DeepAlerts: Deep Learning Based Multi-Horizon Alerts for Clinical Deterioration on Oncology Hospital Wards

Dingwen Li,[1] **Patrick G Lyons,**[2] **Chenyang Lu,**[1] **Marin Kollef**[2]
[1]Washington University in St. Louis, McKelvey School of Engineering
[2]Washington University in St. Louis, School of Medicine
{dingwenli, plyons, lu, kollefm}@wustl.edu

Abstract

Machine learning and data mining techniques are increasingly being applied to electronic health record (EHR) data to discover underlying patterns and make predictions for clinical use. For instance, these data may be evaluated to predict clinical deterioration events such as cardiopulmonary arrest or escalation of care to the intensive care unit (ICU). In clinical practice, early warning systems with multiple time horizons could indicate different levels of urgency, allowing clinicians to make decisions regarding triage, testing, and interventions for patients at risk of poor outcomes. These different horizon alerts are related and have intrinsic dependencies, which elicit multi-task learning. In this paper, we investigate approaches to properly train deep multi-task models for predicting clinical deterioration events via generating multi-horizon alerts for hospitalized patients outside the ICU, with particular application to oncology patients. Prior knowledge is used as a regularization to exploit the positive effects from the task relatedness. Simultaneously, we propose task-specific loss balancing to reduce the negative effects when optimizing the joint loss function of deep multi-task models. In addition, we demonstrate the effectiveness of the feature-generating techniques from prediction outcome interpretation. To evaluate the model performance of predicting multi-horizon deterioration alerts in a real world scenario, we apply our approaches to the EHR data from 20,700 hospitalizations of adult oncology patients. These patients' baseline high-risk status provides a unique opportunity: the application of an accurate model to an enriched population could produce improved positive predictive value and reduce false positive alerts. With our dataset, the model applying all proposed learning techniques achieves the best performance compared with common models previously developed for clinical deterioration warning.

Introduction

Hospital inpatients are at high risk for clinical instability (Escobar and dellinger 2016): about 4% of ward encounters involve transfer to the intensive care unit (ICU) or death (Kipnis et al. 2016; Churpek et al. 2014). A recent study found that this risk is over 9% among oncology inpatients, who also have unique risk factors, such as cancer type and treatment complications (Lyons et al. 2019). Since deterioration is often presaged by abnormal vital signs or test results hours before clinical decompensation (Churpek et al. 2012), which is often unrecognized (J R de Bie et al. 2019), electronic health record (EHR)-based risk prediction models have been designed to identify these patients during windows of potentially preventable instability, facilitating early interventions to prevent or mitigate adverse events (Churpek et al. 2014; Subbe et al. 2001). However, despite the promise of these early warning systems (EWS), they have not consistently improved patient outcomes (Bedoya et al. 2019), which may be related to limitations in predictive accuracy, discrimination, model calibration and the time horizon of the impending clinical deterioration event.

Machine learning methods have been more widely applied to solve clinical event prediction problems with the advent of massive EHR data. Deep learning, in particular, has the potential to achieve superior predicting accuracy for large datasets. In our study of predicting clinical deterioration for oncology patients, the multi-horizon alerts (i.e., 6 hours, 24 hours, and 48 hours prior to an event) provide different levels of urgency to allow clinicians to better plan interventions in advance. Based on the fact that the multi-horizon alerts are all predicting the same type of deterioration events across different (but related) time horizons, the predictive models for the alerts can have similar structure and intrinsic relatedness. This motivates us to effectively learn the multi-horizon alerts model simultaneously via multi-task learning. Multi-task learning utilizes the task relatedness to improve the prediction accuracy by jointly learning a shared model for multiple prediction tasks. For example, multi-task learning and its variants are used to predict clinical outcomes from multiple diseases (Nori et al. 2015). They are also widely used to discover the characteristic patterns of clinical physiology data, namely clinical phenotyping (Liu et al. 2015; Che et al. 2015). While deep multi-task models have thousands or even millions of parameters to learn, the models need to be trained properly to simultaneously achieve optimal performance for all the tasks. In this paper, we explore two technical approaches based on a deep multi-task learning framework that can utilize the tasks' intrinsic characteristics to accurately predict multi-horizon alerts for clinical deterioration across different time horizons. Prior knowledge regularization is adopted to our deep multi-task framework, which exploits the positive effects of task relatedness in the multi-task learning framework. Moreover, we propose a novel task-specific loss balancing technique, which reduces the negative effects of the imbalanced task-specific losses when optimizing the joint objective function.

We evaluate the proposed approaches on the adult oncology patients hospitalized at Barnes-Jewish Hospital. Given that a major drawback of many existing EWS (designed in other populations) involves false positive alerts due to low positive predictive value (PPV), our outcome-rich population provides a high-prevalence setting which could improve PPV significantly when paired with a well-performing model. The task is to learn multi-horizon alerts that can predict patients' deterioration as a composite event of ward death or ICU transfer. The evaluations compare the proposed approaches to (1) the Modified Early Warning Score (MEWS) that is implemented on many general inpatient wards (but which is known to perform poorly among oncology inpatients (Cooksley, Kitlowski, and Haji-Michael 2012)), (2) state-of-the-art machine learning models, such as (a) penalized logistic regression, (b) random forest, and (c) gradient boosted tree, and (3) deep learning models trained separately for each time horizon or jointly for all time horizons via multi-task learning. The prior knowledge regularization and task-specific loss balancing both exhibit their superior performance over the benchmark models. The performance is further improved when combining the two techniques together.

In this paper, we present to our knowledge the first multi-horizon alert system based on a deep multi-task learning framework. Specifically, the contributions of this work are four-fold: (1) we apply prior knowledge regularization to the deep multi-task learning framework and demonstrate its ability of improving prediction performance over the state-of-the-art models as well as generating valid multi-horizon alert sequences; (2) we propose the task-specific loss balancing to eliminate the negative effects brought by jointly training deep multi-task model; (3) we evaluate the proposed multi-horizon alerts model on a large and real oncology inpatient dataset. The significant predictive performance improvement and the clinical relevance of high-impact features demonstrate the feasibility of applying deep multi-task learning to accurately predict clinical deterioration; (4) by using relevance back-propagation, we show the interpretability of the prediction results and verify the significance of including second order time series features as the input to the predictive models.

Related Work

Deterioration Prediction on Wards

As EHR data become widely available, machine learning methods have been increasingly adopted to predict outcomes including clinical deterioration events. Machine learning on EHR data is challenging, since the data is heterogeneous and contains missing values. Previous works focus on applying logistic regression and its variants to predict unplanned ICU transfer (Wellner et al. 2017; Zhai et al. 2014; Churpek et al. 2016) or other deterioration events on the wards (Churpek et al. 2014; Bailey et al. 2013; Jeffery et al. 2018). Some of these studies report better discriminatory performance over simpler early warning models, such as MEWS (Churpek et al. 2016) and the VitalPAC$^{\text{TM}}$ Early Warning Score (Churpek et al. 2014) (ViEWS). Random forest approaches have also been applied (Churpek et al. 2016; Jeffery et al. 2018) and achieve better discrimination than logistic regression (Churpek et al. 2016) for various outcomes, including ICU transfers and ICU readmissions. Ensemble methods, such as adaptive boosting and gradient boosting, proposed in recent works (Desautels et al. 2017; Rubin et al. 2018), achieve better outcomes compared with earlier models. Recently, more attention has been drawn to deep models (Hu et al. 2016; Wellner et al. 2017; Lin et al. 2019). Deep models are able to discover complex underlying patterns from massive heterogeneous data, permitting inclusion of complex interaction effects, previously unconsidered patterns from metadata, and complicated temporal relationships. However, previous models either predict alerts without a specific time horizon, or predict events in a single time horizon. To our knowledge our work produces the first clinical warning model to generate alerts over multiple time horizons.

Multi-task Learning

For applications with tasks that are related to each other, multi-task models can achieve significant performance improvement over the single-task models. Graph Laplacian-based regularization incorporates any relational information in the prior knowledge as a weighted graph (Che et al. 2015), which can be applied as regularization term in the objective function. Multi-task learning has been applied to clinical phenotyping (Liu et al. 2015; Che et al. 2015) and prediction of multiple clinical events (Nori et al. 2015), but has not to our knowledge been applied towards development of an early warning system, which either uses single or undefined event horizons. Here, we exploit multi-task learning for multi-horizon alerts for clinical deterioration, a new problem that has not been addressed in clinical data mining.

A key challenge of applying multi-task models is that they are hard to train properly (Chen et al. 2018). Imbalanced task-specific losses in a multi-task model hinder the performance (Kendall, Gal, and Cipolla 2018). Previous literature proposes to use uncertainty as the loss weights for task-specific losses (Kendall, Gal, and Cipolla 2018), but it is computationally intensive to optimize the loss weights. GradNorm (Chen et al. 2018) proposes a new objective function for choosing the optimal loss weights at each training step. However, the numerical optimization for loss weights adds substantial computational burden to the learning process. Our proposed task-specific loss balancing has a closed-form analytic solution, which does not need additional numerical optimization for loss weights and can be easily integrated into the deep model training procedure.

Methodology

Multi-horizon Alerts Problem

The EHR data, including demographics, comorbidity diagnoses (via ICD-9 and ICD-10 codes from prior hospital admissions), patient location, and time-stamped vital signs, lab values, cultures, medications, and procedures, from a 6-hour time window prior the predicting horizon is used as input to the predictive models (a discrete-time framework to account for interval censoring of clinical data such as vital signs and laboratory measurements). The clinical deterioration event (a composite of death on the wards or transfer to the ICU) occurring at future time points (e.g. 6 hours, 24 hours and 48 hours after the end of an input window) is defined as the positive label in our multi-horizon alerts problem. It is important to note that the alerts with different horizons are related to each other. For instance, if a patient experienced an alert 24 hours prior to a true outcome, the patient would likely have remained at high risk (i.e., had a high predicted probability of deterioration) 18 hours later, at the 6-hour mark. Relatedly, it is unlikely for a 6-hour alert to be followed by no alert in the next 24 hours. Therefore, we employ a multi-task learning approach to exploit the relatedness of alerts of different time horizons. The precedent constraints between the alerts are encoded as co-occurrence matrix in the prior knowledge regularization. Furthermore, we incorporate this prior knowledge regularization into the learning process to improve the accuracy of our predictions.

In the following, we describe the details of the approaches to improve deep multi-task model for the mutli-horizon alerts. We first exploit the task relatedness via incorporating prior knowledge as graph Laplacian-based regularization (Che et al. 2015). Then, we introduce the approach of removing negative effects induced by task loss imbalance, and devise an auxiliary loss to balance the task-specific losses during training. The block coordinate optimization with a closed-form analytic solution is derived to optimize the joint loss function with prior knowledge regularization and auxiliary loss for task-specific loss balancing.

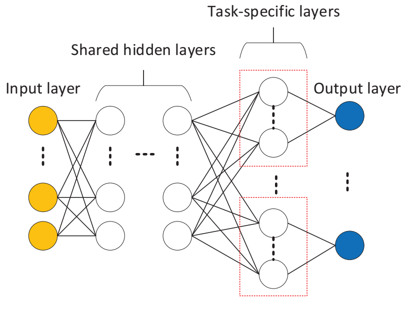

Figure 1: Deep multi-task model with hard parameter sharing

Exploiting Task Relatedness via Prior Knowledge

Suppose we have a dataset of N instances, each with a D-dimensional feature vector derived from EHRs and K binary labels for K different learning tasks. The labeled instances can be represented as $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$, where $\mathbf{x}^{(i)} \in \mathbb{R}^D$, $\mathbf{y}^{(i)} \in \{0,1\}^K$. The model used in our study is a feed-forward fully connected neural network with L hidden layers. We use $\mathbf{\Theta}_l = (\mathbf{W}_l, \mathbf{b}_l)$ to denote the neural network parameters in lth layer. As shown in Figure 1, the parameters of hidden layers $\{\mathbf{\Theta}_l\}_{l=1}^{L-1}$ from the first to the $(L-1)$th layer are shared by all the tasks to form a common feature representation, which is also known as hard parameter sharing. There are K separate Lth task-specific hidden layers with their own parameters $\mathbf{\Theta}_{L,k}$. The parameters in each task-specific layer are updated simultaneously while optimizing the joint objective function.

In addition, we add the non-linearity activation, Rectified Linear Unit (ReLU),

$$\mathbf{z}_l = \max(0, \mathbf{W}_l \mathbf{z}_{l-1} + \mathbf{b}_l) \tag{1}$$

to each hidden layer, where $\mathbf{z}_{l-1}$ denotes the activation outputs from the $(l-1)$th layer. The ReLU takes care the vanishing gradient issue when training the neural network and accelerates the convergence of gradient descent method. Since the clinical deterioration events in our study have binary outcomes (e.g. deteriorating within a specific time frame or not), the sigmoid function

$$\sigma(\mathbf{z}_{L,k}) = 1/(1+\exp(-\mathbf{z}_{L,k})) \tag{2}$$

is applied to the task-specific activation functions in the last layer. The cross entropy loss $\mathcal{L}_{CEk}$ is used for each task, which is referred as task-specific loss in the paper. The cross entropy loss for the whole multi-task model is the sum of task-specific cross entropy loss:

$$\mathcal{L}_{CE} = \sum_{k=1}^{K} \mathcal{L}_{CEk} = \sum_{k=1}^{K}\sum_{i=1}^{N} \Big[y_k^{(i)} \log \sigma(\mathbf{W}_{L,k}\mathbf{z}_{L-1}^{(i)} + \mathbf{b}_{L,k}) + (1-y_k^{(i)}) \log(1-\sigma(\mathbf{W}_{L,k}\mathbf{z}_{L-1}^{(i)} + \mathbf{b}_{L,k}))\Big] \tag{3}$$

where $y_k^{(i)}$ is the ground-truth label for the kth task of instance i, $\mathbf{z}_{L-1}^{(i)}$ is the vector of activation outputs from the $(L-1)$th layer.

For those with known relatedness among the multiple prediction tasks, the similarity of tasks can be utilized as the prior information to regularize the optimization objective. Graph Laplacian-based regularization is a natural way to incorporate cross-task relatedness via graph representation (Che et al. 2015). Let $\mathbf{S} \in \mathbb{R}^{K \times K}$ denote graph adjacency matrix encoding the similarity between the kth and k'th tasks. Then $\mathbf{L}$ is the Laplacian matrix of $\mathbf{S}$, such that $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{D}$ is a diagonal "degree" matrix of $\mathbf{S}$ whose diagonal elements are $\mathbf{D}_{k,k} = \sum_{k'} \mathbf{S}_{k,k'}$. Thus, the graph Laplacian-based regularization can be written as:

$$\mathcal{R}_{Lap} = \mathrm{tr}(\mathbf{\Theta}_L^T \mathbf{L} \mathbf{\Theta}_L) \tag{4}$$

$$\mathcal{R}_{Lap} = \frac{1}{2}\sum_{k=1}^{K}\sum_{k'=1}^{K} \mathbf{S}_{k,k'} \left\| \mathbf{\Theta}_{L,k} - \mathbf{\Theta}_{L,k'} \right\|_2^2 \tag{5}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace, $\mathbf{\Theta}_L = \{\mathbf{\Theta}_{L,k}\}_{k=1}^{K}$ is a vector of weights from all K task-specific layers. The intention is to regularize the model parameters in the task-specific layers according to the extent of similarity among tasks represented by $\mathbf{S}$. In (5), the element $\mathbf{S}_{k,k'}$ indicates the extent of similarity between task k and k'. The high value of $\mathbf{S}_{k,k'}$ will penalize $\|\boldsymbol{\Theta}_{L,k} - \boldsymbol{\Theta}_{L,k'}\|_2^2$ and thus enforce $\boldsymbol{\Theta}_{L,k}$ and $\boldsymbol{\Theta}_{L,k'}$ to be alike. There are various heuristics to determine the similarity matrix $\mathbf{S}$. However, we use the co-occurrence matrix (Che et al. 2015) throughout the study, which is defined by:

$$S_{k,k'} = \frac{1}{N}\sum_{i=1}^{N} I(y_k^{(i)} = y_{k'}^{(i)}) \tag{6}$$

$I(y_k^{(i)} = y_{k'}^{(i)})$ is an indicator function, which is evaluated to 1 if task k and task k' have the same label for instance i. The Frobenius norm is also applied as a regularization term, which prevents the model from overfitting by enforcing small model weights:

$$\mathcal{R}_F = \|\boldsymbol{\Theta}\|_F^2 \tag{7}$$

The overall objective function so far is the summation of the task-specific cross entropy losses, graph Laplacian-based prior knowledge regularization and Frobenius norm, which is defined as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1\mathcal{R}_{Lap} + \lambda_2\mathcal{R}_F \tag{8}$$

where λ_1, λ_2 are the regularization weights that can be tuned respectively.

Reducing Negative Effects from Task Imbalance

The deep multi-task neural network is challenging to train properly. As the joint learning process is often dominated by the task-specific loss with the largest value, in the training process the model parameters of the fast-converging tasks may be overfitted while those of the slowly-converging tasks are still underfitted. To address this problem, our goal is to find a shared representation, so the losses of all tasks converge to their own optimum. Task-specific loss weights have been introduced to mitigate the imbalance among the loss functions. The general expression of multi-task loss function is:

$$\mathcal{L} = \sum_k \nu_k\mathcal{L}_k + \mathcal{L}_{aux} + \lambda_1\mathcal{R}_{Lap} + \lambda_2\mathcal{R}_F \tag{9}$$

where $\mathcal{L}_k$ is the task-specific loss for task k, ν_k is the loss weight for task-specific loss k, $\mathcal{L}_{aux}$ is the auxiliary loss to balance the rate of change of task-specific losses. One example choice of task-specific loss is the cross entropy loss $\mathcal{L}_{CEk}$ for classification problem.

As for multivariate differentiable functions, the directional derivative along a certain vector $\mathbf{v}$ is a scalar function that describes the rate of change in the direction of vector $\mathbf{v}$. The directional derivative of the multivariate function $\mathcal{L}_k(\boldsymbol{\Theta}_{L,k})$ along direction $\mathbf{v}$ is defined as:

$$\begin{aligned}\nabla_{\mathbf{v}}\mathcal{L}_k(\boldsymbol{\Theta}_{L,k}) &= \mathbf{v}\cdot\nabla\mathcal{L}_k(\boldsymbol{\Theta}_{L,k}) \\ &= \lim_{t\to 0}\frac{\mathcal{L}_k(\boldsymbol{\Theta}_{L,k}+t\mathbf{v}) - \mathcal{L}_k(\boldsymbol{\Theta}_{L,k})}{t}\end{aligned} \tag{10}$$

The rate of change of task-specific loss is actually the directional directive along the direction of $\Delta\boldsymbol{\Theta}_{L,k}$, which is the change of model parameter from last epoch. The directional derivative of loss function $\mathcal{L}_k(\boldsymbol{\Theta}_{L,k})$ along unit directional vector $\mathbf{v} = \Delta\boldsymbol{\Theta}_{L,k}/\|\Delta\boldsymbol{\Theta}_{L,k}\|$ is:

$$\begin{aligned}&\nabla_{\mathbf{v}}\mathcal{L}_k(\boldsymbol{\Theta}_{L,k}) = \\ &\lim_{\|\Delta\boldsymbol{\Theta}_{L,k}\|\to 0}\frac{\mathcal{L}_k(\boldsymbol{\Theta}_{L,k}+\Delta\boldsymbol{\Theta}_{L,k}) - \mathcal{L}_i(\boldsymbol{\Theta}_{L,k})}{\|\Delta\boldsymbol{\Theta}_{L,k}\|}\end{aligned} \tag{11}$$

When implementing the auxiliary loss $\mathcal{L}_{aux}$, we approximate the numerator and denominator of (12) by task-specific loss differences ($\Delta\mathcal{L}_k$) and model parameter updates ($\Delta\boldsymbol{\Theta}_{L,k}$) from the current and last epoch. The auxiliary loss is designed in such way that smaller value results in more balanced task-specific losses. The auxiliary loss should consider the balance among all pairs of tasks. Hence, it is a sum of squared differences among all pairs of approximate directional derivatives:

$$\mathcal{L}_{aux} = \sum_{i\neq j}\left(\frac{\nu_i\Delta\mathcal{L}_i}{\|\Delta\boldsymbol{\Theta}_{L,i}\|} - \frac{\nu_j\Delta\mathcal{L}_j}{\|\Delta\boldsymbol{\Theta}_{L,j}\|}\right)^2 \tag{12}$$

where $i, j \in 1, ..., K$, $\Delta\mathcal{L}_i/\|\Delta\boldsymbol{\Theta}_{L,i}\|$ and $\Delta\mathcal{L}_j/\|\Delta\boldsymbol{\Theta}_{L,j}\|$ are approximate directional derivatives of parameters in ith and jth task-specific layer.

Optimization

The joint objective function with prior knowledge regularization $\mathcal{R}_{Lap}$ and the auxiliary loss $\mathcal{L}_{aux}$ for task-specific loss balancing we optimize for the multi-horizon alerts is:

$$\begin{aligned}\min_{\boldsymbol{\Theta},\nu}\mathcal{L}(\boldsymbol{\Theta},\nu) = \min_{\boldsymbol{\Theta},\nu}&\sum_k\nu_k\mathcal{L}_{CEk}(\boldsymbol{\Theta}) + \mathcal{L}_{aux}(\boldsymbol{\Theta}) \\ &+\lambda_1\mathcal{R}_{Lap}(\boldsymbol{\Theta}) + \lambda_2\mathcal{R}_F(\boldsymbol{\Theta})\end{aligned} \tag{13}$$

The objective function $\mathcal{L}(\boldsymbol{\Theta},\nu)$ is non-convex with respect to $\boldsymbol{\Theta}$ and ν. The block coordinate descent method is applied to find the optimal value of objective function along a direction one at a time.

When solving variable block $\boldsymbol{\Theta}$ with fixed variable block ν, the optimization problem becomes the common neural network parameter learning. Various gradient based methods, such as stochastic gradient descent, can be used to solve the optimization problem. The model parameters $\boldsymbol{\Theta}$ are updated via back-propagation.

When solving variable block ν with fixed variable block $\boldsymbol{\Theta}$, the objective function is a multivariate quadratic function with respect to ν. The optimal ν has a closed-form analytic solution. We define $\bar{\nu_i}$ as the ith task-specific loss weight to achieve optimal value of $\mathcal{L}$:

$$\begin{aligned}\bar{\nu_i} = \operatorname*{argmin}_{\nu_i}&\,\mathcal{L} = \\ &\frac{\|\Delta\boldsymbol{\Theta}_{L,i}\|}{n_{i\neq j}\Delta\mathcal{L}_i}\left(\sum_{i\neq j}\frac{\nu_j\Delta\mathcal{L}_j}{\|\Delta\boldsymbol{\Theta}_{L,j}\|} - \frac{\mathcal{L}_i\|\Delta\boldsymbol{\Theta}_{L,i}\|}{2\Delta\mathcal{L}_i}\right)\end{aligned} \tag{14}$$

where $n_{i\neq j}$ denotes the number of distinct task pairs (i, j). In practical implementation, the loss weights ν are bounded to positive values to avoid the exploding gradient issue. Assume $\nu_i \in [\nu_1, \nu_2]$ for $\forall i$, where $\nu_1, \nu_2 > 0$, the optimal ν_i

is:

$$\nu_i = \begin{cases} \nu_1 & \bar{\nu_i} < \nu_1 \\ \nu_2 & \bar{\nu_i} > \nu_2 \\ \bar{\nu_i} & \nu_1 < \bar{\nu_i} < \nu_2 \end{cases} \tag{15}$$

The complete procedure of learning the neural network model with prior knowledge regularization and task-specific loss balancing is summarized as Algorithm 1. Since optimizing the objective function with respect to ν has a closed-form solution, updating the loss weights ν only adds constant computation complexity to the neural network optimization. The computational complexity of the optimization algorithm is much lower than other approaches that optimize loss weight via a separate gradient based optimization.

Algorithm 1: Model Optimization with Task-specific Loss Balancing

```
Input: Training Data $\mathbf{x} \in \mathbb{R}^{N \times D}$, Training Label
       $\mathbf{y} \in \{0,1\}^{N \times K}$
Output: Model Parameters $\mathbf{\Theta}$, Loss Weights $\nu$
Compute similarity matrix $\mathbf{S}$ by (6);
Initialize $\mathbf{\Theta}$ and $\nu$ ;
for $t = 1, ..., T$ do
    Update $\mathbf{\Theta}$ via back-propagation by optimizing (9);
    Update $\nu$ by using (14) and (15);
end
```

Experimental Evaluation

To evaluate our proposed prior knowledge regularization and task-specific loss balancing, we performed multiple experiments on a real clinical dataset. In this section, we compare the performance of applying the proposed techniques with the commonly used models for prediction of clinical deterioration on the ward. Then we show that our proposed model can be interpretable in terms of utilizing clinically significant features.

Dataset

The dataset used to evaluate the prediction models is an EHR dataset of adult oncology patients from Barnes-Jewish Hospital. The dataset contains patient demographics, comorbidities, vital signs, laboratory results, and medications of all adult hospitalizations from 2014 to 2017 for cancer or stem cell transplant. The dataset consists of 20,700 distinct encounters with hospitalization longer than 48 hours, including 1,939 encounters with deterioration during the hospitalization.

Data Preprocessing

The patient's demographics data, comorbidity diagnoses, time-stamped vital signs and their second order features, lab values, cultures, medications, and procedures from that data window are used as the input to the predictive models. For each encounter with no deterioration, we extract one example with the EHR data in the 6 hours starting from the beginning of hospitalization as the input data. Since the encounters do not contain any deterioration event (ICU transfer or death), the labels for 6-hour, 24-hour and 48-hour alerts are all zeros. For each encounter with deterioration, we extract three examples corresponding to 48, 24, and 6 hours, respectively, prior to the deterioration event. Figure 2 illustrates how the three examples are generated associated with a deterioration event. We segment the timeline into 6-hour windows. If a clinical deterioration event occurs within a time window, a positive label is generated at the end of the time window. For the first example, the input data includes the EHR data in the time window ending at 48 hours before the positive label corresponding to the deterioration event. The labels for the 6-hour, 24-hour, and 48-hour alerts are negative, negative, and positive, respectively. For the second example, the input data is from the time window ending 24 hours prior to the label generated for the deterioration event. Hence, the label for the 6-hour alert is negative, followed by a positive label for the 24-hour alert. If the patient remains in the ICU or is deceased 24 hours after the label generated for the initial deterioration event, the label for the 48-hour alert is positive. Otherwise, the patient has been discharged from the ICU and the 48-hour alert is labeled negative. The third example is extracted in a similar fashion, except the time window slides 24 hours further. The examples extracted from the encounters form the dataset for training and evaluating the model performance with random train-test split. For the time-series data, we performed last value carry-forward followed by median imputation to address missing values in the raw data. All the features used as input to the model are normalized when feeding into the models.

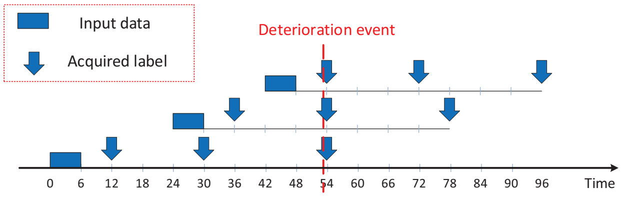

Figure 2: Extracting examples from a hospitalization encounter with clinical deterioration

Model Evaluation

In this section, the evaluation results are reported over 20 experiments with random 75% vs. 25% train-test split. For the training data, we further split it as 10% of training data used as a validation set for hyperparameter tuning. We apply random over-sampling to the training data to handle the imbalanced dataset problem.

Comparison with Common Models We compare the performance of Modified Early Warning Score (MEWS), logistic regression with elastic net regularization (LR-EN), random forest (RF), gradient boosted tree (GBT), single-task neural network (DNN) trained separately for each alert and multi-task neural network (DMNN) with our proposed methods (DMNN$_{\text{pb}}$). In addition, we evaluate the individual effect of prior knowledge regularization (DMNN$_{\text{p}}$) and task-specific loss balancing (DMNN$_{\text{b}}$) by applying each of them alone. The deep models used in our evaluation are the four-layer fully connected neural network. For the multi-task variants, the first three layers are shared among all the tasks, while the fourth layer is the task-specific layer that has separate parameters for each task. We show the performance metrics of all models for predicting 6-hour, 24-hour and 48-hour alerts with specificity fixed around 0.95 in Table 1, since an important clinical goal is to correctly identify risk patients with fewer false alerts.

Since our dataset lacks the AVPU score (”Awake, Verbal, Pain, Unconscious” - a common nursing scale used for evaluating consciousness), the MEWS used in our dataset was calculated without AVPU score. The area under the ROC curve (AUROC) of the MEWS model is around 0.5 for all three horizons of alert, which indicates the MEWS model without the AVPU score actually cannot predict the deterioration events in our study. The superior results from machine learning models suggest they can predict the upcoming deterioration events, even if the AVPU score is not recorded for some of the study cases. When comparing the state-of-the-art machine learning models, gradient boosted trees have better AUROC values than those of penalized logistic regression and random forest. Although penalized logistic regression does not have high AUROC values, it has better sensitivity and precision for 24-hour and 48-hour alerts.

The deep models (DNN), which are trained separately for each horizon, do not have any advantages compared to gradient boosted trees in terms of AUROC, sensitivity and precision. The results demonstrate the superior predictive power of deep multi-task models, compared to the ”shallow” machine learning models, such as penalized logistic regression, random forest and gradient boosted trees. DMNN$_b$ outperforms DMNN in terms of 6-hour (p=0.027), 24-hour ($p<0.001$), 48-hour ($p<0.001$) sensitivity; 6-hour ($p<0.001$), 24-hour ($p<0.001$), 48-hour ($p<0.001$) precision. DMNN$_p$ outperforms DMNN in terms of 6-hour ($p<0.001$), 24-hour ($p<0.001$), 48-hour ($p<0.001$) sensitivity; 6-hour ($p<0.001$), 24-hour ($p<0.001$), 48-hour ($p<0.001$) precision; 24-hour AUROC (p=0.030). The results demonstrate the performance improvement after incorporating the proposed techniques. In our evaluation, the DMNN$_{pb}$ model has the best performance among all the models for the three horizon alerts in terms of all the metrics. The DMNN$_{pb}$ model has the highest precision compared to the MEWS and the state-of-the-art models previously used for clinical early warning. DMNN$_{pb}$ can achieve high AUROC (0.9493) for 24-hour alert as well as high sensitivity (0.5419) and precision (0.7062) while fixing the specificity at around 0.95. One interesting finding is that the prediction of 24-hour alert is the most accurate among the three horizon alerts for all models except MEWS.

Alert Sequence

The goal of multi-horizon alerts is to inform clinicians of the level of urgency of an impending deterioration event to facilitate planning for intervention. For multi-horizon alerts to be effective and accepted by clinicians, the resulting alert sequence should be consistent with real-world clinical experiences. In clinical practice, a deteriorated patient is rarely discharged from the ICU within hours after the deterioration event (when the patient was transferred to the ICU or became deceased). As a result, it is unusual for a positive alert to be followed by a negative alert in the sequence of alerts generated by a multi-horizon alert model. Henceforth, we use an *unexpected sequence* to refer to an alert sequence in which a positive alert precedes a negative alert, and an *expected sequence* to refer to an alert sequence without any negative alert following a positive one. Figure 3 shows all expected alert sequences and unexpected alert sequences. Only 0.77% of the examples in our dataset have unexpected sequences. Therefore, a multi-horizon alert model should avoid predicting unexpected sequences that may reduce the trust and acceptance of the model among clinicians.

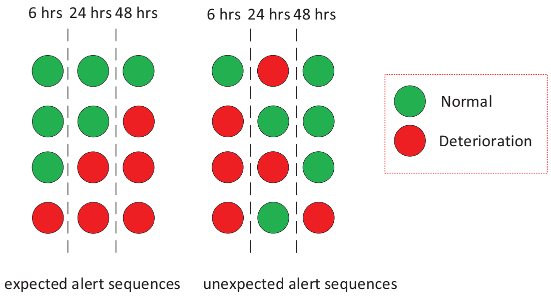

Figure 3: Expected alert sequences (left) and unexpected alert sequences (right)

We define *sequence concordance* as the number of expected alert sequences over the total number of alert sequences. Table 2 lists the Sequence Concordance for DNN, DMNN, DMNN$_p$, DMNN$_{pb}$. The results indicate the effectiveness of incorporating prior knowledge to regularize the model and generate expected alert sequences that are concordant with clinical experiences. We also investigate the models' ability to generate correct alerts for each individual patient. The one-alert, two-alert and three-alert concordances represent the percentage of correctly predicted alerts for at least one, two or three, respectively. As summarized in Table 2, the DMNN$_p$ model can achieve the 0.7755 three-alert concordance, which means 77.55% of the alert sequences are correct for all three alerts over different horizons. The results suggest that prior knowledge can exploit the event relatedness and constrain the predicted multi-horizon alerts to be expected sequences. Furthermore, the concordance results of DMNN$_{pb}$ in Table 2 show that task-specific loss balancing does not have a negative impact on the prior knowledge regularization in enforcing expected alert sequences.

Predictive Features

An additional important aspect of clinical event prediction involves conveying the predictive results to clinical decision-makers. Deep learning models are known for their opaqueness on generating the predictions. In this section, we apply Layer-wise relevance propagation (Bach et al. 2015), which has been widely used for interpreting the outcomes

Table 1: Performance evaluation of predictive models with mean and standard deviation reported (specificity fixed around 0.95)

Model	6-hour alert			24-hour alert			48-hour alert		
	AUROC	Sensitivity	Precision	AUROC	Sensitivity	Precision	AUROC	Sensitivity	Precision
MEWS	.5013(.0048)	.0662(.0072)	.1014(.0110)	.4993(.0035)	.0647(.0045)	.1884(.0128)	.4999(.0040)	.0645(.0039)	.1847(.0113)
LR-EN	.8852(.0022)	.3050(.0122)	.3976(.0106)	.9216(.0013)	.4620(.0129)	.6772(.0074)	.9106(.0025)	.4262(.0254)	.6520(.0131)
RF	.8660(.0033)	.1370(.0112)	.2270(.0161)	.9309(.0015)	.3110(.0119)	.5830(.0089)	.9090(.0022)	.1480(.0182)	.3917(.0300)
GBT	.9122(.0019)	.3228(.0140)	.4205(.0118)	.9442(.0014)	.4493(.0356)	.6668(.0090)	.9294(.0019)	.3560(.0635)	.5911(.0237)
DNN	.8604(.0051)	.2588(.0132)	.3633(.0114)	.9263(.0063)	.4400(.0086)	.6382(.0121)	.9165(.0029)	.3481(.0103)	.5964(.0139)
DMNN	.9144(.0018)	.3393(.0098)	.4158(.0088)	.9478(.0014)	.4944(.0168)	.6715(.0065)	.9329(.0016)	.4053(.0347)	.6220(.0159)
DMNN$_p$	.9142(.0020)	.3576(.0116)	.4325(.0098)	.9489(.0013)	.5400(.0160)	.7061(.0077)	.9334(.0016)	.4403(.0284)	.6483(.0155)
DMNN$_b$	.9136(.0019)	.3552(.0175)	.4335(.0079)	.9487(.0012)	.5215(.0089)	.7009(.0046)	.9332(.0019)	.4405(.0344)	.6566(.0343)
DMNN$_{pb}$	**.9147(.0019)**	**.3629(.0097)**	**.4346(.0087)**	**.9493(.0014)**	**.5419(.0171)**	**.7062(.0083)**	**.9345(.0016)**	**.4413(.0276)**	**.6567(.0140)**

Table 2: Concordance evaluation with mean and standard deviation reported

	DNN	DMNN	DMNN$_p$	DMNN$_{pb}$
Seq.	.7548(.0031)	.9542(.0023)	.9814(.0020)	**.9817(.0022)**
1-alert	.8745(.0035)	.8839(.0028)	.8857(.0019)	**.8861(.0020)**
2-alert	.8039(.0027)	.8162(.0024)	.8206(.0029)	**.8212(.0032)**
3-alert	.7592(.0033)	.7677(.0028)	.7755(.0031)	**.7758(.0028)**

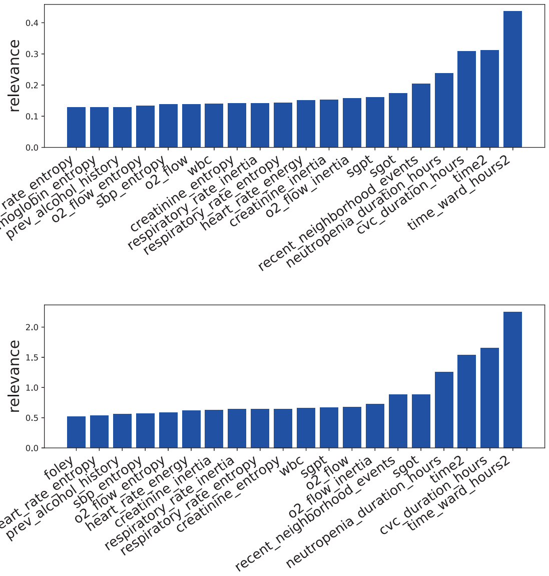

Figure 4: Top 20 features with the largest impacts on positive outcomes (top) and negative outcomes (bottom)

of deep neural networks. Figure 4 shows the top 20 features with the highest relevance to the positive and negative outcomes. `time_ward_hours2` (the square of hours on the ward) has the greatest impact on both outcomes, followed by `cvc_duration_hours` (the duration a central venous catheter had been in place) and `time2` (the square of total time in the hospital). It seems reasonable that these are the most predictive features in the model, since patients remaining in the hospital or on the ward for a long time are likely to be sicker, and thus more likely to suffer clinical deterioration. Among the top 20 relevant features listed for positive outcomes, 10 of them are second order statistical features from the time series data, which are entropy, energy and inertia generated from heart rate, hemoglobin, oxygen flow, creatinine, respiratory rate, and systolic blood pressure. The result is similar for negative outcomes. The findings suggest the second order statistical features have significant impact on the prediction outcomes, which support our claim of incorporating second order statistical features into the model input is beneficial. The results also suggest that extracting temporal features from time series data is extremely useful for training predictive models.

Conclusion and Future Work

We applied two new approaches - prior knowledge regularization and task-specific loss balancing - to a deep multi-task prediction model for clinical deterioration among a high-risk group of hospitalized patients. The evaluation on a large dataset of adult oncology patients demonstrates that the proposed techniques effectively improve the prediction accuracy of multi-horizon alerts. Our review of highly-predictive features includes many expected to contribute to a patient's high-risk status, as well as a number of complex features not being used in previous early warning systems. The application of such a model to a high-risk cohort of patients may be quite beneficial clinically, as common early warning systems are limited by low positive predictive values. Another advantage of our approach is that simultaneous multi-horizon alerts may be clinically useful by allowing the planning and triage of diagnostic tests and interventions earlier in a patient's clinical course. Because these approaches have only been evaluated retrospectively in a single cohort of patients, future evaluations should be considered. External evaluations could enhance the generalizability of this approach to other hospitals and patient groups, and prospective evaluations could ensure temporal validity prior to implementing for real-time use.

Acknowledgement

This research is supported, in part, by the Fullgraf Foundation.

References

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE* 10:1–46.

Bailey, T. C.; Chen, Y.; Mao, Y.; Lu, C.; Hackmann, G.; Micek, S. T.; Heard, K. M.; Faulkner, K. M.; and Kollef, M. H. 2013. A trial of a real-time alert for clinical deterioration in patients hospitalized on general medical wards. *Journal of Hospital Medicine* 8(5):236–242.

Bedoya, A. D.; Clement, M. E.; Phelan, M.; Steorts, R. C.; O'Brien, C.; and Goldstein, B. A. 2019. Minimal impact of implemented early warning score and best practice alert for patient deterioration. *Critical Care Medicine* 47(1).

Che, Z.; Kale, D.; Li, W.; Bahadori, M. T.; and Liu, Y. 2015. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 507–516. ACM.

Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; and Rabinovich, A. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks.

Churpek, M. M.; Yuen, T. C.; Huber, M. T.; Park, S. Y.; Hall, J. B.; and Edelson, D. P. 2012. Predicting cardiac arrest on the wards: a nested case-control study. *Chest* 141(5):1170–1176.

Churpek, M. M.; Yuen, T. C.; Park, S. Y.; Gibbons, R.; and Edelson, D. P. 2014. Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards*. *Critical care medicine* 42(4):841–848.

Churpek, M. M.; Yuen, T. C.; Winslow, C.; Meltzer, D. O.; Kattan, M. W.; and Edelson, D. P. 2016. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical care medicine* 44(2):368–374.

Cooksley, T. J.; Kitlowski, E.; and Haji-Michael, P. 2012. Effectiveness of modified early warning score in predicting outcomes in oncology patients.

Desautels, T.; Das, R.; Calvert, J.; Trivedi, M.; Summers, C.; Wales, D. J.; and Ercole, A. 2017. Prediction of early unplanned intensive care unit readmission in a uk tertiary care hospital: a cross-sectional machine learning approach. *BMJ Open* 7(9).

Escobar, G., and dellinger, R. 2016. Early detection, prevention, and mitigation of critical illness outside intensive care settings: Critical illness outside the icu. *Journal of Hospital Medicine* 11:S5–S10.

Hu, S. B.; Wong, D. J. L.; Correa, A.; Li, N.; and Deng, J. C. 2016. Prediction of clinical deterioration in hospitalized adult patients with hematologic malignancies using a neural network model. *PloS one* 11(8):e0161401–e0161401.

J R de Bie, A.; Subbe, C. P.; Bezemer, R.; Cooksley, T.; G Kellett, J.; Holland, M.; Bouwman, R. A.; Bindels, A.; and H M Korsten, E. 2019. Differences in identification of patients' deterioration may hamper the success of clinical escalation protocols. *QJM : monthly journal of the Association of Physicians* 112.

Jeffery, A. D.; Dietrich, M. S.; Fabbri, D.; Kennedy, B.; Novak, L. L.; Coco, J.; and Mion, L. C. 2018. Advancing in-hospital clinical deterioration prediction models. *American journal of critical care : an official publication, American Association of Critical-Care Nurses* 27(5):381–391.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kipnis, P.; Turk, B.; Wulf, D.; LaGuardia, J.; XLiu, V.; Churpek, M.; Brufau, S.; and J. Escobar, G. 2016. Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the icu. *Journal of Biomedical Informatics* 64.

Lin, Y.-W.; Zhou, Y.; Faghri, F.; Shaw, M. J.; and Campbell, R. H. 2019. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLOS ONE* 14(7):1–22.

Liu, C.; Wang, F.; Hu, J.; and Xiong, H. 2015. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 705–714. ACM.

Lyons, P. G.; Klaus, J.; McEvoy, C. A.; Westervelt, P.; Gage, B. F.; and Kollef, M. H. 2019. Factors associated with clinical deterioration among patients hospitalized on the wards at a tertiary cancer hospital. *Journal of oncology practice* 15(8):e652–e665.

Nori, N.; Kashima, H.; Yamashita, K.; Ikai, H.; and Imanaka, Y. 2015. Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 855–864. ACM.

Rubin, J.; Potes, C.; Xu-Wilson, M.; Dong, J.; Rahman, A.; Nguyen, H.; and Moromisato, D. 2018. An ensemble boosting model for predicting transfer to the pediatric intensive care unit. *International Journal of Medical Informatics* 112:15 – 20.

Subbe, C. P.; Kruger, M.; Rutherford, P. J.; and Gemmel, L. 2001. Validation of a modified early warning score in medical admissions. *QJM : monthly journal of the Association of Physicians* 94 10:521–6.

Wellner, B.; Grand, J.; Canzone, E.; Coarr, M.; Brady, P. W.; Simmons, J.; Kirkendall, E.; Dean, N.; Kleinman, M.; and Sylvester, P. 2017. Predicting unplanned transfers to the intensive care unit: A machine learning approach leveraging diverse clinical elements. *JMIR Med Inform* 5(4):e45.

Zhai, H.; Brady, P.; Li, Q.; Lingren, T.; Ni, Y.; Wheeler, D. S.; and Solti, I. 2014. Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children. *Resuscitation* 85(8):1065 – 1071.