Deep Recurrent Neural Networks for Mortality Prediction in Intensive Care using Clinical Time Series at Multiple Resolutions

Suparna Ghanvatkar
National University of Singapore, sups.ghanvatkar@gmail.com

Vaibhav Rajan
National University of Singapore, vaibhav.rajan@nus.edu.sg

Deep Recurrent Neural Networks for Mortality Prediction in Intensive Care using Clinical Time Series at Multiple Resolutions

Short Paper

Suparna Ghanvatkar, Vaibhav Rajan
Department of Information Systems and Analytics
School of Computing, National University of Singapore
{suparnag, vaibhav}@comp.nus.edu.sg

Abstract

Mortality models in Intensive Care Units (ICU) are important for clinical decision support tasks such as identifying high-risk patients and prioritizing their care. Previous mortality models have used predictive variables mainly from Electronic Medical Records (EMR) where each patient observation can be represented as a sparse multivariate time series. Bedside monitors are another common data source in ICUs containing high-resolution time series, which have not been explored in combination with EMR data for mortality modelling. We take the first step towards building such a model. Specialized techniques developed for sparse time series cannot be used to model multiple time series at different resolutions. To address this problem, we develop MTS-RNN, a new deep recurrent neural network architecture. Our preliminary experiments on real clinical data show that MTS-RNN outperforms state-of-the-art mortality models in predictive accuracy, highlighting the importance of using clinical time series at multiple resolutions for ICU mortality prediction.

Keywords: Intensive Care Units; Electronic Medical Records; Time Series; Deep Neural Network

Introduction

An Intensive Care Unit (ICU) has the most critically ill patients in a hospital, who require continuous monitoring and clinical support, the economic burden of ICU care is enormous. For instance, in the United States, although ICUs have less than 10% of hospital beds, ICUs contribute to 22% of total hospital costs and 5% of the total healthcare cost (Halpern and Pastores 2010). As the need for ICUs has grown worldwide, more ICUs have been created, but the availability of resources, both clinical staff and monitoring equipment, remain limited due to practical constraints (Halpern et al. 2013).

The importance of predictive models for ICU mortality (risk of death) is well recognized. At an individual patient level, it can be used to assess the severity of illness and thereby identify high-risk patients. Identifying high–risk patients can not only aid critical decisions during ICU stay such as interrupting treatments or providing Do– Not–Resuscitate orders but also enable effective triage, making ICU resources available to patients in need. Mortality models can also be used as a quality indicator for benchmarking ICU performance across hospitals and play an important role in risk adjustment and risk stratification in randomized controlled trials (Power and Harrison 2014). Further, they are used to determine the value of novel treatments, interventions and health care policies (Pirracchio et al. 2015). Thus, accurate mortality models and their effective use can play an important role in potentially reducing the clinical and economic burden of ICU care.

Continuous monitoring of ICU patients generates a wealth of clinical data presenting both, opportunities to build predictive models for decision support, as well as challenges in effectively integrating the available information (Badawi et al. 2014). Two large sources of patient data in the ICU are Electronic Medical Records (EMR) and bedside patient monitors. EMR includes, among others, data from laboratory investigations, clinical notes, medication orders, and radiology investigations. Repeated measurements

induce temporality in most of the variables in EMR. Unlike most longitudinal panel data, these time series are very sparse as they contain measurements at irregular intervals, with highly variable lengths and frequencies both across measurements and across patients. Further, there are recording biases, e.g., sicker patients are measured more often, which leads to informative missingness patterns. The other source of clinical time series, bedside monitors, have high-resolution physiological signals, also called "waveforms", such as electrocardiograms (ECG). Their frequencies can range from 1 to 125 Hz.

Previous ICU mortality models have been developed using many predictive variables mainly from EMR data. Machine learning models typically pose the problem as a supervised binary time series classification task. Each patient is associated with a binary label indicating the presence or absence of in-hospital mortality. A classifier is trained on historical data consisting of sparse time series data from EMR. Note that static (i.e., non-temporal) demographic variables, such as age and gender, are also commonly included during classifier training. A trained classifier can be deployed in Clinical Decision Support Systems to trigger alerts when the predicted label indicates high mortality risk.

Heterogeneity and temporality make modeling EMR data challenging (Huddar et al. 2016). To model sparse time series, a timescale (e.g., every hour) is chosen and missing values are imputed before classifier training. Imputation can be done in various ways, such as through the use of statistical summaries (mean, median) or the last measured value. Empirical studies have found that modeling temporal correlations during imputation leads to better imputation and forecasting performance in clinical time series (Sengupta et al. 2017). A principled approach that can model such correlations is through the use of Multi-Task Gaussian Process (MGP) (Bonilla et al. 2008) that has been effectively used with EMR data for mortality prediction in (Ghassemi et al. 2015). Recently, models such as MGP-RNN perform simultaneous imputation and classifier training that can also model the correlation of missingness patterns with the target label (Futoma et al. 2017).

To our knowledge, no previous work has used high-resolution waveforms along with EMR data for mortality modeling, and we take the first step towards building such a model. Since dense waveforms provide patient information that complements EMR-based information, our hypothesis is that by incorporating data from this additional source, we can obtain better patient models, which in turn may lead to improved mortality prediction. However, we face a modeling challenge: while sparse time series data from EMR can be modeled well using methods designed for such data, high-resolution waveform data cannot be directly used in them. For example, MGP is designed for sparse data and training them is computationally intractable for such dense time series. To address this limitation, we develop a new Multi Time Scale Recurrent Neural Networks (MTS-RNN) architecture that effectively models both dense and sparse time series.

Our architecture has several advantages. It generalizes easily: all the sparse time series are modeled using MGP while each dense time series is modeled using individual RNNs. The interaction between the RNNs and end-to-end training enables MGP and each RNN to learn from heterogeneous input streams. The use of MGP also facilitates easy incorporation of expert-based domain knowledge and personalization to individual patients through the use of kernels. Further, the architecture can model varying temporal gaps in each measurement, including those in the dense time series. Our preliminary experiments on real ICU data of 1,950 patients show promising results suggesting that the addition of dense waveforms to EMR data is indeed valuable for ICU mortality prediction. In our experiments, MTS-RNN outperforms state-of-the-art models that were designed primarily for EMR-based sparse clinical time series.

Related Work

Traditional scoring systems for ICU mortality risk, such as APACHE (Zimmerman et al. 2006) have been developed to assess how care procedures, medications and other clinical factors affect mortality in ICUs. They are not calibrated for patient specific predictions. The efficacy of machine learning models over such risk scoring systems was demonstrated in the Physionet ICU Mortality Prediction Challenge (Silva et al. 2012) and the following research using the same dataset (e.g., Pirracchio 2015; Bhattacharya et al. 2017).

The key challenge of modeling sparsity due to missing values in clinical time series continues to remain an active research area. Most classical time series models assume regularity of measurements and cannot be directly used to model such clinical time series, and so, Ghassemi et al. (2015) proposed the use of MGP to model such data. It has been found that the missing values are not random and the missingness patterns

themselves are correlated with target labels in clinical tasks, which has led to models that combine imputation and classifier training in end-to-end models (Futoma et al. 2017; Che et al. 2018).

State-of-the-art results for various predictive tasks using EMR data have been achieved by deep learning models based on convolutional or recurrent neural networks (e.g., Choi et al. 2016; Lipton et al. 2015). A survey on deep learning for clinical data modeling can be found in (Shickel et al. 2018). For ICU mortality prediction, LSTM-based architectures have shown the best results (Harutyunyan et al. 2019; Rajkomar et al. 2018; Song et al. 2018). However, all these models only use data available in EMR comprising sparse time series, static variables and nursing notes. The combination of another text-based modality, i.e., nursing notes, in addition to laboratory investigations showed improved performance in previous studies on mortality modelling (Ghassemi et al. 2015; Rajkomar et al. 2018), that motivates the exploration of an additional previously unused source of information, high-resolution waveforms, for mortality modeling. Moreover, it has been shown that information in these signals strongly correlates to the mortality of patients (Zhang et al. 2016).

High resolution waveforms have been independently studied and used in several predictive models. For instance, Sadeghi et al. (2018) extract 12 statistical features from such signals to train classifiers such as SVM, decision trees, and random forest to predict mortality. To the best of our knowledge, very few works have integrated high-resolution waveforms with sparse clinical time series from EMR data. A recent work of Xu et al. (2018) develops an architecture for the closely related application of decompensation prediction but has not been evaluated for mortality prediction. A recent review of deep learning methods for using physiological signals in clinical predictions can be found in Faust et al. (2018).

Background

Multitask Gaussian Process

Gaussian process (GP) naturally models irregular spacing and differing number of observations per time series in clinical data. Multi-task Gaussian Process (MGP) is an extension of GP to model multiple time series simultaneously. We briefly describe MGP following the notation used in Futoma et al. (2017).

Let $f_{im}(t)$ be a latent function representing the true values of variable m for patient i at time t. We denote the total number of time series variables by M and the length of each patient's record by T_i. All the sparse time series variables of the i^{th} patient are denoted by $\mathbf{Y}_i = [\mathbf{y}_{i1},\ldots,\mathbf{y}_{iM}] \in \mathrm{R}^{T_i \times M}$, where $y_{im} \in \mathrm{R}^{T_i}$ is a univariate time series of the m^{th} variable. Note that only a subset of the M variables are recorded at each observation time where the recording times differ across patients. Further, the number of observations also differ across patients, i.e., usually $T_i \neq T_{i'}$ for $i \neq i'$.

Consider the regression model $\mathbf{y}_{im}(t) = f_{im}(t) + \epsilon$, where $\epsilon \sim \mathrm{N}(0,\sigma_m)$ is a noise term. A GP is characterized by a mean function (usually assumed to be zero) and a covariance function, and can be viewed as a prior probability density over all possible functions for the given sequence. MGP places GP priors over the latent functions, with a shared correlation function over time. Assuming centered data, the likelihood of the observed multivariate time series is given by a multivariate Gaussian (N):

$$\mathrm{vec}(\mathbf{Y}_i) \equiv \mathbf{y}_i \sim N(0, \Sigma_i), \quad \Sigma_i = K^M \otimes K^{T_i} + D \otimes I$$

where $\mathbf{y}_i$ is a stacked vector of all M variables at the T_i observation times, denotes $\otimes$ the Kronecker product, K^M is a full-rank $M \times M$ covariance matrix, K^T is a $T_i \times T_i$ correlation matrix for the observation times t_i (specified by a kernel) and D is a diagonal matrix of noise variances $\{\sigma_m^2\}_{m=1}^M$. More details can be found in Bonilla et al. (2008). A key aspect of the model is that K^M allows information from more frequently sampled variables to improve the learning of more sparse variables. Also note that the covariance between different variables and between different points in time is explicitly separated. The $MT_i \times MT_i$ covariance matrix Σ_i only needs to be computed at the observed values, yielding a sparse matrix.

Given the likelihood of the observed data and the GP prior, the posterior distribution can be computed for evenly spaced time-points (X_i). This posterior latent time-series is represented by Z_i over the grid points X_i. For $\boldsymbol{z}_i = \mathrm{vec}(\boldsymbol{Z}_i)$, the posterior is normally distributed with mean (μ_{z_i}) and covariance (Σ_{z_i}):

$$\mu_{z_i} = (K^M \otimes K^{X_i T_i})\Sigma_i^{-1} y_i \,, \;\; \Sigma_{z_i} = (K^M \otimes K^{X_i}) - (K^M \otimes K^{X_i T_i})\Sigma_i^{-1}(K^M \otimes K^{T_i X_i})$$

where $K^{X_i T_i}$ is the correlation matrix between evenly spaced grid times x_i and observation times t_i, and K^{X_i} is the correlation matrix between the grid times x_i with itself, specified by the chosen kernel. Following Futoma et al. (2017), we choose the Ornstein-Uhlenbeck kernel function, with a single length-scale hyperparameter η that has been used previously for noisy physiological time series. MGP parameters include the kernel hyperparameters and noise variances, which are typically learnt by maximum-likelihood estimation (Bonilla et al. 2008).

Thus, MGP enables us to transform sparse irregularly sampled multivariate time series (**y**) to a representation (**z**) that is evenly spaced in time (e.g., every hour). The hyperparameters of MGP trained on clinical time series have been used as features for ICU mortality prediction (Ghassemi et al. 2015). MGP is an elegant model for sparse multivariate time series but its computational complexity, mainly due to the inversion of Σ_i (which takes $O(M^3 T_i^3)$ time and is required for both learning and inference) in posterior computation, renders it impractical for dense time series that have large T_i.

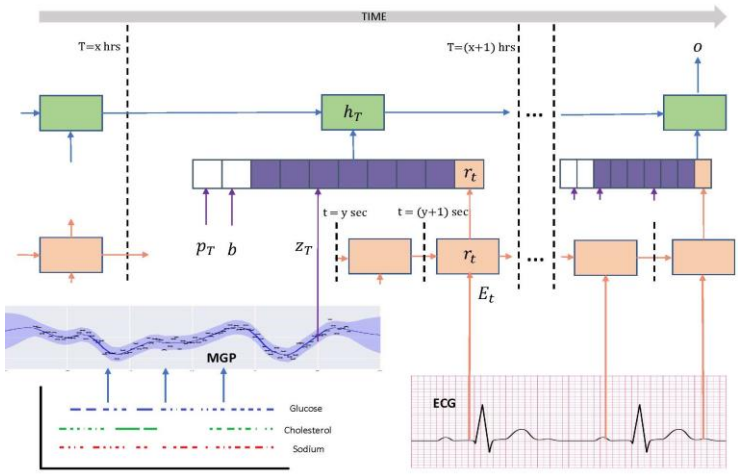
MGP-RNN

MGP-RNN takes in raw sparse time series data, transforms it through MGP to a uniform representation on an evenly spaced grid, and uses the latent function values (**z**) as input to an RNN. MGP-RNN can be used for time series classification and has been evaluated for sepsis prediction (Futoma et al. 2017). Unlike previous works that separate MGP training from the final classification task, MGP-RNN is an end-to-end classifier that learns both the MGP parameters as well as RNN parameters jointly. The advantage of such an approach is that the uncertainty in the imputed values is modeled during RNN training, enabling the classifier to rely more on the observed values and less on the imputed values. Such uncertainty modeling also indirectly models the correlations of the missing patterns with the target label. The latent function values are concatenated with other patient information – static baseline variables and medication information, and the concatenated vector is used as an input to the RNN. Thus, the RNN is able to learn complex time-varying interactions among the clinical data.

Our Model: MTS-RNN

To model sparse and dense time series simultaneously, we design a new RNN architecture called Multi Time Scale RNN (MTS-RNN). We determine grid points at two different time-scales for sparse and dense data, e.g., every hour and every second. We use statistical features (mean and standard deviation) of the dense time series for each interval (i.e. every second) as input to an RNN and obtain hidden representations of the signal at each grid point (i.e. every hour). These representations are concatenated with the latent representations obtained using MGP for sparse time series, as described above, and used as input to a second RNN which is used for binary classification. Figure 1 shows a schematic of the model.

Let $D_i = \{b_i, t_i, Y_i, E_i, P_i, o_i\}$ represent the i^{th} patient's data, where b_i represents static baseline variables (e.g., demographics that are measured during hospital admission), t_i represents the time points of measurements of the sparse time series, Y_i represents the sparse time series measurements corresponding to the t_i (e.g., laboratory measurements), E_i represents the dense time series measurements (e.g., ECG), P_i represents the medications administered and o_i represents the binary label of in-hospital mortality. The variables b_i and P_i are modeled differently: b_i is not temporal; P_i is a P-dimensional vector containing the amounts of each of the P medications administered at every grid point (that, unlike physiological variables, does not have a latent structure). Statistical features obtained every second from the dense time series E_i are used as input to an RNN which learns the hidden representation r_{ij} at each j^{th} grid point. These grid points correspond to the grid points of MGP that is used to model the sparse time series $\{t_i, Y_i\}$. At each grid point (indexed by j) the learnt latent function z_{ij} from the MGP is concatenated with the vectors b_i, P_{ij} and r_{ij}. This concatenated vector is used as input to the second RNN. The entire model is trained end-to-end using the procedure described in Futoma et al. (2017).



p_T = medications at grid point T | z_T = Posterior of MGP at grid point T | r_t, h_T = hidden representations of RNN

b = baseline demographics | E_t = waveform input at time t | o = predicted label

Figure 1. Schematic for MTS-RNN model

MTS-RNN can be seen as a variant of stacked RNNs (El Hihi and Bengio 1996) that are constructed by stacking multiple recurrent hidden layers, with the goal of capturing representations at different timescales in each recurrent level. In contrast, in our model, each RNN models a distinct temporal process with different inputs at each time step. Thus, MTS-RNN can also be viewed as multiple interacting RNNs.

Our model has numerous advantages. By using MGP-based representations it inherits all the benefits of MGP-RNN, viz., uncertainty-aware modeling of sparse time series and the propagation of uncertainty in RNN training. Due to the interaction between RNNs, through end-to-end training, the RNN for dense time series also 'learns' from other patient data; thus, even the dynamics of the dense time series is modeled better and may be useful in forecasting applications. Moreover, additional information specific to the dense time series can be used as input directly to this RNN. There may be multiple dense time series: each dense time series can be added as another RNN and the architecture can generalize easily. Furthermore, there can be temporal gaps of varying frequency and length between measurements in the dense time series as well. The MTS-RNN architecture can also model such temporal gaps – grid intervals that do not have dense time series measurements can have only latent functions from the sparse time series, with the previous hidden representation (r_{ij}) of the dense time series propagated unmodified.

Experiments

Data

We use a publicly available EMR dataset, MIMIC-III (Johnson et al. 2016), containing clinical data of over 40,000 patients admitted to the ICUs in Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. The data is extracted from EMR and de-identified in compliance with HIPAA standards (Johnson et al. 2016). This data includes vital sign measurements, laboratory test results, medications and clinical notes and other demographic and billing related information. The MIMIC III Waveform Database (Goldberger et al. 2000) contains multiple physiological signals ("waveforms"), sampled at high frequencies such as 125 Hz, and time series for vitals ("numerics"), sampled at lower frequencies of 1 second or 1 minute. These have been collected from bedside monitors and include ECG, continuous arterial blood pressure (ABP), pulse plethysmograms and other signals. From this database, 10,282 patient records have been matched to the MIMIC-III EMR dataset.

We follow the data preprocessing steps described in Harutyunyan et al. (2019) that use MIMIC-III EMR data. Out of the 10,282 matched patients, we exclude all neonatal and patients under 16 years, as critical care protocols are different for these patients. ICU stays with duration less than 24 hours and conflicting

diagnoses are excluded. Also, patients whose demographics are missing are excluded. Among the waveforms, ECG-II signal was present in the maximum number of patients. For our preliminary experiments, we choose only those patients who have ECG-II signal. As no other selection criteria was applied, the resulting cohort is heterogeneous and hence representative of a real clinical setting. We use the first 24 hours of clinical data to predict mortality risk at the end of the first day of ICU stay for all patients.

The ECG-II recordings available are sometimes duplicate or split into multiple recordings for a given patient's ICU stay. For a patient encounter, we merge multiple ECG-II recordings with maximum overlap of 5 minutes for every ICU stay. We eliminate patient encounters with duplicate ECG-II recordings. This selected cohort consists of 7317 patient encounters, with 788 positive labels for in-hospital mortality.

Table 1. List of features used

Sparse clinical time series variables (Y_i)	Medications (P_i)	Static (b_i)
Capillary refill rate, Cholesterol, Creatinine, Diastolic blood pressure, Eosinophils, Fraction inspired oxygen, Glucose, Heart Rate, Height, Hemoglobin, Magnesium, Mean blood pressure, Oxygen saturation, Potassium, Red blood cell count, Respiratory rate, Sodium, Systolic blood pressure, Temperature, Weight, pH	Sodium Chloride 0.9%, Propofol, Insulin, Heparin, Potassium Chloride	Ethnicity, Gender, Age, Height, Weight

In addition to the data cleaning steps described in Harutyunyan et al. (2019), we eliminate duplicate entries (multiple measurements of the same variable at the same time point) and remove entries containing ranges (e.g., 3-7) instead of single values. Wherever multiple units of measurements were used, appropriate unit conversions were performed. The binary label of in-hospital mortality is set to 1 if the Date of Death is between the hospital admission and discharge times. The EMR variables chosen are listed in Table 1. These features are among the most frequently measured variables in the ICU, including this database, and have been used in previous literature (Silva et al. 2012; Harutyunyan et al. 2019). We use a single dense time series ECG-II from the Waveform Database. Thus, the i^{th} patient encounter ($D_i = \{b_i, t_i, Y_i, E_i, P_i, o_i\}$) has static baseline variables (b_i), sparse time series of clinical measurements (Y_i) at observation times (t_i), medication administration information (p_i), dense time series of ECG measurements (E_i) and the mortality label (o_i).

Evaluation

We compare the performance of MTS-RNN on ICU mortality prediction with that of the following baseline methods. The LSTM model (Harutyunyan et al. 2019; Song et al. 2018) is the state-of-art model for mortality prediction. Since MTS-RNN extends the architecture of MGP-RNN, we also use MGP-RNN as another baseline. All the three methods use LSTM for the hourly time series. The difference lies in the imputation method (i.e. MGP) and additional LSTM for waveform (in case of MTS-RNN). Similar to the approach used by Sadeghi et al. (2018), we extract statistical features from clinical time series and use standard classifiers (k-nearest neighbor (KNN), support vector machine (SVM) and multilayer perceptron (MLP)) as baselines.

MGP-RNN can handle all the input datatypes except dense time series. The time and memory requirement for using high frequency waveform data are used is impractically high. The LSTM model has been designed for sparse time series and does not use medication information or dense time series. Hence we only use sparse time series data and static variables for predicting using this model. Note that their model imputes missing values with previous measured values. For the waveforms, we impute missing values with mean of the signal, before obtaining the mean and standard deviation for each second. For the remaining classifiers, each patient encounter is transformed into a single vector by using summary statistics of both sparse and dense time series that are concatenated with static variables. The summary statistics are count, mean, standard deviation, minimum value, median and maximum value.

Five-fold cross validation is used to evaluate all the models, by randomly dividing the patient encounters into five groups. To address the imbalance in the dataset, we use an under-sampling strategy during training, within each fold. To avoid losing information through subsampling we use the Condensed Nearest Neighbour undersampling method (Hart, 1968), that uses a 1-nearest neighbor (1-NN) rule to iteratively

decide which majority class sample should be used. The method begins with a training set comprising all the minority class samples. Then using the 1-NN rule on the training set, only the misclassified samples are iteratively added from the majority class into the training set.

We use the following evaluation metrics: *Sensitivity* or *Recall* measures the proportion of actual positives that are correctly identified, and *Specificity* measures the proportion of actual negatives that are correctly identified. *Precision* measures the proportion of correctly identified positives among the predicted positives. The *F1-score*, an overall measure of accuracy, is the harmonic mean of Precision and Recall. These metrics are for a specific threshold used with a classifier (set to 0.5 in our experiments). Threshold independent metrics include Area Under Curve of the Receiver Operating Characteristic Curve (*AUROC*) and Area Under Precision-Recall Curve (*AUPR*). ROC and precision-recall curves show the tradeoffs between, respectively, sensitivity-specificity and precision-recall for different thresholds. The area under these curves indicate overall performance over all thresholds. In all metrics higher values indicate better performance. Thus, higher AUPR and F1 scores indicate that the classifier is returning accurate results (high precision), and a large proportion of all positive results (high recall, implying better identification of high-risk patients).

Results

Table 2 shows the classification performance of MTS-RNN and all the baseline methods.

Table 2. Evaluation results: Averages over 5-fold CV (Standard Deviation)

	AUROC	**AUPR**	**Sensitivity**	**Specificity**	**Precision**	**F1 Score**
SVM	0.5 (0)	0.1076 (0.0075)	0 (0)	1 (0)	0 (0)	0 (0)
KNN	0.512676 (0.0075)	0.112754 (0.0112)	0.0531 (0.0162)	0.9724 (0.0135)	0.2057 (0.0666)	0.08210 (0.0229)
MLP	0.51154 (0.0193)	0.11048 (0.0107)	0.3949 (0.5185)	0.6282 (0.4993)	0.08262 (0.0805)	0.0984 (0.1043)
LSTM	0.68026 (0.0177)	0.2121 (0.0269)	0.2812 (0.1221)	0.89347 (0.0476)	0.2423 (0.0312)	0.25096 (0.0569)
MGP-RNN	0.7096 (0.0306)	0.2341 (0.0465)	0.4865 (0.0324)	0.2186 (0.0159)	0.0698 (0.0049)	0.1220 (0.0081)
MTS-RNN	0.7094 (0.0193)	0.24216 (0.0122)	0.4946 (0.0435)	0.7876 (0.0129)	0.21903 (0.01412)	0.3031 (0.0167)

We observe that SVM, MLP and KNN, that only use statistical summaries and do not explicitly model temporality have poor performance. Their AUROC values are close to 0.5, indicating almost no improvement over random guesses, and their lack of discrimination results in low sensitivity. All the other methods – LSTM, MGP-RNN and MTS-RNN, explicitly model temporality and have much better performance. Both the models that use MGP to model sparse time series outperform LSTM in AUROC, AUPR and sensitivity. This could be due to better imputation of missing values in the time series using MGP, which models the correlations between different clinical time series and/or due to additional features – static variables, medications and dense time series, used in MGP-RNN and MTS-RNN. Finally, MTS-RNN is better than that of MGP-RNN with respect to AUPR, F1-score, sensitivity, specificity and precision, and they have comparable AUROC. This suggests that the multiple time scale architecture of MTS-RNN leads to improvements in classification performance. This also lends evidence to the trend observed previously: explicit modeling of the temporality of dense time series in MTS-RNN improves performance compared to statistical summaries within MGP-RNN. Note that MGP-RNN cannot model dense time series directly, nor can it scale with an increasing number of measurements due to its cubic time complexity. The performance improvement of MTS-RNN over LSTM comes at the cost of higher computational cost in training due to the

$O(M^3T_l^3)$ factor in MGP, and an additional LSTM with more time steps (O(W) per time step). With respect to F1-score, MTS-RNN significantly outperforms other baselines (p-value 0.001204, using Friedman Test).

Initial Contributions and Future Plan

In this study, we develop the MTS-RNN architecture to simultaneously model dense and sparse clinical time series along with other clinical measurements found in ICUs. To our knowledge, this is the first model to predict mortality using combined patient data from two large sources in the ICU: Electronic Medical Records (containing static data and sparse time series) and Bedside Monitors (containing dense time series). MTS-RNN can effectively model multiple input data types comprising static data, sparse and dense time series in a generalizable architecture that learns from all the different input data modalities. Our initial experiments suggest that such an architecture for combining dense and sparse time series can improve the accuracy of ICU mortality prediction over previously used architectures for sparse time series only.

Our model can be used to assess the severity of illness and identify high-risk patients, which in turn can aid critical decision-making tasks within the ICU and has the potential to improve the efficacy of clinical decision support systems. Further, through the use of multiple modalities as inputs, our model can potentially counter the biases in individual data inputs which has been recognized as a problem in systems based on Electronic Health Records (Hripscak and Albers, 2012). The incorporation of data from bedside monitors, that is not controlled directly by clinical staff and is therefore unbiased, can address this problem.

This work can be extended in many ways. Further experiments using larger cohort sizes, more clinical variables and data from other hospitals are required to conclusively establish the superior predictive performance of MTS-RNN. Additional clinical variables include the use of more dense time series variables as well as other modalities such as clinical notes and imaging data. Experiments are also required to evaluate how well MTS-RNN can model varying temporal gaps in each of the dense time series. Another avenue of exploration is the development of combined knowledge-based and data-driven models, which is facilitated by the use of kernels in MGP. Prior knowledge about the correlations of the clinical investigations can improve MGP representations that in turn may improve the classification. This is particularly useful if the number of available training samples is low. Knowledge graphs contain relationships between various clinical entities (such as medications, diseases, and lab investigations) in the form of heterogeneous graphs. Correlations between clinical investigations can be derived from such graphs and used for kernel initialization. Kernels are also useful in personalization: if there are known patient subtypes, then different MGP models can be learnt on these different groups of patients before being used to train MTS-RNN.

References

Badawi O, Brennan T, Celi LA, Feng M, Ghassemi M, Ippolito A, Johnson A, Mark RG, Mayaud L, Moody G, Moses C, Naumann T, Nikore V, Pimentel M, Pollard TJ, Santos M, Stone DJ, Zimolzak A. 2014. "Making Big Data Useful for Health Care: A Summary of the Inaugural MIT Critical Data Conference", *JMIR Med Inform* (2:2), e22

Bhattacharya, S., Rajan, V., and Shrivastava, H. 2017. "ICU mortality prediction: A classification algorithm for imbalanced datasets," in *Thirty-First AAAI Conference on Artificial Intelligence*.

Bonilla, E. V., Chai, K. M., and Williams, C. 2008. "Multi-task Gaussian process prediction," in *Advances in Neural Information Processing Systems,* pp. 153–160.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. 2018. "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports* (8:1), p. 6085.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. 2016. "Doctor AI: Predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference,* pp. 301–318.

El Hihi, S. and Bengio, Y. 1996. "Hierarchical recurrent neural networks for long-term dependencies," in *Advances in Neural Information Processing Systems,* pp. 493–499.

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., and Acharya, U. R. 2018. "Deep learning for healthcare applications based on physiological signals: A review," *Computer Methods and Programs in Biomedicine* (161), pp. 1–13.

Futoma, J., Hariharan, S., and Heller, K. 2017. "Learning to detect sepsis with a multitask Gaussian process RNN classifier" in *Proceedings of the 34th International Conference on Machine Learning* (70), pp. 1174-1182.

Ghassemi, M., Pimentel, M. A., Naumann, T., Brennan, T., Clifton, D. A., Szolovits, P., and Feng, M. 2015. "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence,* pp. 446-453

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. 2000. "PhysioBank, PhysioToolkit, and PhysioNet," *Circulation* (101:23).

Halpern, N. A. and Pastores, S. M. 2010. "Critical care medicine in the United States 2000–2005:an analysis of bed numbers, occupancy rates, payer mix, and costs," *Critical Care Medicine* (38:1), pp. 65–71.

Halpern, N., Pastores, S., Oropello, J.,and Kvetan, V. 2013. "Critical care medicine in the United States: addressing the intensivist shortage and image of the specialty," *Critical Care Medicine* (41:12), pp. 2754–2761.

Hart, P. 1968. "The Condensed Nearest Neighbor Rule (Corresp.)," *IEEE Transactions on Information Theory* (14:3), Citeseer, pp. 515–516.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. 2019. "Multitask Learning and Benchmarking with Clinical Time Series Data," *Scientific Data* (6:1), p. 96.

Hripcsak, George, and David J. Albers. 2012. "Next-generation phenotyping of electronic health records." *Journal of the American Medical Informatics Association* (20:1), p. 117-121.

Huddar, V., Desiraju, B. K., Rajan, V., Bhattacharya, S., Roy, S., and Reddy, C. K. 2016. "Predicting complications in critical care using heterogeneous clinical data," *IEEE Access* (4), pp. 7988–8001.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. 2016. "MIMIC-III, a freely accessible critical care database," *Scientific Data* (3) 2016, p. 160035.

Lipton, Z. C., Kale, D. C., Elkan, C., and Wetzell, R. 2016. "Learning to diagnose with LSTM recurrent neural networks," in *International Conference on Learning Representations*.

Pirracchio, R., Petersen, M. L., Carone, M., Rigon, M. R., Chevret, S., and Laan, M. J. van der 2015. "Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study," *The Lancet Respiratory Medicine* (3:1), pp. 42–52.

Power, G. S. and Harrison, D. A. 2014. "Why try to predict ICU outcomes?" *Current Opinion in Critical Care* (20:5), pp. 544–549.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. 2018. "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine* (1:1), p. 18.

Sadeghi, R., Banerjee, T., and Romine, W. 2018. "Early hospital mortality prediction using vital signals," *Smart Health* (9), pp. 265–274.

Sengupta, A., Prathosh, A., Shukla, S. N., Rajan, V., and Reddy, C. K. 2017. "Prediction and imputation in irregularly sampled clinical time series data using hierarchical linear dynamical models," in *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* pp. 3660–3663.

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. 2018. "Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE Journal of Biomedical and Health Informatics* (22:5), pp. 1589–1604.

Silva, I., Moody, G., Scott, D. J., Celi, L. A., and Mark, R. G. 2012. "Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012." *Computing in Cardiology* (39), pp. 245–248.

Song, H., Rajan, D., Thiagarajan, J., and Spanias, A. 2018. "Attend and diagnose: Clinical time series analysis using attention models," in *32nd AAAI Conference on Artificial Intelligence*, pp. 4091–4098.

Xu, Y., Biswal, S., Deshpande, S. R., Maher, K. O., and Sun, J. 2018. "RAIM: Recurrent Attentive and Intensive Model of Multimodal Patient Monitoring Data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2565–2573.

Zhang, D., Shen, X., and Qi, X. 2016. "Resting heart rate and all-cause and cardiovascular mortality in the general population: a meta-analysis," *CMAJ* (188:3), E53–E63.

Zimmerman, J. E., Kramer, A. A., McNair, D. S., and Malila, F. M. 2006. "Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients," *Critical Care Medicine* (34:5), pp. 1297–1310.