# Predicting ICU Mortality Risk by Grouping Temporal Trends from a Multivariate Panel of Physiologic Measurements

**Yuan Luo**
Northwestern University
yuan.luo@northwestern.edu

**Yu Xin** and **Rohit Joshi**
MIT
{yuxin,rjoshi}@mit.edu

**Leo Celi**
Harvard Medical School
lceli@bidmc.harvard.edu

**Peter Szolovits**
MIT
psz@mit.edu

**Abstract**

ICU mortality risk prediction may help clinicians take effective interventions to improve patient outcome. Existing machine learning approaches often face challenges in integrating a comprehensive panel of physiologic variables and presenting to clinicians interpretable models. We aim to improve both accuracy and interpretability of prediction models by introducing Subgraph Augmented Non-negative Matrix Factorization (SANMF) on ICU physiologic time series. SANMF converts time series into a graph representation and applies frequent subgraph mining to automatically extract temporal trends. We then apply non-negative matrix factorization to group trends in a way that approximates patient pathophysiologic states. Trend groups are then used as features in training a logistic regression model for mortality risk prediction, and are also ranked according to their contribution to mortality risk. We evaluated SANMF against four empirical models on the task of predicting mortality or survival 30 days after discharge from ICU using the observed physiologic measurements between 12 and 24 hours after admission. SANMF outperforms all comparison models, and in particular, demonstrates an improvement in AUC (0.848 vs. 0.827, $p<0.002$) compared to a state-of-the-art machine learning method that uses manual feature engineering. Feature analysis was performed to illuminate insights and benefits of subgraph groups in mortality risk prediction.

## 1. Introduction

Mortality risk prediction and early recognition of clinical trends can identify actionable items for improving patient survival (Buist et al. 2002; McNeill; Bryden 2013; Chan et al. 2010). This problem has particular importance in the Intensive Care Unit (ICU). Modern ICUs generate multivariate time series data for a patient using an increasing number of monitoring devices and laboratory tests. The close attention required from critical care providers exposes ICU patients to human errors known to be common in hospital admissions (Levinson; General 2010). Thus automated tools are needed to help clinicians interpret such data in a timely fashion, quickly assemble effective treatment plans, and ultimately improve patient outcome.

In the ICU, monitor sensitivity is often favored over specificity, thus alerts based on whether the value of a single parameter crosses a threshold may result in a prevalence of false alarms (McIntosh 2002). Better trade-off between sensitivity and specificity can be achieved by considering multiple variables comprehensively (Zong et al. 2004). The assumption is that more volatile patients display concerted abnormalities in multiple variables, which are associated with a high risk of mortality. Calibrated clinical models can be built by mining archived ICU physiologic time series.

Despite methodological advances in machine learning, clinicians often regard learnt models as black boxes. This flaw lies partially in the difficulty of translating complex clinical events to model features. For example, vital measurements and laboratory test values fluctuate as time progresses (e.g., glucose level may increase from 158 mg/dL to 189 mg/dL after 53 minutes, then fall to 172 mg/dL after another 62 minutes). We refer to these events as temporal trends. In contrast, the standalone numerical measurements (e.g., 158 mg/dL glucose level) are single time point snapshots. Snapshot measurements have been widely used due to their simple extraction and robust statistical properties. However, they are less informative and interpretable than temporal trends. Temporal trends are more expressive and informative, but their extraction is often cumbersome. For better modeling, temporal trends often need to be considered in groups because the underlying pathophysiologic evolution of a patient (e.g. kidney failure) usually manifests itself through multiple physiologic variables (e.g., abnormalities in glomerular filtration rate, creatinine, etc.).

## 2. Related Work

Previous work in predicting mortality risk based on ICU patients' physiological status generally falls into two categories. Score-based methods (e.g., $SAPS_{II}$ (Le Gall et al. 1993) and APACHE (Knaus et al. 1991)) assume a re-

source-limited ICU setting and aim to select a limited set of commonly measured clinical predictors that can (often manually) be aggregated into a severity score and best associated to a particular outcome. Others adopt a broader modeling perspective. Hug et al. (Hug; Szolovits 2009) considered a comprehensive set of physiologic measurements from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) clinical dataset (Saeed et al. 2011) and manually defined a set of trend patterns. Quinn et al. (Quinn et al. 2009) developed a factorial switching linear dynamic system to model the patient states underlying 8 physiologic measurements. Cohen et al. (Cohen et al. 2010) used hierarchical clustering to extract 10 clusters as clinically relevant patient states from physiologic measurements, over a set of 17 patients and 14 measurements. However, these methods either model temporal trends independently without capturing correlated trends manifesting underlying pathophysiologic states, or lack scalability to many more physiologic variables. Recently, the 2012 PhysioNet/Computing in Cardiology Challenge tackled the problem of in-hospital mortality prediction on MICMIC-II dataset with 36 physiologic time series from first 2 days of admission (Silva et al. 2012; Johnson et al. 2012; Lee et al. 2012; Citi; Barbieri 2012; Xia et al. 2012; McMillan et al. 2012; Vairavan et al. 2012; Krajnak et al. 2012; Severeyn et al. 2012; Pollard et al. 2012; Hamilton; Hamilton 2012; Bera; Nayak 2012). The challenge used the minimum of precision and recall (score1) as official evaluation metrics with highest score1 around 0.53. Several systems also reported AUC scores from 0.82 (e.g., (McMillan et al. 2012), score1=0.46) to 0.86 (e.g., (Johnson et al. 2012), score1=0.53). Joshi et al. (Joshi; Szolovits 2012) extended to 54 physiologic time series to predict 30-day mortality for ICU patients. They manually clustered the physiologic measurements into organ specific patient states by associating each measurement with the status of a particular organ, and achieved a state-of-the-art 30-day mortality AUC of 0.91 on the MIMIC-II dataset. Despite improving scalability, their manual feature clustering applies to only snapshot measurements and can be a subjective call. For example, low hematocrit may be linked to blood loss, bone marrow problems or kidney problems, thus is hard to assign to a specific organ. Addressing these challenges, we study how to group temporal trends instead of single time point measurements, and how evidence-based grouping can be performed over a comprehensive set of physiologic variables. Our representation of temporal trends falls into the category of time series abstraction that discretizes time series into sequences of symbols and attaches meaning to the symbols (Lin et al. 2007; Dagliati et al. 2014; Combi et al. 2012; Moskovitch; Shahar 2014; Sacchi et al. 2015). We chose our discretization approach to be consistent with the state-of-the-art comparison model (Joshi; Szolovits 2012), which showed that a customized z-score was an effective discretization on MIMIC-II data. Our approach also handles time series with irregularly sampled time points, though we expect augmenting it with some of the above abstraction methods may lead to further improvements and leave this as future work.

# 3. Methods

In this work, we develop an unsupervised feature learning method in order to build machine learning models that are both more accurate and more interpretable to clinicians. The model applies non-negative matrix factorization to discover groups of subgraph-encoded temporal progression trends, hence the name Subgraph Augmented Non-negative Matrix Factorization (SANMF).

## 3.1 Workflow of SANMF

We first outline the workflow of the SANMF algorithm in *Fig 1*. ICU physiologic time series are first converted to graph representations. The graph representation is derived by discretizing time and measurement axes for physiologic measurements. We use a frequent subgraph mining tool to collect important subgraphs where the subgraphs are identified as common temporal trends of the physiologic variables. With such representations, subgraphs encode temporal trends, and we use "subgraphs" and "temporal trends" interchangeably within the context of this paper. We model the correlation between the subgraphs, and apply non-negative matrix factorization to discover groups of subgraphs and patients, and then train a logistic regression model to predict the mortality risk using subgraph groups as features. We next explain each step in more detail.

## 3.2 Representing Time Series as Graphs

In representing the time series as graphs, we focus on the data from the second half of the first day after patients' admissions to ICU. We exclude the first half of the first day because many measurements are not yet available in that time period. The time series of different variables are often sparse and irregularly sampled, and contaminated by a variety of noise and human error. Thus before converting time series into graphs, we perform discretization on both the time axis and the measurement axis. We discretize the time axis by linearly interpolating the time series and resampling at regularly spaced time intervals. We expect more advanced imputation algorithms such as EM or Gaussian processes inference may lead to better performance at the expense of more parameter tuning and leave these to future work. We determined empirically (by 5-fold cross-validation over choices of 1, 2, 4, or 6 hour intervals) that a two-hour time interval was best in our experiment. With the interpolated time series, we compute a customized z-score (z'-score) where we define measurements

within the reference range of a certain test to be 0. For a physiologic variable $x$, let $x_l$ and $x_h$ be the low and high ends of the reference range, let $j$ index different ICU patient stays, and $\mu(x)$ and $\sigma(x)$ be the mean and standard deviation of variable $x$ across different ICU patient stays, the z'-score is calculated using the following equations

$$z(x_j) = \left(x_j - \mu(x)\right)/\sigma(x) \quad (1)$$

$$z'(x_j) = \begin{cases} 0 & if\ z(x_l) < z(x_j) < z(x_h) \\ z(x_j) - z(x_h) & if\ z(x_j) > z(x_h) \\ z(x_j) - z(x_l) & if\ z(x_j) < z(x_l) \end{cases} \quad (2)$$

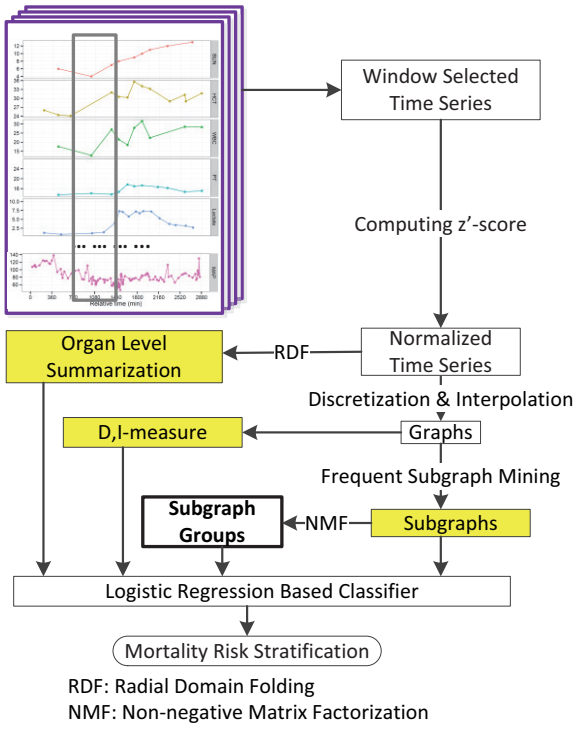



*Fig 1 Workflow of SANMF. We focus on the physiologic time series from the second half of the first day in ICU admission, balancing the trade-off between early detection of clinical trends and data availability. In the flow chart, shaded blocks indicate comparison models, see section 2 for organ level summarization, section 3.2 for D,I-measure, section 3.3 for subgraphs. The block with bold fonts corresponds to the features output by SANMF.*

Interpolating-then-normalizing aims to address the sampling bias that normal physiologic variables are sampled less frequently (email correspondence with MIMIC-II curators). Admittedly, interpolations are not real measurements, but they are reasonable estimates to counterbalance sampling bias, as ICU monitoring should catch most sudden/drastic changes. Each individual measurement is then discretized based on whether its value is within the reference range (label 0), within one $\sigma$ outside the reference range (label $\pm 1$), or beyond one $\sigma$ outside the reference range (label $\pm$ 2). Such discretization is essentially a thresholded round-up from equation ( 2 ). After discretization, we generate the time series graph for each measurement by connecting the discretized measurement values that are adjacent on the time axis. We use three types of edges to distinguish changes between adjacent nodes, namely *up*, *down* and *same*, and to encode partial directionality in temporal progression.

### 3.3 Frequent Subgraph Mining

Frequent subgraph mining (FSM) is a technique for pattern mining and has seen applications in natural language processing (Luo et al. 2014; Luo et al. 2015; Liu et al. 2013). In this work, we adapt FSM to produce the temporal trends that are common in the dataset. Intuitively, similar patients undergo similar physiologic trajectories during their ICU stays. Compared to temporal pattern mining (e,g, motif mining in (McMillan et al. 2012) and temporal variation summarization in (Hug; Szolovits 2009)), FSM supports flexible sequence/graph size and frequency based selection. FSM is defined on the notion of graph subisomorphism. We say one graph $G_s$ is subisomorphic to another graph $G$ if all its nodes $V_s$ and edges $E_s$ match with part of the other one. A subgraph occurs once in a corpus whenever it is subisomorphic to a graph in that corpus. Frequent subgraph mining identifies those subgraphs whose occurrences in a corpus are above a given threshold. In this work, we use the frequent subgraph miner MoSS (Borgelt; Berthold 2002) with frequency threshold empirically chosen (by 5-fold cross-validation on choices including 5, 10 or 15) to be 10 (i.e., subgraphs must occur at least 10 times in the dataset). Example frequent subgraphs are shown in Fig 3. The frequent subgraph mining step takes 1 min on a 2.7GHz CPU 8GB RAM laptop and generates 5534 frequent subgraphs. Among them, smaller subgraphs may be subisomorphic to other larger frequent subgraphs. When a larger subgraph is frequent; all of its subgraphs are necessarily also frequent. Furthermore, if a patient case has a larger subgraph, then both the larger and smaller subgraphs are counted for that patient. This may cause the signal from larger subgraphs to be overwhelmed by the signal from many smaller subgraphs. Therefore, we kept only the larger subgraphs in such pairs *when a patient case has both*. Note that such filtering is different from the notion of mining maximal frequent subgraphs, where only subgraphs that are not a part of any other frequent subgraphs at all are collected (Huan et al. 2004). It is cost prohibitive to perform a full pairwise check because the subisomorphism comparison between two subgraphs is already NP-complete (Nijssen; Kok 2005), and a pairwise approach would ask for over 15 million such comparisons for our task. However, in our case, we only need to compare subgraph pairs from the same physiologic variable. Furthermore, subgraph subisomorphism comparison can be simplified into string matching because our subgraphs are essentially sequences of maximum length 6. Combining the two observations, the algorithm for determining the

subisomorphism relation among frequent subgraphs is shown in Fig 2. The above filtering step takes 0.5 min and in fact does exclude some small subgraphs completely, leaving the final number of subgraphs at 5387.

```
Subisomorphism for set of subgraphs
input:  S - set of subgraphs
output: m - adjacency matrix of subisomorphism
among subgraphs in S
1  categorize subgraphs in S by variables
2  foreach v in variables:
3    stable sort S_v in ascending order of #nodes

4    for i = 1 to length(S_v)-1
5      for j = i+1 to length(S_v)
6        ids = S_v[i] // index of smaller subgraph
7        idb = S_v[j] // index of bigger subgraph
8        if subStringMatch(S[ids], S[idb])
9          m[ids, idb] = 1
10 return m
```

*Fig 2 Algorithm for determining subisomorphism relation among time series subgraphs. The simplification mainly comes from variable partition (line 1-2) and reduction of subisomorphism to substring match (line 8) for time series subgraphs.*

### 3.4 Subgraph Augmented NMF

Non-negative Matrix Factorization (NMF) has been a highly effective unsupervised method to cluster similar patients (Hofree et al. 2013) and sample cell lines (Müller et al. 2008), and to identify subtypes of diseases (Collisson et al. 2011). We explore a novel application of NMF based on the observation that a patient's underlying pathophysiologic evolution usually manifests itself in a group of temporal progression trends concerning multiple physiologic variables. This motivates us to use NMF to group time series subgraphs by factorizing the patient-by-subgraph count matrix, hence the name Subgraph Augmented NMF (SANMF). A schematic view of SANMF is shown in Fig 3. Let $M$ be the patient-by-subgraph count matrix of dimension $P \times S$, where $P$ is the number of patients and $S$ is the number of subgraphs. NMF approximates $M$ using two lower ranked matrices $U$ (of dimension $P \times S_g$ where $S_g$ is the number of subgraph groups) and $V$ (of dimension $S_g \times S$), as formalized in the following equation.

$$\min_{U,V} \|M - UV\|_F^2 \quad st.\ U \geq 0, V \geq 0 \qquad (3)$$

where $\|\cdot\|_F^2$ indicates squared Frobenius norm (squared summation of all entries in a matrix) and $U \geq 0$ means that $U$ is entry-wise non-negative. Intuitively, each row of $V$ gives the composition of each subgraph group, each column of $U$ reveals how each patient may be viewed as having a mixture of subgraph groups (approximating pathophysiologic evolutions).

As we are focusing on count data that is by definition nonnegative, we use NMF instead of other grouping methods such as k-means or principal component analysis (PCA) that do not have a built-in nonnegative constraint. The NMF solver we used is the projected gradient NMF (Lin 2007) implemented in Scikit-learn (Pedregosa et al. 2011). We use nonnegative double singular value decomposition as a deterministic initialization method (Boutsidis; Gallopoulos 2008). We also enforce sparsity on subgraph groups (Hoyer 2004) so that a group has only a limited number of non-zero weighted subgraphs and places most weight on only a few subgraphs, which is easier to interpret for clinicians.

### 3.5 Feature Group Discovery Using SANMF

In SANMF, the column vectors in the subgraph factor matrix $V$ specify the grouping of subgraphs. Such groupings can be viewed as mixtures of subgraphs, as they allow sharing of a subgraph among different groups as specified by its fractional weights across groups. In Fig 3 two example subgraph groups are shown. The top ranked subgraphs in subgraph group 1 indicate a general progression to an improved state. The top ranked subgraphs in subgraph group 2 indicate a general progression to a worse state. The motivation is to identify some subgraph groups that can indicate concerted progression patterns of physiologic variables as driven by the patient's underlying pathophysiologic evolution. The subgraph groups as specified in $V$ are used as features in logistic regression with the instance-feature matrix being $U$. Using the trained regression model, we rank the subgraph groups by their regression coefficients and focus on the top subgraph groups that are associated with high mortality risk.

### 3.6 Evaluating the Groups Discovered by SANMF

Because there is no innate way to determine whether the groupings of subgraphs discovered by SANMF are good or poor, we evaluate their utility as features, abstracted from the base data, in a prediction model. We assume that good features will improve prediction and will give us some insights into which temporal trend patterns are indicative of patient mortality risk. We use physiologic time series from the MIMIC-II database (Saeed et al. 2011). The time series include laboratory test values and physiologic measurements captured from patients monitored in the ICU at Beth-Israel Deaconess Medical Center (BIDMC) (see appendix A-Table 1 for a list of variables and their interpretation). Our dataset is a subset of the one used by Joshi et al. (Joshi; Szolovits 2012) (patients from the year 2000 to 2008); we only include those patients who have at least one day length of time series data. We predict whether a patient survives or dies in the ICU or within 30 days after ICU discharge, from data available for each patient during the period between 12 and 24 hours after their admission to the ICU. Choosing the 30-day mortality instead of in-hospital mortality emphasizes our motivation to detect clinical trends early on. We partitioned the cases equally, stratified by mortality, into a training set (3932 cases total) and a test set (3931 cases total), as shown in Table 1.

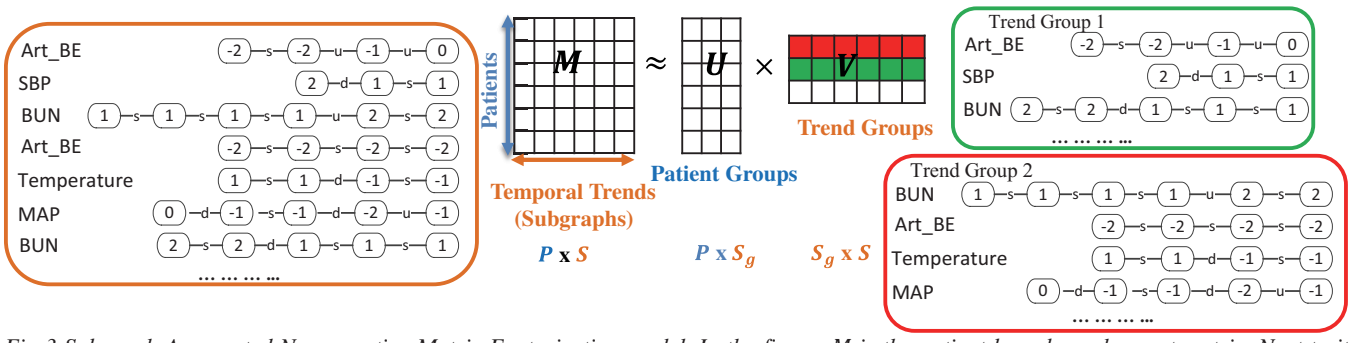



*Fig 3 Subgraph Augmented Non-negative Matrix Factorization model. In the figure, M is the patient-by-subgraph count matrix. Next to it are some example subgraph (trends). We also show example trend group 1 and trend group 2 after factorization. It is often desirable to have some trend groups indicate a general progression to the better states (e.g., trend group 1), or to the worse states (e.g., trend group 2).*

| Patient ICU Stays | | | |
|---|---|---|---|
| **Mortality** | **Total Cases** | **Training Cases** | **Test Cases** |
| ≤ 30 days | 788 | 383 (9.7%) | 405 (10.3%) |
| > 30 days or alive | 7075 | 3549 (90.3%) | 3526 (89.7%) |

*Table 1. Statistics of experiment data. The table includes the ICU patients' 30-day mortality distribution. The dataset is split equally into a training set and a test set.*

To evaluate the effectiveness of SANMF in abstracting raw data into more predictive features, we use 5-fold cross-validation on the training set to choose the number of subgraph groups, and use these subgraph groups as independent features to train a logistic regression model. We chose logistic regression over alternatives such as SVM or random forests for its capability to generate deterministic weights for individual features. As we have already grouped the subgraphs by accounting for their correlations, logistic regression provides a convenient way to directly assess subgraph group contribution. We then evaluate the model on the held-out test cases, and compare its performance against the following models: (a) as a baseline, 30-day mortality prediction by a logistic regression model using an approximation of the $SAPS_{II}$ score and its logarithm as predictors, where the $SAPS_{II}$ variable "chronic diseases" is approximated using ICD9 codes and the variable "type of admission" is approximated using the ICU service type (Hug; Szolovits 2009); (b) a state-of-the-art organ-level summarization model (Joshi; Szolovits 2012) for which we obtained their code and adapted it to account for our use of 12 hours of data rather than snapshot measurements by replacing the binary representation of whether an organ system is in a specific state with the number of times it is in that state during the 12 hours; (c) the D,I-measure based on our discretized (D) and interpolated (I) data values, where we also count the number of times each physiologic variable took on a discretized value during the 12 hours; and (d) a model based on treating subgraphs as independent features. The comparison models (b-d) are shaded in Fig 1. We compare the Area Under the ROC Curve (AUC) of our model against these four others.

# 4. Results

## 4.1 Performance on ICU Mortality Prediction

When using NMF to identify latent groups of temporal trend features, the number of groups needs to be empirically determined. We set this parameter to 100 by 5-fold cross-validation on the training data and considered a range of groups between 10 and 120 (at increments of 10), see appendix A-Fig 1 for detail. NMF with 100 groups takes 12 min in total for training and test data on a 2.7GHz CPU 8GB RAM laptop.

The AUC performance results of SANMF, comparison models and the baseline on held-out test data are shown in Fig 4. Comparing all the models and baseline, we can see that $SAPS_{II}$ approximation has an AUC of 0.673, which is lower than what is generally reported for $SAPS_{II}$ in the literature (Le Gall et al. 1993; Hug; Szolovits 2009; Joshi; Szolovits 2012) (We address this issue in detail in the Discussion). Nevertheless, all models that abstract the measured data by discretizing and aggregating them perform better by a large margin, each with an AUC greater than 0.8. The predictive model based on our SANMF-derived trend groups has the best performance, with an AUC of 0.848, significantly better ($p<0.002$ by random permutation test (Noreen 1989)) than the next-best model (AUC=0.827) based on abstraction by organ-system.

## 4.2 Important Subgraph Groups

Using the method in the feature group discovery section, we identified the top two subgraph groups that are associated with high mortality risk as listed in Table 2. These subgraph groups typically contain physiologic trends that stay at constant discretized values or sometimes progress to more severe states. In addition, they collectively indicate problematic pathophysiologic processes that involve one organ or multiple organs simultaneously, while still retaining the temporal trend details at the physiologic variable level.

For example, the first associated subgraph group has

several subgraphs suggesting that the patient mainly has pulmonary problems (continuously low minute ventilation, high plateau pressure, fluctuating airway resistance, and high level of positive end-expiratory pressure set on the ventilator). On the other hand, this group also has Glasgow Coma Scale staying very low, meaning that the patient is probably unconscious or sedated. Thus this group may be interpreted as unconscious or sedated patients with severe pulmonary problems. The second associated subgraph group displays abnormal trends related to problems in multiple organs including kidney, lung, and heart.

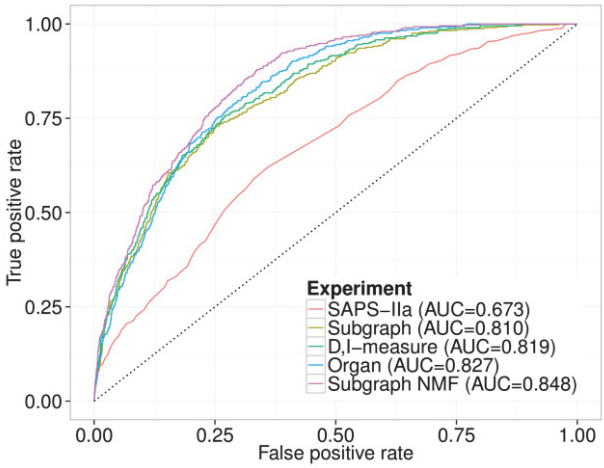



*Fig 4 ROC curves for proposed method SANMF, comparison models including subgraphs as independent features, discretized & interpolated measures (D,I-measure), and organ level status, as well as the baseline model using $SAPS_{II}$ approximation. SANMF significantly outperforms the next-best model (AUC 0.848 vs 0.827, p<0.002).*

An interesting observation is that top ranked subgraph groups contributing to high mortality risk usually involve problems in multiple organs rather than a single organ, which is more likely to happen in real ICU settings. Such grouping is difficult to achieve using manual grouping according to only organ status as done by Joshi et al. (Joshi; Szolovits 2012) and is considered one of the benefits of using NMF to automatically group temporal progression trends in an evidence-driven fashion.

## 5. Limitation, Discussion and Future Work

We observe that the AUC for our approximation to $SAPS_{II}$ is lower than in previously reported studies (Le Gall et al. 1993; Hug; Szolovits 2009; Joshi; Szolovits 2012). We believe that this is due to the large amount of missing data in our data set and the approximations we make because our data do not include the exact parameters used in SAPSII. Moreover, we built $SAPS_{II}$ predictions on 12hr data and obtained a 30-day mortality AUC=0.67, whereas (Joshi; Szolovits 2012) used all the data and reported an AUC=0.77 by $SAPS_{II}$. Putting system performance into context, our system outperforms (AUC 0.848 vs. 0.827) the system by (Joshi; Szolovits 2012) adapted to use 12hr data (Fig 4 organ model). Their AUC drop from 0.91 to 0.827 is consistent with using less data (totality vs. 12hr). In addition, both 0.848 and 0.827 30-day mortality AUCs are inferior to those for in-hospital mortality prediction (e.g., AUC=0.86, (Johnson et al. 2012) from PhysioNet/CinC2012), highlighting that 30-day mortality prediction is a more difficult task. Note that we combined perspectives from previous studies by reducing 48hr data in Physionet/CinC2012 to 12hr and predicting 30-day mortality. Although both extensions increase task difficulty, they offer potential clinical value as observed separately by previous studies.

| | | |
|---|---|---|
| **30-day Mortality 1st Subgraph Group** | | |
| 0.1000 | Glasgow Coma Scale | -2 -2 -2 -2 -2 -2 |
| 0.0085 | Minute Ventilation | -2 -2 -2 -2 -2 -2 |
| 0.0082 | Minute Ventilation | -1 -1 -1 -1 -1 -1 |
| 0.0081 | PEEPSet | 2 2 2 2 2 2 |
| 0.0066 | Airway Resistance | 1 0 |
| 0.0060 | Airway Resistance | 0 1 1 |
| 0.0059 | Plateau Pressure | 2 2 2 2 2 2 |
| 0.0052 | PEEPSet | 1 1 1 1 1 1 |
| 0.0047 | PaO2/FiO2 | 0 2 |
| 0.0040 | Airway Resistance | 1 1 0 |
| **30-day Mortality 2nd Subgraph Group** | | |
| 0.1634 | BUN/Creatinine | 2 2 2 2 2 2 |
| 0.0481 | BUN | 2 2 2 2 2 2 |
| 0.0155 | Albumin | -2 -2 -2 -2 -2 -2 |
| 0.0040 | Arterial CO2 | 1 1 1 1 1 1 |
| 0.0040 | Heart Rate | 0 -1 |
| 0.0038 | Na | 2 2 2 2 2 2 |
| 0.0034 | Na | 1 1 1 1 1 |
| 0.0033 | Arterial CO2 | 2 2 2 2 2 |
| 0.0032 | Arterial Base Excess | 2 1 |
| 0.0029 | Delivered Tidal Volume | -1 -1 -1 0 |

*Table 2 Top trend groups associated with high mortality risks. Trends are converted into a sequence to save space. For each trend such as "0.1000 Glasgow Coma Scale -2 -2 -2 -2 -2 -2", 0.1000 is the membership coefficient, Glasgow Coma Scale is the measurement label, "-2 -2 -2 -2 -2 -2" is the trend (flat for this case). Abbreviations used in the table include: PEEPSet – positive end-expiratory pressure set on ventilator; PaO2 – arterial oxygen tension; FiO2 – fraction of inspired oxygen; BUN – blood urea nitrogen; Na – sodium level. Please refer to appendix A-Table 1 for the detailed description and interpretation of all the variables.*

In this work, we use 30-day mortality (including mortality in hospital or within 30 days after discharge) as an obtainable ground truth in order to demonstrate the efficacy of SANMF as an unsupervised feature learning algorithm. Similar methods may be applicable to improve not only mortality predictions but also predictions that indicate specific types of patient deterioration (e.g., anticipating hypotension, kidney injury, hepatic failure, sepsis) and identifying therapeutic opportunities (e.g., ability to wean from a ventilator, an intra-aortic balloon pump, vasopressors). Such improved models can provide decision support

for treatment planning and informed staffing. A systematic comparison of published systems, for both mortality and customized outcome prediction, along the directions of both FSM and temporal pattern methods (Hug; Szolovits 2009; McMillan et al. 2012), can be informative to the research community and is part of our future work.

In our experiment, FSM takes 1.5 min and NMF takes 12 min on a 2.7GHz CPU 8GB RAM laptop. NMF is most time consuming but necessary for better interpretability and accuracy, as demonstrated by Fig 4 and Table 2. Note that we only included physiologic time series, in order to make a fair comparison to the baseline model $SAPS_{II}$, and did not include treatment. However, ICU mortality risk can be better stratified by considering the interplay between observations and interventions. In addition, we want to further model trend-intervention and trend-trend relative changes such as "if temperature rose before the change in BUN". Capturing both aspects may require interconnecting trend and intervention sequences, thus generating proper graphs and subgraphs. The intention to pursue this direction motivated us to resort to the reasonably fast subgraph mining instead of sequence mining in the first place. We expect further improved accuracy and interpretability by predicting outcomes for patient groups who have similar underlying pathophysiologic evolutions and who have undergone similar treatment regimens. This step asks for better heuristics in FSM and our subgraph subisomorphism check, and perhaps more refined discretization and interpolation, but is a promising direction.

## 6. Conclusions

We proposed a novel unsupervised feature learning algorithm named Subgraph Augmented Non-negative Matrix Factorization (SANMF), designed for analyzing temporal progression patterns in clinical time series, and showed it to improve both the accuracy and the interpretability of the learnt model for ICU mortality risk prediction. In summary, subgraph mining on multivariate time series leads to unsupervised extraction of multivariate temporal trends, which are more informative than snapshot measurements. The ensuing NMF-based step groups correlated temporal trends of different physiologic variables. This leads to better interpretability and improved accuracy. We compared SANMF with four different models using features with varying granularities and time spans. SANMF outperforms all the comparison models and in particular demonstrates an AUC improvement from 0.827 to 0.848 ($p < 0.002$), compared to a similarly motivated state-of-the-art model on MIMIC-II dataset that explores manual feature engineering. A detailed feature analysis of the subgraph groups that are generated by SANMF offers more clinical insights about multiple organ problems associated with high mortality risk, all being automatically identified from the data.

## Acknowledgements

The work described was supported in part by Grant Number U54LM008748 from the National Library of Medicine. The intensive care data are from a data set distributed under a limited data use agreement, which was approved by the Beth Israel Deaconess Hospital's IRB.

## References

Bera, D.; Nayak, M.M., 2012. Mortality risk assessment for ICU patients using logistic regression. In Computing in Cardiology (CinC). 493–496, IEEE.

Borgelt, C.; Berthold, M.R., 2002. Mining molecular fragments: Finding relevant substructures of molecules. In Proceedings. 2002 IEEE International Conference on Data Mining. 51–58, IEEE.

Boutsidis, C.; Gallopoulos, E., 2008. SVD based initialization: A head start for nonnegative matrix factorization. Pattern Recognition 41(4): 1350–1362.

Buist, M.D. et al., 2002. Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: preliminary study. BMJ 324(7334): 387–390.

Chan, P.S. et al., 2010. Rapid response teams: a systematic review and meta-analysis. Archives of internal medicine 170(1): 18–26.

Citi, L.; Barbieri, R., 2012. PhysioNet 2012 Challenge: Predicting mortality of ICU patients using a cascaded SVM-GLM paradigm. In Computing in Cardiology (CinC). 257–260, IEEE.

Cohen, M.J. et al., 2010. Research Identification of complex metabolic states in critically injured patients using bioinformatic cluster analysis. Critical Care 14(1): R10.

Collisson, E.A. et al., 2011. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nature medicine 17(4): 500–503.

Combi, C.; Pozzi, G.; Rossato, R., 2012. Querying temporal clinical databases on granular trends. Journal of biomedical informatics 45(2): 273–291.

Dagliati, A. et al., 2014. Temporal data mining and process mining techniques to identify cardiovascular risk-associated clinical pathways in Type 2 diabetes patients. In Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on. 240–243, IEEE.

Le Gall, J.-R.; Lemeshow, S.; Saulnier, F., 1993. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA: the journal of the American Medical Association 270(24): 2957–2963.

Hamilton, S.L.; Hamilton, J.R., 2012. Predicting in-hospital-death and mortality percentage using logistic regression. In Computing in Cardiology (CinC). 489–492, IEEE.

Hofree, M. et al., 2013. Network-based stratification of tumor mutations. Nature methods 10(11): 1108–1115.

Hoyer, P.O., 2004. Non-negative matrix factorization with sparseness constraints. The Journal of Machine Learning Research 5: 1457–1469.

Huan, J. et al., 2004. Spin: mining maximal frequent subgraphs from graph databases. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 581–586, ACM.

Hug, C.W.; Szolovits, P., 2009. ICU acuity: real-time models versus daily models. In AMIA Annual Symposium Proceedings. 260–264, American Medical Informatics Association.

Johnson, A.E. et al., 2012. Patient specific predictions in the intensive care unit using a Bayesian ensemble. In Computing in Cardiology (CinC). 249–252, IEEE.

Joshi, R.; Szolovits, P., 2012. Prognostic Physiology: Modeling Patient Severity in Intensive Care Units Using Radial Domain Folding. In AMIA Annual Symposium Proceedings. 1276–1283, American Medical Informatics Association.

Knaus, W.A. et al., 1991. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. CHEST Journal 100(6): 1619–1636.

Krajnak, M. et al., 2012. Combining machine learning and clinical rules to build an algorithm for predicting ICU mortality risk. In Computing in Cardiology (CinC). 401–404, IEEE.

Lee, C.H. et al., 2012. An imputation-enhanced algorithm for ICU mortality prediction. In Computing in Cardiology (CinC). 253–256, IEEE.

Levinson, D.R.; General, I., 2010. Adverse events in hospitals: national incidence among Medicare beneficiaries. Department of Health and Human Services Office of the Inspector General.

Lin, C.-J., 2007. Projected gradient methods for nonnegative matrix factorization. Neural computation 19(10): 2756–2779.

Lin, J. et al., 2007. Experiencing SAX: a novel symbolic representation of time series. Data Mining and knowledge discovery 15(2): 107–144.

Liu, H. et al., 2013. Approximate Subgraph Matching-Based Literature Mining for Biomedical Events and Relations. PloS one 8(4): e60954.

Luo, Y. et al., 2014. Automatic Lymphoma Classification with Sentence Subgraph Mining from Pathology Reports. Journal of the American Medical Informatics Association (JAMIA) 2014 21(5): 824–832.

Luo, Y. et al., 2015. Subgraph Augmented Non-Negative Tensor Factorization (SANTF) for Modeling Clinical Text. Journal of the American Medical Informatics Association ocv016.

McIntosh, N., 2002. Intensive care monitoring: past, present and future. Clinical medicine 2(4): 349–355.

McMillan, S. et al., 2012. ICU mortality prediction using time series motifs. In Computing in Cardiology (CinC). 265–268, IEEE.

McNeill, G.; Bryden, D., 2013. Do either early warning systems or emergency response teams improve hospital patient survival? A systematic review. Resuscitation 84(12): 1652–1667.

Moskovitch, R.; Shahar, Y., 2014. Classification of multivariate time series via temporal abstraction and time intervals mining. Knowledge and Information Systems 1–40.

Müller, F.-J. et al., 2008. Regulatory networks define phenotypic classes of human stem cell lines. Nature 455(7211): 401–405.

Nijssen, S.; Kok, J.N., 2005. The gaston tool for frequent subgraph mining. Electronic Notes in Theoretical Computer Science 127(1): 77–87.

Noreen, E.W., 1989. Computer-intensive methods for testing hypotheses: an introduction, Wiley.

Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 12: 2825–2830.

Pollard, T.J. et al., 2012. 2012 PhysioNet Challenge: An artificial neural network to predict mortality in ICU patients and application of solar physics analysis methods. In Computing in Cardiology (CinC). 485–488, IEEE.

Quinn, J.A.; Williams, C.K.; McIntosh, N., 2009. Factorial switching linear dynamical systems applied to physiological condition monitoring. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(9): 1537–1551.

Sacchi, L.; Dagliati, A.; Bellazzi, R., 2015. Analyzing Complex Patients' Temporal Histories: New Frontiers in Temporal Data Mining. In Data Mining in Clinical Medicine. 89–105, Springer.

Saeed, M. et al., 2011. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. Critical care medicine 39(5): 952–960.

Severeyn, E. et al., 2012. Towards the prediction of mortality in Intensive Care Units patients: A Simple Correspondence Analysis approach. In Computing in Cardiology (CinC). 469–472, IEEE.

Silva, I. et al., 2012. Predicting in-hospital mortality of ICU patients: The physioNet/computing in cardiology challenge 2012. In Computing in Cardiology (CinC). 245–248, IEEE.

Vairavan, S. et al., 2012. Prediction of mortality in an intensive care unit using logistic regression and a hidden Markov model. In Computing in Cardiology (CinC). 393–396, IEEE.

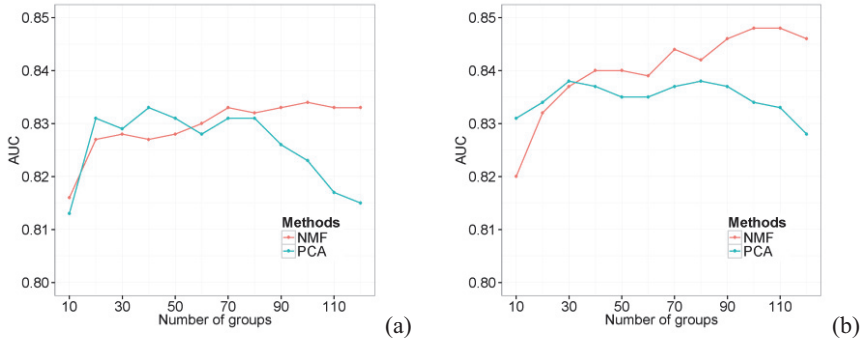
Xia, H. et al., 2012. A neural network model for mortality prediction in ICU. In Computing in Cardiology (CinC). 261–264, IEEE.

Zong, W.; Moody, G.; Mark, R., 2004. Reduction of false arterial blood pressure alarms using signal quality assessement and relationships between the electrocardiogram and arterial blood pressure. Medical and Biological Engineering and Computing 42(5): 698–706.

# Appendix

| Variable | Description | Variable | Description |
|---|---|---|---|
| Age | Age of the patient upon admission | Hemoglobin | Hemoglobin level |
| Airway Resistance | The resistance of the respiratory tract to airflow during inspiration and expiration. | INR | Prothrombin time international normalized ratio |
| Albumin | Albumin in blood | Ion Calcium | Ion Calcium level |
| ALT | Alanine aminotransferase in blood | K | Potassium level |
| Arterial Base Excess | Excess in the amount of base present in arterial blood | Lactate | Lactate level |
| Arterial CO2 | Arterial carbon dioxide | MAP | Mean arterial pressure |
| Arterial PaCO2 | Arterial carbon dioxide tension | Mg | Magnesium level |
| Arterial PaO2 | Arterial oxygen tension | Minute Ventilation | Volume of gas exchanged from lung per minute |
| Arterial pH | pH level in arterial blood | Na | Sodium level |
| AST | Aspartate aminotransferase in blood | PaO2/FiO2 | Partial pressure arterial oxygen / Fraction of inspired oxygen |
| AST/ALT | Aspartate aminotransferase / alanine aminotransferase | PTT | Partial Thromboplastin Time |
| BUN | Blood urea nitrogen | PEEPSet | Positive end-expiratory pressure set on ventilator |
| BUN/Creatinine | Blood urea nitrogen / Creatinine | PIP | Peak inspiratory pressure |
| Ca | Calcium level | Plateau Pressure | Pressure applied (in positive pressure ventilation) to the small airways and alveoli |
| Cardiac Index | Relates the cardiac output (CO) from left ventricle in one minute to body surface area | Platelets | Platelets count |
| Central Venous Pressure | Blood pressure in the thoracic vena cava | Prothrombin Time | Time it takes for plasma to clot |
| Cl | Chloride level | RBC | Red blood count |
| Creatinine | Level of creatinine in the blood | Respiratory Rate | Respiratory rate per minute |
| Delivered Tidal Volume | Air volume of lung without extra effort | RSBI | Rapid shallow breathing index* |
| Diastolic blood pressure | Minimum blood pressure during heartbeat | RSBI Rate | Rapid shallow breathing index rate change |
| Direct bilirubin | Level of bilirubin conjugated with glucuronic acid | SaO2 | Saturation of arterial oxygen |
| eGFR | Estimated glomerular filtration rate | Systolic blood pressure | Maximum blood pressure during heartbeat |
| FiO2Set | Fraction of inspired oxygen set on ventilator | Temperature | Body temperature |
| GCS | Glasgow coma scale | Total Bilirubin | Level of bilirubin |
| Glucose | Glucose level | tProtein | Total protein in the blood plasma |
| Heart Rate | Heart rate per minute | Urine/Hour/Weight | Urine per hour per kg body weight |
| Hematocrit | Hematocrit level | WBC | White blood count |

A-Table 1 Physiologic time series predictor variables from MIMIC II dataset. Demographic information such as age is also included.




A-Fig 1 Comparing NMF and PCA under different number of subgraph groups. (a) 5-fold cross-validation AUC. (b) the held-out test AUC. Show in panel (a) for corresponding number of groups is a single AUC by merging all the responses from the 5 validation subsets.