AttDMM: An Attentive Deep Markov Model for Risk Scoring in Intensive Care Units

Yilmazcan Özyurt
ETH Zürich
8092 Zürich, Switzerland
yozyurt@ethz.ch

Tobias Hatt
ETH Zürich
8092 Zürich, Switzerland
thatt@ethz.ch

Mathias Kraus
FAU Erlangen-Nuremberg
90403 Nuremberg, Germany
mathias.kraus@fau.de

Stefan Feuerriegel
ETH Zürich
8092 Zürich, Switzerland
sfeuerriegel@ethz.ch

ABSTRACT
Clinical practice in intensive care units (ICUs) requires early warnings when a patient’s condition is about to deteriorate so that preventive measures can be undertaken. To this end, prediction algorithms have been developed that estimate the risk of mortality in ICUs. In this work, we propose a novel generative deep probabilistic model for real-time risk scoring in ICUs. Specifically, we develop an attentive deep Markov model called AttDMM. To the best of our knowledge, AttDMM is the first ICU prediction model that jointly learns both long-term disease dynamics (via attention) and different disease states in health trajectory (via a latent variable model). Our evaluations were based on an established baseline dataset (MIMIC-III) with 53,423 ICU stays. The results confirm that compared to state-of-the-art baselines, our AttDMM was superior: AttDMM achieved an area under the receiver operating characteristic curve (AUROC) of 0.876, which yielded an improvement over the state-of-the-art method by 2.2%. In addition, the risk score from the AttDMM provided warnings several hours earlier. Thereby, our model shows a path towards identifying patients at risk so that health practitioners can intervene early and save patient lives.

CCS CONCEPTS
• Computing methodologies → Neural networks; Supervised learning by classification; • Mathematics of computing → Time series analysis; Markov processes; • Applied computing → Health care information systems.

KEYWORDS
disease progression, intensive care unit, mortality prediction, attention network, deep Markov model

1 INTRODUCTION
Intensive care units (ICUs) provide healthcare to patients with severe or life-threatening illnesses. ICUs receive patients directly from an emergency unit, from other wards if their condition deteriorates rapidly, or after surgery. In ICUs, patients require constant care to ensure normal body functions. Yet, due to the severity of underlying illnesses, their health trajectories cannot always be stabilized. Owing to this, mortality rates in ICUs are among the highest across all hospital units [34] and are estimated to be 8–19% [36].

To ensure normal body functions, patients in ICUs are subject to extensive monitoring [21]. Examples are the monitoring of body temperature, heart rate, and blood pressure. Based on this, clinical professionals determine a risk score that identifies the probability of in-hospital mortality. The risk score is crucial for decision-making in ICU practice as it guides the treatment plans [5, 26, 37]. In addition, it provides early warnings when a health condition is about to deteriorate so that preventive measures can be taken.

The clinical literature indicates the development of several methods for risk scoring in ICUs that are nowadays widely used in clinical practice. Among the most widely applied ones is the simplified acute physiology score (SAPS) [32, 35], which assesses the severity of the health condition as defined by the probability of patient in-hospital mortality. For this, the risk score makes use of measurements that indicate vital signs, such as body temperature, heart rate, and blood pressure. However, the aforementioned risk scores are computed through overly simple decision rules that operate only on a few measurements from selected timestamps. In other words, the complete time-series of measurements is ignored, and because of this, the prediction power concerning patient mortality is limited.

Recent works have addressed mortality prediction in ICUs through the use of machine learning. On the one hand, neural networks have been used so that long-term temporal dependencies are captured. Examples of neural networks that have been adapted for ICU predictions are long short-term memory (LSTM) [15, 44, 52] and gated recurrent unit (GRU) [6, 10]. These models have been powerful in representing complex interactions among (high-dimensional) measurements and thus represent the state of the art. On the other hand, latent variable models have been used for ICU prediction [16, 17]. In this case, the latent variables allow latent disease states in the health trajectory to be captured. However, we are not aware of any previous works that have combined the strengths of neural networks and latent variables into a joint model for predicting ICU mortality.

Proposed model: In this work, we propose a novel generative deep probabilistic model for predicting mortality risk in ICUs. Specifically, we develop an attentive deep Markov model called AttDMM. Based on this, our model allows us to jointly capture (1) long-term disease dynamics (via attention) and (2) different disease states in the health trajectory (via a latent variable model). To the best of our knowledge, our model is the first combination

The code is available from https://github.com/oezyurty/AttDMM
of a deep Markov model with an attention mechanism. In addition, AttrDMM is closely aligned with the needs in clinical practice providing confidence interval of the real-time risk score that further facilitates decision-making. Finally, we show how to estimate AttrDMM via an end-to-end training task by a tailored evidence lower bound (ELBO).

**Results:** Our AttrDMM was evaluated on an established baseline dataset from clinical practice, MIMIC-III [21], comprising 53,423 ICU stays. For each ICU stay, we evaluated the performance across two prediction tasks. First, we predicted the mortality risk from measurements spanning the first 48 hours after ICU admission. This prediction task is analogous to that of risk scoring from the clinical literature [e.g., 20, 32, 35, 53] and thus facilitates comparability with prior literature. In this prediction task, state-of-the-art baselines were consistently outperformed. Compared to the baselines, the proposed AttrDMM achieved a performance improvement in the area under the receiver operating characteristic curve (AUROC) by 2.2% and in area under the precision-recall curve (AUPRC) by 2.4%. On top of that, our model achieved the same AUROC (AUPRC) performance as the best baseline 12 hours (6 hours) earlier. Second, we assessed a prediction task in which we make long-term forecasts of mortality risk. This prediction task is demanded by clinical practice as many conditions remain stable for a fairly long time window but afterwards deteriorate suddenly. We found that AttrDMM outperforms state-of-the-art baselines in terms of AUROC by 2.2% and AUPRC, remarkably, by 5.4%.

**Contributions:** Our work advances machine learning for ICU predictions in the following ways:

1. We present a novel generative deep probabilistic model called AttrDMM. For this, we combine a deep Markov model with an attention mechanism for risk scoring. To the best of our knowledge, AttrDMM is the first deep Markov model that is tailored to mortality predictions in ICUs.

2. Our evaluation demonstrated that AttrDMM achieved state-of-the-art performance. Specifically, our model outperformed existing baselines by a considerable margin. Thereby, our work enables precise warnings in critical care so that patient lives can be saved.

3. Our AttrDMM adheres to the needs of clinical practice in critical care by informing about the precision of the prediction. It provides a confidence interval of the real-time mortality risk score. Based on it, practitioners can carefully assess their decision-making with regard to the confidence intervals. Thereby, we overcome a shortcoming of existing works in machine learning for ICU predictions, which return only the point estimate of ICU mortality.

## 2 RELATED WORK

### 2.1 Machine learning in healthcare

Machine learning in healthcare has different objectives. Given the breadth of research, we summarize key streams in the following (see [3] for a detailed literature review). Examples are predicting hospital readmission risk [e.g., 13, 19, 45, 47], forecasting the course of symptoms [e.g., 4, 12, 29, 42], and predicting diagnosis codes of future diseases [e.g., 7–9, 27, 28, 33, 51]. Depending on the objective of the prediction, different sources of clinical data are used. Typical examples comprise electronic health records, sensor measurements (e.g., in ICUs), and patient sociodemographics (often termed “risk factors” in a clinical context). The underlying prediction models also differ, for instance, in whether they handle static and/or time-series data, warrant interpretability, and model prediction intervals (i.e., whether they output confidence intervals to assess the uncertainty around point estimates) [e.g., 8, 28, 51].

The use of machine learning in healthcare allows clinical practitioners to obtain predictions about the current (and future) health condition of a patient. Based on this, clinical practitioners can adapt their treatment plans accordingly (e.g., by planning preventive interventions or choosing different treatments).

### 2.2 Machine learning in intensive care units

**Risk prediction in ICUs:** Machine learning in ICU settings has different objectives, such as predicting adverse events like sepsis [e.g., 22, 23, 40, 41, 48], while a predominant focus in the literature is on predicting mortality risk. For this, one uses various measurements of vital signs, such as body temperature, heart rate, and blood pressure. They are widely regarded as important indicators that describe a patient’s health status in critical care [14, 25].

In practice, there is variability in which measurements are recorded for a specific patient depending on her condition and reason of hospitalization. For instance, for some diseases, sodium and potassium levels are highly indicative of the future course, while for COVID-19, a focus might be placed on measuring respiratory-related thoracic movements. As such, some measurements might not be subject to recording (due to the setting) and, hence, are missing for the complete (or a partial) ICU stay. Hence, the missingness of ICU measurements must be appropriately modeled in the context of ICU prediction [e.g., 6].

**Clinical practice in ICU risk scoring:** In clinical practice, predictions of mortality risk are computed through simple decision rules. A key benefit of decision rules from clinical practice is that they directly output a dynamic risk score. Common examples are the so-called simplified acute physiology score (SAPS) [32, 35], the acute physiology and chronic health evaluation (APACHE) [53], and mortality probability model (MPM) [20]. To allow for straightforward use, the decision rules rely only upon a few sensor measurements and, in particular, additional characteristics of patients (i.e., risk factors) are not considered. In essence, these decision rules determine the mortality risk through a linear combination of different vital signs. However, only the most critical value from the last 48 hours is considered, rather than the complete time-series of vital signs. Hence, the rich information embedded in high-resolution measurements is—to a large extent—ignored which limits the prediction power.

**Machine learning for ICU mortality risk scoring:** Recent works have approached ICU mortality prediction using machine learning. For this, several sequential neural networks have been adapted specifically to ICU settings, namely long short-term memory networks [15, 18] and gated recurrent units [6, 10, 43]. For ICU predictions, state-of-the-art sequential networks are designed to capture long-term dependencies in high-resolution time-series.
However, existing machine learning does not explicitly model different disease states in trajectory, which would require a latent variable approach.

2.3 Latent variable models

Latent variable models have the ability to account for the fact that patient health trajectories undergo different disease states (e.g., acute, stable). However, latent variable models explicitly consider the fact that such disease states cannot be directly observed and must thus be treated as latent [e.g., 4, 12, 42]. In principle, such models assume that the relation between health measurements and the disease states are stochastic. A naive example is a hidden Markov model. However, these models are generally made for modeling a sequence of observations instead of classification tasks.

To adapt latent variable models for prediction tasks, some works [1, 2, 49] encode the risk into the latent variable and adopt sequential hypothesis testing [46]. Others [16, 17] feed latent variables into another prediction algorithm, while even others [30, 31, 50] match discrete latent variables to the risk profiles. However, these approaches have only a limited ability in capturing long-term temporal dynamics, rely on discrete latent variables, and assume linearity in the transition and emission components.

The deep Markov model (DMM) [24] overcomes the above limitations by introducing continuous latent variables, non-linear transition networks, and non-linear emission networks. However, the DMM presents an unsupervised framework, because of which an application of the DMM to the prediction of mortality risk is precluded. Instead, a new model is needed, and for this reason, we develop an attentive deep Markov model to address the above mentioned prediction tasks.

Research gap: To the best of our knowledge, no prior work has combined the strength of neural networks and latent variable approaches for the purpose of ICU prediction. To fill this gap, we propose a novel attentive deep Markov model called AttDMM.

3 ATTDM: ATTENTIVE DEEP MARKOV MODEL FOR ICU RISK SCORING

3.1 Problem statement

In the following, we develop AttDMM, which processes ICU stays in order to predict the ICU mortality label \( y \in \{0, 1\} \). A label \( y = 0 \) denotes a discharge from ICU, whereas \( y = 1 \) denotes that the patient has died during the ICU stay.

The input to our model are ICU stays. Each stay is represented by static and time-series features: (1) Static features are denoted by \( s \). These encode potential risk factors at the patient level (e.g., gender, age) and thus describe the between-patient heterogeneity. (2) Time-series features are denoted by \( \{x_{ti}\}_{t=1}^{T} \). These encode a high-resolution time-series with various measurements for each time step \( t = 1, \ldots, T \) throughout the ICU stay (e.g., in MIMIC-III, they are sampled at a 2-hour resolution). Some measurements might not be subject to recording, and, therefore, are missing for the complete (or a partial) ICU stay. To address this issue, we first decompose \( x_{ti} \) into \( \{x_{ti}\}_{i=1}^{M} \), where \( M \) is the number of features (i.e., types of measurements). Then, for each \( x_{ti} \), we derive the mask variable \( m_{ti} \) that denotes if \( x_{ti} \) will be imputed (\( m_{ti} = 0 \)) or not (\( m_{ti} = 1 \)).

3.2 Specification of AttDMM

AttDMM is a generative deep probabilistic model for predicting mortality risk in ICUs. The model takes lab measurements recorded during an ICU stay as input. The measurements are processed to capture the latent disease state in a patient trajectory. This is achieved via a latent variable model which captures the latent disease state at time \( t \) by a latent variable \( z_t \). Based on the series of latent variables, our model infers the ICU mortality risk of a patient. This is achieved via an attention mechanism.

AttDMM has 4 model components, which are depicted in Figure 1. A (1) transition network specifies the transition probability among consecutive latent variables (i.e., \( z_{t-1} \) and \( z_t \)). It further accommodates between-patient heterogeneity. An (2) emission network models the probability of a lab measurement given the latent variable from a certain time step. In this case, it adheres to the assumption that the lab measurements are stochastically linked to the latent variables. The combination of a transition network and emission network represents the generative part of the model, which is used as an ELBO regularization in the loss function. An (3) attention network outputs the summary representation of the patient trajectory \( \hat{z} \), based on all latent variables \( z_{1:T} \). Finally, a (4) predictor network takes this summary representation as input and predicts the ICU mortality risk of a patient.

In terms of notation, let ReLU denote the rectified linear unit, \( \odot \) denote the element-wise multiplication, and \([\cdot : \cdot]\) denote the concatenation of two vectors.

![Figure 1: AttDMM. Black squares denote neural networks.](image_url)

3.3 Model components

(1) Transition network: This component specifies the transition probability among consecutive latent variables. For this, we build upon the Markov property; i.e., the current latent variable only depends on the previous one. Further, it accommodates the between-patient heterogeneity by making use of the patient’s static variable \( s \).

Formally, the transition network models the distribution of a latent variable \( z_t \) conditional on the inputs \( z_{t-1} \) and \( s \), denoted by \( p(z_t | z_{t-1}, s) \). This distribution is parametrized by a multivariate Gaussian distribution with a diagonal covariance, denoted by
\( N(\mu z_t, \Sigma z_t) \), where \( \Sigma z_t \) is the covariance matrix with \( \sigma_{z_t}^2 \) on the diagonal and 0 otherwise. The mean \( \mu z_t \) is a sum of two terms: (1) a linear contribution of the inputs, denoted by \( \bar{\sigma} z_t \) and (2) a non-linear contribution of the inputs, denoted by \( \tilde{\sigma} z_t \). The weights of these two terms are controlled by a gate \( g z_t \). The diagonal \( \sigma_{z_t}^2 \) is derived from \( \tilde{\sigma} z_t \) (plus a constant\(^2\)). The mathematical formulation of the transition network is as follows:

\[
\begin{align*}
g'_z &= \text{ReLU}(W'_g z_{t-1} \cdot s + b'_g), \\
g_z &= \text{sigmoid}(W'_g g'_z + b'_g), \\
\tilde{\mu}_z &= \text{ReLU}(W'_p [z_{t-1}; s] + b'_p), \\
\hat{\mu}_z &= W'_\mu \tilde{\mu}_z + b'_\mu, \\
\mu_z &= g_z \odot \hat{\mu}_z + (1 - g_z) \odot \tilde{\mu}_z, \\
\sigma_z &= \text{softplus}(W'_\sigma \text{ReLU}(\tilde{\mu}_z) + b'_\sigma) + \text{const},
\end{align*}
\]

with matrices \( W'_g, W'_p, W'_\mu, W'_\sigma \), and \( b'_g, b'_p, b'_\mu, b'_\sigma \).

(2) Emission network: The emission network outputs the probability of a lab measurement given the latent variable. Formally, the emission network models the distribution of \( x_t \) given the latent variable \( z_t \), denoted by \( p(x_t \mid z_t) \). In the distribution, each feature \( x_{ti} \) of \( x_t \) is modeled as conditionally independent from each other, i.e.,

\[
p(x_t \mid z_t) = \prod_{i, m_i = 1} p(x_{ti} \mid z_t),
\]

where \( m_i \) is leveraged to mask out unobserved entries. The distribution \( p(x_t \mid z_t) \) is parameterized by a univariate Gaussian distribution, denoted by \( N(\mu x_{ti}, \sigma_{x_{ti}}^2) \), where \( \mu x_{ti} \) and \( \sigma_{x_{ti}}^2 \) are the \( i \)-th indices of the vectors \( \mu z_t \) and \( \sigma_{z_t}^2 \). With the same model assumptions as the transition network, the mean \( \mu z_t \) is a sum of two terms: (1) a linear contribution of the inputs, denoted by \( \bar{\mu} z_t \), and (2) a non-linear contribution of the inputs, denoted by \( \tilde{\mu} z_t \). The weights of these two terms are controlled by a gate \( g x_t \). The variance \( \sigma_{x_{ti}}^2 \) is derived from \( \tilde{\mu} z_t \) (plus a constant as above). The mathematical formulation of the emission network is as follows:

\[
\begin{align*}
g'_x &= \text{ReLU}(W'_g \tilde{\mu} z_t + b'_g), \\
g_x &= \text{sigmoid}(W'_g g'_x + b'_g), \\
\tilde{\mu}_x &= \text{ReLU}(W'_p \tilde{\mu} z_t + b'_p), \\
\hat{\mu}_x &= W'_\mu \tilde{\mu}_x + b'_\mu, \\
\mu_x &= g_x \odot \hat{\mu}_x + (1 - g_x) \odot \tilde{\mu}_x, \\
\sigma_x &= \text{softplus}(W'_\sigma \text{ReLU}(\tilde{\mu}_x) + b'_\sigma) + \text{const},
\end{align*}
\]

where \( x_{ti} \sim p(x_{ti} \mid z_t) = N(\mu x_{ti}, \sigma_{x_{ti}}^2), \forall i \in \{1, ..., M\} \)

(3) Attention network: The attention network allows the model to assign a different importance to each latent variable when making inferences of mortality risk. The importance is computed via similarity between a query vector and the latent variable. Thereby, the attention network aggregates the latent variables based on their importance. The resulting vector is used as a summary representation of the patient health trajectory, which is fed into a predictor network. Although this overcomes the Markov property, it allows the model to capture long-term dependencies.

Formally, the attention network outputs the aggregation of latent variables \( \hat{z} \) based on the input sequence \( z_{1:T} \). This is formalized via

\[
\hat{z}_t = W z_t + b z_t, \\
y_t = \frac{\exp(\xi K[v, \hat{z}_t])}{\sum_{i=1}^T \exp(\xi K[v, \hat{z}_i])},
\]

with matrix \( W\), bias vector \( b_z \), query vector \( v \), and scalar \( \xi \), where \( K(\cdot, \cdot) \) defines the cosine similarity, i.e.,

\[
K(v, \hat{z}_t) = \frac{v \cdot \hat{z}_t}{||v|| \cdot ||\hat{z}_t||}
\]

(4) Predictor network: The predictor network is the final step of making the inference of mortality risk. For this, it predicts \( \hat{y} \) based on \( \hat{z} \) from the attention network. This yields \( \hat{y} = \text{sigmoid}(U_y \text{ReLU}(W y_{\hat{z}} + b y) + c_y) \)

\( (22) \)

with matrices \( W_y, U_y \), and bias vectors \( b_y \) and \( c_y \). \( \hat{y} \) is the output of \( \text{AttDMM} \).

3.4 Posterior approximation

The computational complexity present in \( \text{AttDMM} \) hinders the exact inference of the latent variables. Because of this, we develop a posterior approximation that builds upon stochastic variational inference to estimate latent variables (see Figure 2).

For \( \text{AttDMM} \), we approximate the posterior distribution of latent variables (denoted as \( q \)) based on the sequence of lab measurements and static variables. It further makes use of the missingness of each lab measurement (i.e., masks).

Figure 2: Posterior approximation of \( \text{AttDMM} \), rolled-out across time steps. Black squares denote neural networks.

Formally, the approximated posterior distribution \( q \) of \( z_T \) is based on the inputs \( z_{1:T}, s, x_{1:T}, \) and \( m_{q_{t,T}} \). For this, we proceed as follows. First, we use a recurrent neural network (RNN) for encoding
with matrices $W_h$, $U_h$ and bias vector $b_h$. Second, a combiner network is applied to the inputs of $z_{t-1}$, $s$, and $h_t$. In this case, $z_{t-1}$ and $h_t$ encapsulate the past and future lab measurements. In addition, $s$ is leveraged to accommodate the between-patient heterogeneity in the posterior approximation.

The combiner network models the approximated distribution of $z_t$, denoted by $q(z_t | z_{t-1}, s, x_{t:T}, m_{1:T})$. This distribution is parameterized by a multivariate Gaussian distribution with a diagonal covariance, denoted by $N(\mu_{z_t}, \Sigma_{z_t})$, where $\Sigma_{z_t}$ is the covariance matrix with $\sigma^2_{z_t}$ on the diagonal and 0 otherwise. The mathematical formulation of the combiner network is given by

\[
\begin{align*}
    c_t &= W_c [z_{t-1}; s] + b_c, \\
    \hat{h}_t &= 0.5 \tanh (c_t + h_t), \\
    \mu_{z_t} &= W_{\mu} \hat{h}_t + b_{\mu}, \\
    \sigma_{z_t} &= \text{softplus}(W_{\sigma} \hat{h}_t + b_{\sigma}) + \text{const.}, \\
    z_t &\sim q(z_t | z_{t-1}, s, x_{t:T}, m_{1:T}) = N(\mu_{z_t}, \Sigma_{z_t})
\end{align*}
\]

with matrices $W_c$, $W_{\mu}$, and $W_{\sigma}$ and bias vectors $b_c$, $b_{\mu}$, and $b_{\sigma}$. Altogether, this approximated the posterior of the latent variables of AttrDMM, which is leveraged to optimize the ELBO via the stochastic variational inference. See Appendix A for the details of the estimation procedure.

### 3.5 Inferring risk scores

AttrDMM estimates the probability of in-hospital mortality at time $T$ as follows. The static features $s$, the lab measurements $x_{1:T}$, and the masking of measurements $m_{1:T}$ are fed into the posterior approximation network. Based on it, AttrDMM constructs the posterior distribution $q(z_t | z_{t-1}, s, x_{t:T}, m_{1:T})$, from which the latent variables $z_{1:T}$ are sampled sequentially. This sampling procedure is repeated $N$ times, where the number $N$ is defined by the user. At the end of the sampling procedure, the set of sampled latent variables $\{z^t_G\}^{N}_{g=1}$ is processed sequentially by the attention network and the predictor network. AttrDMM then produces $N$ samples of in-hospital mortality prediction $\{\hat{y}^g\}^{N}_{g=1}$, whose mean is used for the estimate of the probability of in-hospital mortality.

The complete hospital stay of an ICU patient is tracked by our AttrDMM, thereby producing the probability of in-hospital mortality over time. In this way, AttrDMM can inform the physicians in the ICU about the risk scores of each patient.

AttrDMM yields a full posterior distribution of the prediction (rather than a point estimate). This distribution is acquired by the multiple samples of the mortality prediction $\{\hat{y}^g\}^{N}_{g=1}$. This allows us to compute confidence intervals of the prediction in order to quantify the uncertainty of our prediction. And, in practice, one can use it in order to decide whether the prediction is sufficiently reliable.

### 4 EXPERIMENTAL SETUP

#### 4.1 Dataset

Our evaluation was based on an established reference dataset with ICU measurements, namely, MIMIC-III [21]. MIMIC-III is one of the largest publicly available ICU datasets, comprising 38,597 distinct patients and a total of 53,423 ICU stays. Because of this, MIMIC-III has been used extensively in prior literature for benchmarking [e.g., 6, 10, 15, 18, 22, 23, 38, 40, 41]. For each ICU stay, the dataset includes a time-series of different measurements, such as temperature, heart rate, and blood pressure. In addition, the dataset also reports sociodemographic variables that describe the heterogeneity among patients, such as age and admission type (e.g., scheduled surgery, unscheduled surgery). The available features for prediction are reported in Appendix B. For details on the dataset, we refer the reader to [21].

We follow the preprocessing pipeline established by prior literature [6, 38]. We first remove ICU stays which are shorter than 2 days or longer than 30 days, and ICU stays of patients younger than 15. Further, we only consider the first ICU stay of patients with multiple stays. Afterwards, we applied further preprocessing steps consistent with the aforementioned references [6, 38]. For each ICU stay, we extracted features that were sampled regularly (i.e., every 2 hours), and we filled missing values with a forward-backward linear imputation.

The preprocessed dataset contains 31,895 ICU stays. Out of them, the share of ICU stays with in-hospital mortality amounted to 3,311 stays (i.e., 10.38%). The remaining 28,584 stays (i.e., 89.62%) resulted in a discharge from the ICU. The length of hospital stay after ICU admission had a mean of 189.21 hours (with a standard deviation of 136.44 hours). Of all the ICU stays, 58.78% required a hospital stay of longer than five days. The distribution of hospitalization length is shown in the Appendix.

#### 4.2 Prediction tasks

##### 4.2.1 Task 1: Mortality prediction (48 hours after ICU admission)

In this task, we predict the in-hospital mortality based on the measurement data from the first 48 hours after ICU admission. This represents the reference task in the literature for calibrating and evaluating mortality predictions in ICUs. To examine how the prediction performance changes throughout an ICU stay, we further report the prediction performance at different time periods after the ICU admission, ranging from 12 to 48 hours after the ICU admission. For this, all the predictions have only access to the respective (limited) time frame of ICU measurements. This allows us to assess how precise predictions are at the beginning of the ICU stay.

##### 4.2.2 Task 2: Mortality prediction (complete hospital stay after ICU admission)

A second prediction task is used to evaluate ICU mortality during the complete hospital stay. Typical use cases are diseases in which the health trajectory remains stable for a large amount of time and, only later, deteriorates suddenly. The relevance of the second prediction task is also seen in the summary statistics of our dataset, in which hospital stays with a duration of more than 48 hours are highly common and account for more than 80% of all ICU admissions.

The complete hospital stay of an ICU patient is tracked by our AttrDMM, thereby producing the probability of in-hospital mortality over time. In this way, AttrDMM can inform the physicians in the ICU about the risk scores of each patient.

AttrDMM yields a full posterior distribution of the prediction (rather than a point estimate). This distribution is acquired by the multiple samples of the mortality prediction $\{\hat{y}^g\}^{N}_{g=1}$. This allows us to compute confidence intervals of the prediction in order to quantify the uncertainty of our prediction. And, in practice, one can use it in order to decide whether the prediction is sufficiently reliable.
Formally, we now use the complete time-series of the ICU data during both training and evaluation. We then report the prediction performance in intervals of 2 hours. In this case, we label time with respect to hours to discharge/death. We report the prediction performance for all lead times between ~120 hours and 0 hours. Longer lead times were discarded, as there was not sufficient data.

We later also report an overall performance score. For this, we aggregate the prediction performance across all time-steps via a weighted micro-level average. Thereby, we weigh each time step by the number of available patients for the evaluation at that time step.

4.3 Performance metrics

We compare our AttDMM with other baselines based on two performance metrics. (1) We report area under the receiver operating characteristic curve (AUROC). This was also chosen in prior literature [6, 10, 15]. We tested whether the improvement in AUROC is statistically significant via DeLong’s test [11]. (2) We further report area under the precision-recall curve (AUPRC). The AUPRC is frequently used in the clinical literature as it focuses on the performance relative to detecting a negative event (i.e., mortality).

4.4 Baselines

We compare our AttDMM against an extensive series of baselines that have been carefully crafted for mortality prediction in ICU settings. We first implemented SAPS-II representing current practice [25]. Formally, SAPS-II provides predefined decision rules that make use of the most critical lab measurements of the last 48 hours.

In addition, we implemented naïve baselines for performance comparisons. These are a multilayer perceptron (MLP) and a random forest (RF). Both are fed with summary statistics (mean, max, min, std. dev.) of the same features used by current practice. Hence, the naïve baselines are intended to reflect the power of the prediction heuristics from clinical practice.

We further adapt latent variable models to ICU mortality prediction. A hidden Markov model (HMM) is utilized with two hidden states, referring to discharge and mortality. Additionally, we use an HMM with multiple hidden states and feed these states into a long short-term memory, denoted as HMM+LSTM. Finally, we crafted a DMM for the prediction task, in which the last latent variable is fed into the predictor network (the same network of the AttDMM).

We use the following ICU-specific state-of-the-art models: a long short-term memory network for ICU prediction (ICU/LSTM) [15]; a gated recurrent unit for ICU prediction (ICU/GRU) [10]; and a gated recurrent unit with decay mechanism for ICU prediction (ICU/GRU-D) [6]. These models can, by nature of the underlying model, handle sequential input of variable-length and are, thus, trained with the same data as AttDMM. We use implementations analogous to those in the prior literature that have been tailored to handle the specific time-series structure in ICU settings (e.g., with regard to missing values, sampling frequency). Appendix C describes details about the implementation and evaluation of all baseline models as well as AttDMM.

We further provide a sensitivity analysis in which we study how the prediction performance varies after ICU admission. For this, Figure 3 depicts the evolution of both the AUROC and AUPRC from 12 to 48 hours (in steps of 2 hours) after ICU admission. All the baselines are outperformed by the proposed AttDMM. The improvement of AttDMM over the baselines is particularly pronounced for time periods after the first 24 hours. We further compare the models in their minimum time needed after the ICU admission in order to achieve a specific prediction performance. To achieve an AUROC of 0.85, the ICU/GRU-D needs 44 hours of patient data, whereas the same AUROC is achieved by AttDMM 12 hours earlier. For AUPRC, the /GRU-D performs better shortly after ICU admission, but after 18 hours, it is outperformed by AttDMM. An AUPRC of 0.45 is achieved by the ICU/GRU-D 46 hours after ICU admission, whereas the same AUROC is achieved by AttDMM after 40 hours, i.e., 6 hours earlier.

5 RESULTS

5.1 Task 1: Mortality prediction (48 hours after ICU admission)

Table 1 lists the results of in-hospital mortality prediction based on the measurement data from the first 48 hours after ICU admission. The best baseline is given by the ICU/GRU-D with an AUROC of 0.857. In contrast, AttDMM achieves an AUROC of 0.876, which is an improvement over the best baseline by 2.2%. The improvement is statistically significant (p<0.001). Similarly, AttDMM achieved the highest AUPRC with 0.465. Compared to the best baseline (ICU/GRU-D with an AUPRC of 0.454), this is an improvement of 2.4%.

| Model               | AUROC  | AUPRC  |
|---------------------|--------|--------|
| SAPS-II [25]        | 0.747 ± 0.007 | 0.255 ± 0.014 |
| MLP                 | 0.832 ± 0.014 | 0.395 ± 0.027 |
| Random forest       | 0.826 ± 0.010 | 0.416 ± 0.020 |
| HMM                 | 0.711 ± 0.007 | 0.173 ± 0.003 |
| HMM+LSTM            | 0.774 ± 0.022 | 0.289 ± 0.029 |
| DMM                 | 0.839 ± 0.013 | 0.387 ± 0.036 |
| ICU/LSTM [15]       | 0.831 ± 0.017 | 0.403 ± 0.028 |
| ICU/GRU [10]        | 0.838 ± 0.019 | 0.412 ± 0.029 |
| ICU/GRU-D [6]       | 0.857 ± 0.012 | 0.454 ± 0.014 |
| **Proposed AttDMM** | **0.876 ± 0.004** | **0.465 ± 0.010** |

Higher is better. Best value in bold.

Table 1: Performance in prediction task 1 where in-hospital mortality risk is estimated at 48 hours after ICU admission.

5.2 Task 2: Mortality prediction (complete hospital stay after ICU admission)

Table 2 lists the results of in-hospital mortality prediction based on the complete hospital stay. The best baseline is given by the random forest with an AUROC of 0.846. In contrast, AttDMM achieves an AUROC of 0.865. Therefore, AttDMM yields an improvement over the best baseline by 2.2%. The improvement is statistically significant (p<0.001). For the AUPRC, our proposed model again achieves the best score amounting to 0.545. In this case, AttDMM
yields an improvements over the best baseline (ICU/GRU-D with an AUPRC of 0.517) by 5.4%.

Table 2: Performance in prediction task 2 in which in-hospital mortality risk is estimated throughout the complete hospital stay.

| Model                  | AUROC       | AUPRC       |
|------------------------|-------------|-------------|
| SAPS-II [25]           | 0.829 ± 0.008 | 0.449 ± 0.014 |
| MLP                    | 0.821 ± 0.013 | 0.439 ± 0.022 |
| Random forest          | 0.846 ± 0.010 | 0.467 ± 0.023 |
| HMM                    | 0.672 ± 0.032 | 0.161 ± 0.012 |
| HMM+LSTM               | 0.759 ± 0.022 | 0.249 ± 0.034 |
| DMM                    | 0.815 ± 0.010 | 0.403 ± 0.026 |
| ICU/LSTM [15]          | 0.829 ± 0.012 | 0.470 ± 0.026 |
| ICU/GRU [10]           | 0.842 ± 0.016 | 0.473 ± 0.034 |
| ICU/GRU-D [6]          | 0.834 ± 0.016 | 0.517 ± 0.025 |
| **Proposed AttDMM**    | **0.865 ± 0.010** | **0.545 ± 0.029** |

Higher is better. Best value in bold.

Figure 4 depicts the evolution of prediction performance across different time periods. For both the AUROC and AUPRC, AttDMM yields favorable results. In comparison, the baseline models vary in their performance at different steps. The random forest has an AUROC similar to AttDMM until 84 hours to discharge/death. After that, AttDMM outperforms the random forest. The ICU/GRU and ICU/GRU-D show increases in AUROC for the last 24 hours of hospital stays, but are ranked consistently below the AttDMM.

For the AUPRC, AttDMM and the ICU/GRU-D show a similar performance for the last 24 hours of hospital stays. However, prior to that, the ICU/GRU-D and the other baselines are consistently outperformed by AttDMM. For both AUROC and AUPRC, the other sequential networks (i.e., ICU/LSTM, ICU/GRU, and ICU/GRU-D) show a lower performance level compared to AttDMM when focusing on the time window of more than 24 hours to discharge/death (i.e., from −120 hours to −24 hours). Overall, we see consistent gains from using AttDMM over existing baselines.
Mortality risk score
0.4
0.6
0.6
0
1
1

Thereby, our model provides ICU practitioners with more time for patient over time. Thereby, ICU practitioners can track the course of
Att
provided by early intervention in order to save patient lives.

DMM outperformed existing baselines due to the way the model is formalized. Specifically, our AttDMM is modeled so that it jointly captures (1) long-term disease dynamics (via an attention network) and (2) different latent disease states in patient trajectories (via a latent variable model). By combining these two characteristics, AttDMM presents a powerful tool for modeling health progression in ICU's, as was confirmed in a superior prediction performance.

Clinical relevance: AttDMM outperformed both baselines from clinical practice (e.g., SAPS-II [25]) and state-of-the-art machine learning for ICU mortality prediction [6, 10, 15]. In practice, such an improvement in performance allows the triggering of warnings when health conditions deteriorate several hours earlier. Thereby, our model provides ICU practitioners with more time for early intervention in order to save patient lives.

Precision of real-time risk score: The real-time risk score provided by AttDMM provides the health trajectory of an ICU patient over time. Thereby, ICU practitioners can track the course of the disease, and, by identifying patients at risk, adapt their decision-making concerning treatment planning. Owing to our probabilistic setting, AttDMM additionally outputs the confidence interval of the predicted risk score. Thus, ICU practitioners are informed about the precision of the prediction and they might postpone a discharge until more measurements are available. This is a particular benefit over risk scoring (e.g., SAPS, APACHE) in clinical practice, in which similar confidence intervals are lacking.

Generalizability: We developed AttDMM as a generative deep probabilistic model tailored for ICU mortality. Needless to say, AttDMM is not limited to an ICU setting and might facilitate other use cases, in which inferences from time-series have to be made that have long-term dependencies and are driven by latent dynamics, such as clickstream analytics or churn prediction.

7 CONCLUSION

In intensive care units (ICUs), patients are subject to constant monitoring in order to guide clinical decision-making. While state-of-the-art models can capture long-term dependencies, these cannot formally account for latent disease states in health trajectories. To fill this gap, we developed a novel probabilistic generative model, specifically, an attentive deep Markov model (AttDMM). To the best of our knowledge, this is the first attentive deep Markov model. AttDMM was co-developed with clinical researchers who emphasized the benefit of real-time risk scoring. In our numerical experiments, AttDMM yielded performance improvements in both AUROC and AUPRC over the state of the art by more than 2%. Altogether, this enables more accurate risk scoring so that lives of patients at risk of mortality can be saved.

REFERENCES

[1] Ahmed M Alaa, Jinsung Yoon, Scott Hu, and Mihaela van der Schaar. 2017. Individualized risk prognosis for critical care patients: A multi-task Gaussian process model. arXiv:1705.07674 (2017).
[2] Ahmed M Alaa, Jinsung Yoon, Scott Hu, and Mihaela Van der Schaar. 2017. Personalized risk scoring for critical care prognosis using mixtures of Gaussian processes. IEEE Transactions on Biomedical Engineering 65, 1 (2017), 207–218.
[3] Ahmed Alam, Matthias Dittrich, Anna Smitsova, Dominique Brodieck, and Michael Krauthammer. 2020. Patient similarity analysis with longitudinal health data. (2020). arXiv:2005.06630
[4] Nicola Bartolomeo, Paolo Treteroli, and Gabriella Serio. 2011. Progression of liver cirrhosis to HCC: An application of hidden Markov model. BMC Medical Research Methodology 11, 1 (2011).
[5] D. Christopher Bouch and Jonathan P. Thompson. 2008. Severity scoring systems in the critically ill. Continuing Education in Anaesthesia, Critical Care & Pain 8, 5 (2008), 181–185.
[6] Zhengxing Che, Sanjay Purushotham, Kyungbyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. Scientific Reports 8, 1 (2018).
[7] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk prediction with electronic health records: A deep learning approach. In SIAM ICDM.
[8] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuette, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In NIPS. 3504–3512.
[9] Edward Choi, Andy Schuette, Walter F. Stewart, and Jimeng Sun. 2016. Medical concept representation learning from electronic health records and its application on heart failure prediction. arXiv:1602.03680 (2016).
[10] Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. 2018. Deep ensemble tensor factorization for longitudinal patient trajectories classification. arXiv:1811.10501 (2018).
[11] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics (1988), 837–845.
[12] Stacia M. DeSantis and Dipankar Bandyopadhyay. 2011. Hidden Markov models for zero-inflated Poisson counts with an application to substance use. Statistics in Medicine 30, 14 (2011), 1678–1694.

Figure 5: Evolution of risk scores over time for two example patients, namely a patient with in-hospital mortality (top) and a patient with hospital discharge (bottom).
A ESTIMATION PROCEDURE

The true posterior of the latent variables \( p(z_{1:T} \mid x_{1:T}, m_{1:T}, s) \) is computationally intractable. Because of this, we adopted stochastic variational inference and leverage the approximated posterior \( q(z_{1:T} \mid x_{1:T}, m_{1:T}, s) \) as the proxy of the true posterior. Throughout the iterations, the approximated posterior is getting closer, in terms of Kullback-Leibler (KL) divergence, to the true posterior. This is achieved by maximizing the ELBO. For this, we use stochastic optimization via unbiased Monte Carlo estimates of the gradient, details are found in [39]. For the ICU mortality prediction task, we incorporate the maximization of the ELBO into the loss function as a regularization term. Overall, in the following, we show how to estimate AttDMM via an end-to-end training task by a tailored ELBO.

**Loss function:** The loss function of AttDMM is given by

\[
\mathcal{L}(y, \hat{y}, x) = \ell(y, \hat{y}) - \alpha \text{ELBO}(x), \tag{29}
\]

where the two terms \( \ell(y, \hat{y}) \) and the ELBO are described below.

**Cross-entropy loss:** The term \( \ell(y, \hat{y}) \) denotes weighted cross-entropy loss between the observed label \( y \) and the corresponding mortality prediction \( \hat{y} \). It is given by

\[
\ell(y, \hat{y}) = -\rho y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \tag{30}
\]

with a weight \( \rho = \frac{|\{y \in \{\text{Y=\"discharge\}"}\}|}{|\{y \in \{\text{Y=\"death\}"}\}|} \) denoting the discharge-to-death ratio. Introducing such weight further helps in discriminating the minority class (i.e., in-hospital mortality) in the imbalanced dataset.

**ELBO:** We denote the evidence lower bound via ELBO. It serves as the regularization term of the loss function. The strength of the regularization is parametrized by \( \alpha \). The ELBO formulation is given by

\[
\text{ELBO}(x) = \mathbb{E}_{q(z_{1:T} \mid x_{1:T}, m_{1:T}, s)} [\log p(x_{1:T} \mid z_{1:T})]
- \text{KL}(q(z_{1:T} \mid x_{1:T}, m_{1:T}, s) \parallel p(z_{1:T})). \tag{31}
\]

The expectation term denotes the expected log-likelihood of \( x_{1:T} \) given the latent variables \( z_{1:T} \). The KL-divergence measures the similarity between the posterior approximation and the prior formulation of the latent variables \( z_{1:T} \).

In this ELBO formulation, \( q(z_{1:T} \mid x_{1:T}, m_{1:T}, s) \) is decomposed into further terms, which are produced by the posterior approximation, given by

\[
q(z_{1:T} \mid x_{1:T}, m_{1:T}, s) = \prod_{t=1}^{T} q(z_{t} \mid z_{t-1}, s, x_{1:T}, m_{2:T}). \tag{32}
\]

Similarly, \( p(x_{1:T} \mid z_{1:T}) \) is decomposed into individual terms of \( x_{t} \), which are produced by the emission network. Specifically, we can rewrite

\[
p(x_{1:T} \mid z_{1:T}) = \prod_{t=1}^{T} p(x_{t} \mid z_{t}). \tag{33}
\]

We make further adjustments to handle missing measurements inside the vector \( x_{t} \). For this, AttDMM is specified to train on actual measurements (i.e., ignoring imputed values), which is achieved by

\[
p(x_{t} \mid z_{t}) = \prod_{i : m_{it}=1} p(x_{it} \mid z_{t}). \tag{34}
\]

**Sampling:** Computation of the above ELBO is analytically intractable due to latent variables being modeled by a continuous distribution and the non-linearity encoded in AttDMM. Therefore, we leverage Monte Carlo sampling and, thereby, we estimate ELBO via stochastic variational inference based on sampled latent variables.

B DESCRIPTIVE STATISTICS OF MIMIC-III

Table 3: Features from MIMIC-III used in predictions

| Feature | Type | Missingness (%) |
|---------|------|-----------------|
| Glasgow coma scale | time-series | 70.04 |
| Systolic blood pressure | time-series | 49.36 |
| Heart rate | time-series | 48.69 |
| Body temperature | time-series | 70.46 |
| PaO2/FiO2 | time-series | 93.88 |
| Urinary output | time-series | 57.16 |
| Serum urea nitrogen level | time-series | 88.88 |
| White blood cells count | time-series | 89.84 |
| Serum bicarbonate level | time-series | 89.11 |
| Sodium level | time-series | 87.63 |
| Potassium level | time-series | 85.29 |
| Bilirubin level | time-series | 97.47 |
| Age | static | 0.00 |
| Acquired immunodeficiency syndrome | static | 0.00 |
| Hematologic malignancy | static | 0.00 |
| Metastatic cancer | static | 0.00 |
| Admission type | static | 0.00 |

Figure 6: Histogram of length of hospital stay after ICU admission. Recall that hospital stays with less than 48 hours were removed during preprocessing as in [6, 38]

C IMPLEMENTATION DETAILS

The dataset is split randomly into 5 separate folds. Each fold has the same ratio of in-hospital mortality and hospital discharge. To quantify the uncertainty regions for the predictions, we used the following approach: each model was trained 5 times on each fold, in which one fold represents the test set and the remaining 4 folds were randomly split into a training set (3 folds) and a validation set (1 fold; for early stopping). Afterwards, the performance was
Table 5: Hyperparameters used for tuning the baseline models.

| Model         | Tuning parameters       | Tuning range                      |
|---------------|-------------------------|-----------------------------------|
| MLP           | Dimension of hidden layer | 2, 4, 8, 12                      |
|               | Number of hidden layers  | 1, 2, 3                           |
|               | Learning rate           | 0.00002, 0.0002, 0.001, 0.01      |
|               | Batch size              | 64, 128, 256                      |
| RF            | Number of trees         | 50, 100, 500, 1000                |
|               | Maximum depth           | 3, 5, 7, 9, 11                    |
|               | Class weight            | none, balanced                    |
| HMM+LSTM*     | Number of states        | 2, 3, ..., 10                     |
| ICU/LSTM,     | Dimension of cell state | 8, 16, 24, 48                     |
| ICU/GRU,      | Dropout rate            | 0, 0.1, 0.2, 0.5                  |
| ICU/GRU-D     | Learning rate           | 0.00002, 0.0002, 0.001, 0.01      |
|               | Batch size              | 64, 128, 256                      |

* LSTM of HMM+LSTM is tuned based on the same parameter space of ICU/LSTM.

aggregated over the different folds via macro averaging. Later, we report the standard deviation across folds.

The hyperparameters of the models are tuned according to performance on validation set. We use the hyperparameter set that provides the best overall validation score. Details on hyperparameter tuning are presented in Appendix D.

All the models were implemented in Python. We employed the following libraries: Pyro for AttDMM; PyTorch for the LSTM, GRU and MLP; Tensorflow for the GRU-D; and Scikit-learn for the random forest. Training and evaluations were performed on TITAN V GPU from NVIDIA with 12 GB of memory.

D TRAINING DETAILS

Due to the large hyperparameter space across all the models, it was computationally expensive to perform a grid search. Therefore, we adopted a ceteris paribus strategy which formally means that we tuned each parameter individually while keeping the other parameters fixed. We ran this procedure for a few loops, until we had observed the convergence of the score. The tuning parameters with the corresponding tuning ranges are listed for the AttDMM in Table 4 and for the baselines in Table 5.

For both AttDMM and the baselines of the sequential neural network (ICU/LSTM, ICU/GRU, and ICU/GRU-D) we used the Adam optimizer. These models were trained using early stopping with patience set to 20 epochs.

Table 4: Hyperparameters used for tuning AttDMM.

| Tuning parameters                      | Tuning range |
|----------------------------------------|--------------|
| Dimension of latent variable           | 8, 16, 24, 48|
| Dimension of emission hidden layer     | 16, 32, 48, 96|
| Dimension of transition hidden layer   | 32, 64, 96, 192|
| Dimension of RNN cell                  | 8, 16, 24, 48|
| Dimension of attention mechanism       | 16, 24, 36, 48|
| Dimension of MLP hidden layer          | 2, 4, 8, 12 |
| Regularization strength of ELBO (α)    | 0.001, 0.01, 0.1, 1 |
| Number of Monte Carlo samples (N)      | 1, 2, 5, 10, 20 |
| Learning rate                          | 0.000002, 0.00002, 0.001, 0.01 |
| Batch size                             | 64, 128, 256 |