottom panel: lactate (red, mmol/L), creatinine (Cr, green, mg/dL).

Individual Patient Case Studies: State Assignment over Time for Individual Patients Figure 3 shows an example of two test set patients, one who died in the hospital and another who survived the hospital stay, and the corresponding inferred states throughout the first 24 hours. The patient who survived hospital stay had a few hours of being in the medium- and high-risk states (states 1 and 3, in light purple and red) in the first half of their stay, but transition to a lower-risk state (state 2, green). We can also see that the patient's return to the low-risk state seems to coincide with an increase in SpO2, stabilization of blood pressure in response to fluids and vasopressor administration, and a decrease in lactate. On the other hand, the patient in (b) started off with a small proportion of time in the lower-risk state 2 (green), but transition to spent significant amount of time in the high and medium risk state 3 and 1 (red and purple). We can also see that this patient had an increasing trend in respiratory rate and lactate, with low blood pressure.

6. Discussions and Conclusion

In this work, we introduced a constrained inference framework that combines the predictive strength of deep neural networks with the interpretability of latent variable models for clinical outcome prediction. Specifically, we leveraged knowledge distillation to distill the knowledge from a high-performing neural network (teacher) to an autoregressive hidden Markov model (student), enabling the student model to learn discrete, interpretable state representations from multivariate clinical time series while achieving high predictive performance for downstream clinical outcomes. By incorporating a similarity constraint within a variational inference framework, we guided the AR-HMM to learn latent state dynamics that are predictive of the downstream clinical outcomes, while preserving the model's interpretability and generative performance.

We evaluated our method on the MIMIC-IV database across two clinically relevant tasks: predicting hospital mortality and fluid overload in sepsis and in respiratory failure patients undergoing mechanical ventilation. In both cohorts, the proposed KD-AR-HMM trained via constrained inference achieved improved predictive performance in comparison to the baselines, while providing clinically interpretable latent states that are prognostic of downstream outcomes. These findings suggest the potential of our method for real-time risk monitoring and decision support, where both accuracy and interpretability are important.

More broadly, this work highlights a general strategy for enhancing the interpretability of machine learning models in healthcare by integrating black-box predictors with structured probabilistic models. It also demonstrates the value of knowledge distillation via constrained inference as a flexible framework for training interpretable models that achieve competitive predictive performance.

Future work will address several limitations of the current study. First, our evaluation was limited to a single critical care database; validating the approach across diverse patient populations and institutions will be important for assessing generalizability. Second, we selected LSTM as the teacher model due to its strong performance and widespread use in clinical time series modeling, but future work will explore alternative architectures as teacher models. Future directions also include incorporating structured clinician review and further validating the clinical relevance of the learned latent states to assess their utility and relevance in real-world settings. Finally, we aim to extend this framework to

sequential decision-making tasks, where interpretable latent states could support treatment policy evaluation and individualized treatment planning.

7. Acknowledgments

The authors are grateful for Professor Roger Mark, our clinical collaborators and the reviewers for their valuable feedback, and Fareed Sheriff at MIT for his technical assistance. The authors acknowledge NIH grants R21HL177773, R01HL181348, and R01EB030362.

References

David M Blei and Jon D McAuliffe. Supervised topic models. *arXiv preprint arXiv:1003.0783*, 2010.

A. Boussina, S.P. Shashikumar, A. Malhotra, et al. Impact of a deep learning sepsis prediction model on quality of care and survival. *npj Digital Medicine*, 7:14, 2024. doi: 10.1038/s41746-023-00986-6.

Jianshu Chen, Ji He, Yelong Shen, Lin Xiao, Xiaodong He, Jianfeng Gao, Xinying Song, and Li Deng. End-to-end learning of lda by mirror-descent back propagation over a deep architecture. *arXiv preprint arXiv:1508.03398*, 2015.

Edward Choi, Mohammad Taha Bahadori, Annie Schuetz, Walter Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Bayesian nonparametric methods for learning markov switching processes. *IEEE Signal Processing Magazine Special Issue*, 2010. URL https://people.eecs.berkeley.edu/~jordan/papers/fox-etal-ieeespm10.pdf.

E. Fox, M. Hughes, E. Sudderth, and M. Jordan. Joint modeling of multiple time series via the beta process with application to motion capture segmentation. *The Annals of Applied Statistics*, 8(3), 2014. URL https://doi.org/10.1214/14-AOAS742.

Zoubin Ghahramani and Geoffrey E. Hinton. Variational learning for switching state-space models. In Michael I. Jordan, editor, *Graphical Models: Foundations of Neural Computation*, chapter 13. MIT Press, 2001.

M Ghassemi, M Wu, MC Hughes, P Szolovits, and F Doshi-Velez. Predicting intervention onset in the icu with switching state space models. In *AMIA Summits on Translational Science Proceedings*, 2017.

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

M. Hughes, G. Hope, L. Weiner, T. McCoy Jr, R. Perlis, E. Sudderth, and F. Doshi-Velez. Semi-supervised prediction-constrained topic models. In *AISTATS*, pages 1067–1076, 2018. URL https://proceedings.mlr.press/v84/hughes18a.html.

AEW Johnson, L Bulgarelli, L Shen, A Gayles, A Shammout, S Horng, TJ Pollard, B Moody, B Gow, LH Lehman, LA Celi, and RG Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Nature Scientific Data*, 2023.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference, 2016.

L. Lehman, R. Adams, L. Mayaud, G. Moody, A. Malhotra, R. Mark, and S. Nemati. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. *IEEE Journal of Biomedical and Health Informatics*, 19(3), 2015a. URL https://ieeexplore.ieee.org/document/6846269.

L. Lehman, R. Mark, and S. Nemati. A model-based machine learning approach to probing autonomic regulation from nonstationary vital-sign time series. *IEEE Journal of Biomedical and Health Informatics*, 22(1), 2018. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5896770/.

L. H. Lehman, Shamim Nemati, Ryan P Adams, and Roger G Mark. Discovering shared dynamics in physiological signals: Application to patient monitoring in ICU. In *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5939–5942, 2012.

L. H. Lehman, Shamim Nemati, Ryan P Adams, George Moody, Atul Malhotra, and Roger G Mark. Tracking progression of patient state of health in critical care using inferred shared dynamics in physiological time series. In *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013.

L. H. Lehman, S. Nemati, G. B. Moody, T. Heldt, and R. G. Mark. Uncovering clinical significance of vital sign dynamics in critical care. In *Proceedings of the Computing in Cardiology*, 2014.

L. H. Lehman, S. Nemati, and R. G. Mark. Hemodynamic monitoring using switching autoregressive dynamics of multivariate vital sign time series. In *Proceedings of the Computing in Cardiology*, 2015b.

LH Lehman, MJ Johnson, S Nemati, RP Adams, and RG Mark. Bayesian nonparametric learning of switching dynamics in cohort physiological time series: application in critical care patient monitoring. *Advanced State Space Methods for Neural and Clinical Data*, pages 257–282, 2015c.

Kevin Murphy. Learning Switching Kalman Filter Models. Technical report, Compaq Cambridge Research Lab Tech Report, 1998.

S Nemati, LH Lehman, and RP Adams. Learning outcome-discriminative dynamics in multivariate physiological cohort time series. In *IEEE EMBC*, 2013.

Shamim Nemati and Ryan P Adams. Supervised learning in dynamic bayesian networks. In *Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Representation Learning*, 2014.

Shamim Nemati and Ryan P Adams. Identifying outcome-discriminative dynamics in multivariate physiological cohort time series. *Advanced State Space Methods for Neural and Clinical Data*, pages 283–301, 2015.

Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D. Stanley, Gari D. Clifford, and Timothy G. Buchman. An interpretable machine learning model for accurate prediction of sepsis in the icu. *Critical Care Medicine*, 46(4):547–553, April 2018. doi: 10.1097/CCM.0000000000002936.

D. Placido, B. Yuan, J.X. Hjaltelin, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature Medicine*, 29:1113–1122, 2023. doi: 10.1038/s41591-023-02230-8.

John A. Quinn and Christopher K.Ĩ. Williams. *Physiological Monitoring with Factorial Switching Linear Dynamical Systems*, pages 182–204. Cambridge University Press, Cambridge, UK, 2011.

A. Rajkomar, E. Oren, K. Chen, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1:18, 2018. doi: 10.1038/s41746-018-0029-1.

P. Resnik, A. Garron, and R. Resnik. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1348–1353, 2013. URL https://aclanthology.org/D13-1133/.

A. Saeedi, Y. Utsumi, L. Sun, K. Batmanghelich, and L. wei Lehman. Knowledge distillation via constrained variational inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8132–8140, 2022. URL https://doi.org/10.1609/aaai.v36i7.20786.

Gideon E Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978. doi: 10.1214/aos/1176344136.

F Shamout, T Zhu, and DA Clifton. Machine learning for clinical outcome prediction. *IEEE Rev Biomed Eng.*, 2021.

M. Singer, C.S. Deutschman, C.W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, and et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, February 2016. doi: https://doi.org/10.1001/jama.2016.0287.

Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Ioan Stanculescu, Christopher K. I. Williams, and Yvonne Freer. Autoregressive hidden markov models for the early detection of neonatal sepsis. *IEEE Journal of Biomedical and Health Informatics*, 18(5):1560–1570, 2014.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967. doi: 10.1109/TIT.1967.1054010.

M Wu, M Ghassemi, M Feng, L Celi, P Szolovits, and Doshi-Velez F. Understanding vasopressor intervention and weaning: risk prediction in a public heterogeneous clinical time series database. *Journal of the American Medical Informatics Association*, 24(3):488–495, 2017.

Hong Xiong, Feng Wu, Leon Deng, Megan Su, and Li-wei H Lehman. G-Transformer: Counterfactual Outcome Prediction under Dynamic and Time-Varying Treatment Regimes. In *Proceedings of Machine Learning for Healthcare*, 2024.

Appendix A. Data and Cohort

In this work, we used data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database, which contains medical records from hospital admissions and ICU stays at the Beth Israel Deaconess Medical Center (BIDMC) (Johnson et al., 2023). For our patient cohort, we selected patients meeting the sepsis-3 criteria. Under this criteria, a patient is defined to have sepsis if they have both an episode of suspected infection and a Sequential Organ Failure Assessment (SOFA) score of 2 or more points. An episode of suspected infection is defined as either (a) an antibiotic was given and a culture was sampled within 24 hours or (b) a culture was sampled and an antibiotic was administered within 72 hours.

The dataset excludes patients whose time of suspected infection was more than 24 hours after ICU admission. It also excludes patients who were admitted after cardiac, vascular, or trauma surgery since those surgeries pose risks that could lead to different mortality outcomes. Additionally, if a patient had more than one ICU stay, only the first stay was used. The dataset also excludes patients who did not have documented pre-ICU fluids. Finally, we removed patients who died within 24 hours of entering the ICU, and patients who did not have all 24 hours of data.

After ensuring patients met all of these criteria, we were left with 7,663 patients. We put 80% of patients in the training set, and 10% in each of the validation and testing sets. The mean age of patients was 65.10, the median age was 67.0, and 4135 of the patients were male.

For each patient, we have 24 hours of hourly data, with covariates including heart rate, blood pressure, SOFA score, and other clinical variables such as glucose, creatinine, potassium, and more. We also have information about the actual treatment given to these patients at each hour, such as if patients are on mechanical ventilation or dialysis and the amount and dosage (if any) of fluids, vasopressors, and diuretics given. Finally, we have information about outcomes such as if the patient has pulmonary edema, is on diuretics,

dialysis, or mechanical ventilation, or if they die in the hospital. We fill in missing covariate values by either extending the previous covariate measurement for that patient if it exists, or filling it in with the population average value for that covariate.

For each feature, we define a valid range using the values in the dataset, as well as general domain-knowledge based expectations. Any data point outside the range is considered an outlier, and are then removed from the dataset. Next, missing values were filled using the most recent non-missing value from earlier in the sequence (ie) forward fill). This helps preserve temporal continuity in the time-series or sequential data. In cases where missing value occurred at the very beginning of the sequence, the missing entry was instead filled using the mean of that feature from the training set. Finally, each feature is normalized via either standard z-score normalization, log-transformed z-score normalization, or Yeo-Johnson power transformer based on the data distribution and histogram inspection of the dataset.

Other Clinical Outcomes. The final datasets had the following total number of remaining cohort (after excluding patients who already have the target outcome during first 24-hours of their ICU stay), and percentage of patients with the adverse outcomes within the cohort (i.e. patients who ultimately experienced that outcome within 48 hours after their first day in the ICU): 5,034 (19.5%) for pulmonary edema, 7,462 (2.3%) for dialysis, 4,166 (5.5%) for mechanical ventilation, and 6,413 (15.9%) for diuretics.

Appendix B. KD-AR-HMM Model Selection and Additional Results

In Table 5 we show the performance of different variants of the KD-AR-HMM model. We tested different total number of latent states ($K = 5$ vs. $K = 10$), as well as different AR order (AR(1) vs. AR(2)). The three models, KD-AR(1)-HMM with 5 states, KD-AR(2)-HMM with 5 states, and KD-AR(1)-HMM with 10 states exhibit very similar average test AUCs of 0.792, 0.788, and 0.787 respectively. KD-AR(1)-HMM with 5 states, however, had better performing log likelihood and BIC scores. Thus, leading us to choose KD-AR(1)-HMM with 5 states as our student model.

BIC, or the Bayesian information criterion, is a penalized likelihood metric that balances model fit and complexity. The BIC is formally defined as

$$BIC = k \ln(n) - 2 \ln(L)$$

where L is the maximized likelihood of the model, k is the number of parameters, and n is the number of observations. Lower BIC values indicate a more simple and better-fitting model. For us, we use BIC to help select the AR-HMM variants, by penalizing more complex models (Schwarz (1978)).

For model selection, in general the best model is chosen by taking the one with the best performing validation AUC from 10 seeds. In Table 6 we display the full results of the 10-seed experiment for our KD-AR-HMM student model. Seed 132 was the best performing model via validation AUC, which is also the model that is used for analyses and interpretation in Section 5.

Table 5: MIMIC Sepsis Test Set Performance (AUROC and Log Likelihood) in hospital mortality prediction over 10 seeds, using other AR order and number of states. For KD-AR-HMM, we use the LSTM model with the best validation AUC as the LSTM Teacher model (which has a test AUC of 0.851).

Model	Val AUC	Test AUC	Val Log-Lik	Test Log-Lik	BIC
LSTM (Teacher)	0.857	0.851	N/A	N/A	N/A
LSTM (10-seeds)	0.822 ± 0.047	0.836 ±0.026	N/A	N/A	N/A
KD-AR(1)-HMM, K=5	**0.807 ± 0.004**	**0.792 ± 0.009**	**-268.805 ± 20.175**	**-256.290 ± 19.465**	**832.266 ± 40.351**
KD-AR(2)-HMM, K=5	0.808 ± 0.012	0.788 ± 0.007	-435.767 ± 53.621	-420.646 ± 54.795	1215.298 ± 107.242
KD-AR(1)-HMM, K=10	0.794 ± 0.007	0.787 ± 0.009	-532.760 ± 16.277	-522.661 ± 15.431	2145.924 ± 32.556

Table 6: Sepsis Hospital Mortality Prediction: Validation/Test AUC and Log-Likelihoods across different seeds of KD-AR-HMM trainings, with AR(1) and 5 states. Hyperparams are fixed and chosen after 40 seeds of random grid search.

Seed	Val AUC	Test AUC	Val Log-Lik	Test Log-Lik
123	0.8116	0.7842	-252.70	-240.93
124	0.8063	0.8015	-269.97	-258.07
125	0.8012	0.7916	-313.37	-297.91
126	0.8057	0.7920	-246.11	-234.23
127	0.8055	0.7821	-282.07	-270.50
128	0.8113	0.7818	-257.42	-242.34
129	0.8116	0.7871	-265.57	-253.72
130	0.8042	0.8024	-280.29	-266.99
131	0.8030	0.7908	-247.86	-236.24
132	0.8121	0.8053	-272.70	-261.98

Table 7: MIMIC Sepsis Cohort Other Clinical Outcomes: Performance (Log Likelihoods) in Clinical Outcome Predictions. Mean and standard deviation of test Log Likelihoods averaged over 10 seeds reported. MV = Mechanical Ventilation. Edema refers to pulmonary edema.

	Edema	Dialysis	MV	Diuretics
AR-HMM	-6187.512 ± 1440.118	997.811 ± 0.009	501.372 ± 2.938	207.064 ± 7.060
DISC-AR-HMM	-1099.987 ± 475.503	-1512.941 ± 793.635	-1315 ± 906.737	-872.471 ± 1048.093
KD-AR-HMM	-700.874 ± 36.047	-483.433 ± 76.907	-404.758 ± 56.308	-217.457 ± 14.272

Appendix C. Additional Results Sepsis Outcome Prediction

The significance of latent states are also qualitatively supported by observing the generated latent state distribution across all patients in the dataset. In Figure 4, we see that state 2 is much less likely to be found in patients who died, than for patients who lived. On the contrary, states 1 and 3 were more likely in patients who died than patients who lived.

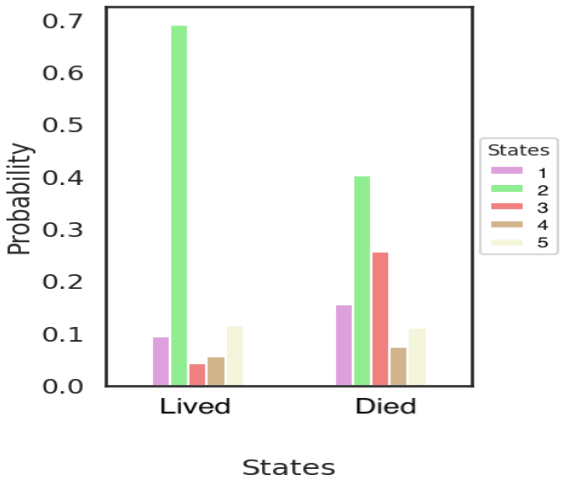


Figure 4: Sepsis Cohort Hospital Mortality Prediction Task: Latent state distribution over all patients in the dataset for the KD-AR-HMM model for hospital mortality prediction, using state marginals via the latent posterior given the final learned global model parameters.

Table 8: MIMIC Sepsis cohort Other Clinical Outcomes: AUPRCs in Outcome Predictions. Mean and standard deviation of test AUPRCs averaged over 10 seeds reported. MV = Mechanical Ventilation. Edema refers to pulmonary edema.

	Edema	**Dialysis**	**MV**	**Diuretics**
LSTM	0.868 ± 0.068	0.717 ± 0.093	0.972 ± 0.021	0.794 ± 0.010
AR-HMM	0.514 ± 0.008	0.150 ± 0.006	0.826 ± 0.000	0.473 ± 0.000
DISC-AR-HMM	0.568 ± 0.025	0.456 ± 0.186	0.821 ± 0.043	0.528 ± 0.025
KD-AR-HMM	0.588 ± 0.010	0.700 ± 0.017	0.856 ± 0.015	0.564 ± 0.016

C.1. Sepsis Other Clinical Outcomes

Pair-wise Similarity Matrices for Sepsis Cohort. The effectiveness of the knowledge distillation framework and the similarity-based constraint for the additional outcomes is further supported by the similarity matrices shown in Figure 5.

The distillation constraint encourages the student model's (KD-AR-HMM) pairwise similarity matrix to match that of the teacher LSTM model. Compared with the unsupervised baseline AR-HMM and the discriminative AR-HMM (DISC-AR-HMM), the constrained distillation approach generally yields pairwise similarity matrices that more closely align with those of the teacher model for predicting MV and Dialysis outcomes. We note that the test set similarity matrices shown are from the best-performing seed for each model (selected based on best validation performance). While the baseline DISC-AR-HMM can sometimes produce matrices resembling the teacher's (e.g. for MV and Dialysis outcomes), its performance varies greatly across random seeds. We recall that all AR-HMM–based models perform worse than the LSTM teacher in predicting Pulmonary Edema and Di-

uretics (in AUROC), and the KD-AR-HMM similarly struggles to replicate the teacher's similarity structure for these outcomes. Investigating strategies to improve performance in these challenging prediction tasks remains a direction for future work.

C.2. Respiratory Failure Dataset

Qualitatively, from the generated similarity matrices provided in Figure 6 for 28-day mortality outcome, the KD-AR-HMM finds it more challenging to learn the underlying latent structure from the teacher LSTM in comparison to the sepsis mortality prediction task. However, the KD-AR-HMM latent structure does have a stronger resemblance to the teacher LSTM than the baseline models, which could be evidence that the knowledge distillation framework is still effective even for more difficult prediction tasks.

Appendix D. Hyperparameter Tuning

Table 9 lists the hyperparameter search space for the Teacher LSTM model. The best hyperparameter setting after the grid search is bolded. Table 10 lists the hyperparameter search space for the AR-HMM models. Exhaustively searching all parameter combinations was computationally prohibitive; therefore, for each seed, we randomly sampled values from the search space.

Table 9: Hyperparameter search space in LSTM Teacher model.

	Hyperparameters	**Range**
LSTM-Teacher	Number of Layers	2, 3, 4
	Hidden Dim.	8, 16, 32, 64, 128
	Learning Rate	0.01, 0.001, 0.0001

Table 10: Hyperparameter Settings. This table shows the options for various hyperparameters that we tried for our AR-HMM models.

Hyperparameter	**Settings**
Discriminator coefficient	1e5, 1e7, 1e9, 1e11
Similarity coefficient	1e9, 1e11, 1e13, 1e15
Log-likelihood coefficient	1, 1e3, 1e6
determinants coeff	1, 1e3, 1e6, 1e9
global vars entropy coeff	1, 1e3, 1e6
local vars entropy coeff	1, 1e3, 1e6
Priors coefficient	1, 1e5, 1e10

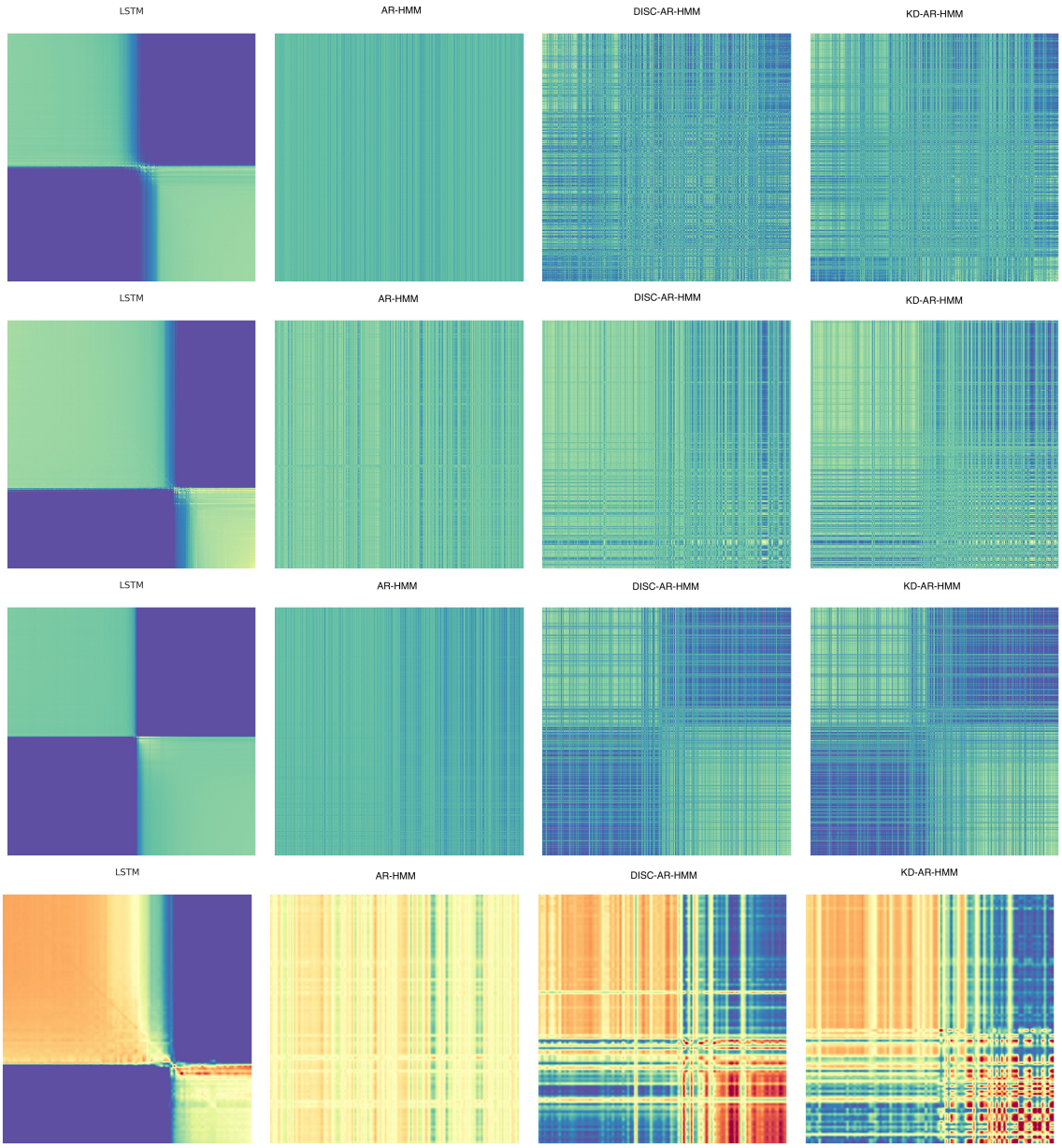


Figure 5: Sepsis Cohort Other Clinical Outcome Prediction: Comparing pair-wise similarity matrices representing the feature representations of the teacher LSTM, KD-AR-HMM and other baseline AR-HMM models. From left to right: LSTM, AR-HMM, DISC-AR-HMM, KD-AR-HMM. These are generated from the models trained to predict other non-mortality outcomes on the sepsis cohort: from top to bottom Pulmonary Edema, need for Diuretics, need for MV, and need for Dialysis.

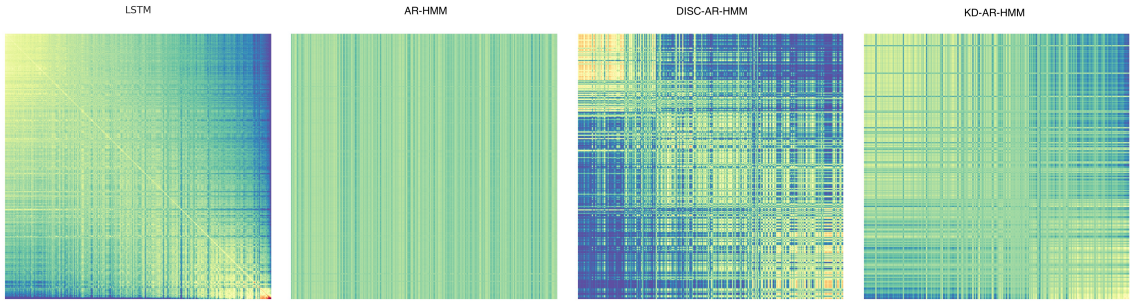


Figure 6: Respiratory Failure Cohort: Generated similarity matrices representing the feature representations of the teacher LSTM and student KD-AR-HMM models. From left to right: LSTM, vanilla AR-HMM, DISC-AR-HMM, KD-AR-HMM. These are generated from the models trained on the MV-wean cohort to predict 28-day mortality.

Appendix E. Data Covariates

Table 11 shows the variables that were used as inputs for both the teacher LSTM and student AR-HMM models. The vasopressor amount variable measures the total amount of vasopressors used during that time period. Vasopressors are standardized by comparing their relative strength to norepinephrine, also known as noradrenaline or norad. The vasopressors included in this standardization are norepinephrine (or levophed), epinephrine, vasopressin, phenylephrine, and dopamine. The measurement unit used is mcg/kg/minute, except for vasopressin, which is expressed as units/minute. The standardization process involves adjusting the dosage rate of each vasopressor by multiplying it with a scaling constant based on the typical dosing of each drug. Norepinephrine is typically administered at a dosage range of 0-1 mcg/kg/minute. If multiple vasopressors were used during the same time period, the combined total dose for each hour is reported.

E.1. Respiratory Failure Cohort Mechanical Ventilation Dataset

Covariates in the respiratory failure MV dataset, include heart rate, systolic blood pressure (SBP), mean BP (MBP), diastolic BP (DBP), respiratory rate, SpO2, pH, baseexcess, total CO2, temperature, lactate, GCS, aniongap, bicarbonate, creatinine, hematocrit, hemoglobin, BUN, WBC, PEEP, tidal volume (set and observed), minutes volume, peak inspiratory pressure, plateau pressure, mean airway pressure, inspiratory time, FiO2, vasopressor amount, PaO2, pCO2, AaDO2 (alveolar-arterial oxygen difference), P/F ratio, driving pressure, SOFA scores, patient weight, patient height, elixhauser comorbidity score, and patient age.

Table 11: MIMIC time-varying variables that were used as inputs to both the teacher LSTM and student AR-HMM models.

Variable Name	**Variable Type**	**Units**
Heart Rate	Continuous	beats/min
Diastolic Blood Pressure	Continuous	mmHg
Systolic Blood Pressure	Continuous	mmHg
Mean Blood Pressure	Continuous	mmHg
Minimum Diastolic Blood Pressure	Continuous	mmHg
Minimum Systolic Blood Pressure	Continuous	mmHg
Minimum Mean Blood Pressure	Continuous	mmHg
Temperature	Continuous	°C
SOFA Score	Treated as Continuous	N/A
Glasgow Coma Score	Treated as Continuous	N/A
Platelet	Continuous	counts/10^9L
Hemoglobin	Continuous	g/dL
Calcium	Continuous	mg/dL
BUN	Continuous	mmol/L
Creatinine	Continuous	mg/dL
Bicarbonate	Continuous	mmol/L
Lactate	Continuous	mmol/L
Potassium	Continuous	mmol/L
Bilirubin	Continuous	mg/dL
Glucose	Continuous	mg/dL
pO2	Continuous	mmHg
SO2	Continuous	%
SpO2	Continuous	%
pCO2	Continuous	mmHg
Total CO2	Continuous	mEq/L
pH	Continuous	Numerical[1,14]
Base excess	Continuous	mmol/L
Weight	Continuous	kgs
Respiratory Rate	Continuous	breaths/min
Total Fluids	Continuous	mL
Urine Output	Continuous	mL
Total Urine Output	Continuous	mL
Fluid Bolus	Continuous	mL
Vasopressor Amount	Continuous	mcg/kg/min

Table 12: MIMIC time-varying variables that were only used for the teacher LSTM model, and not for the student AR-HMM models.

Variable Name	**Type**	**Units**
Minimum Mean Blood Pressure from Baseline	Continuous	mmHg
Glasgow Coma Score - Motor	Ordinal	N/A
Glasgow Coma Score - Verbal	Ordinal	N/A
Glasgow Coma Score - Eye	Ordinal	N/A
O2 requirement level	Ordinal [0,6]	N/A
Pulmonary Edema Indicator	Binary	N/A
Cumulative Edema	Binary	N/A
Diuretics Indicator	Binary	N/A
Diuretics Amount	Continuous	mg
Dialysis Indicator	Binary	N/A
Mechanical Ventilation Indicator	Binary	N/A
Bolus Indicator	Binary	N/A
Vasopressor Indicator	Binary	N/A