Phenotyping with Prior Knowledge using Patient Similarity

Asif Rahman ASIF.RAHMAN@PHILIPS.COM
Philips Research North America
Cambridge, MA, USA

Yale Chang YALE.CHANG@PHILIPS.COM
Philips Research North America
Cambridge, MA, USA

Bryan Conroy BRYAN.CONROY@PHILIPS.COM
Philips Research North America
Cambridge, MA, USA

Minnan Xu-Wilson MINNAN.XU@PHILIPS.COM
Philips Research North America
Cambridge, MA, USA

Abstract

Prior medical knowledge, like the relationships between diseases or treatments and their corresponding risk factors are widely available in electronic health records (EHR), can be generated by domain experts, and extracted from knowledge graphs. Although informative for predictive modeling tasks, most of the patient-specific knowledge in EHR are not utilized because of practical constraints on data availability or cost of acquiring the data to make inferences. We present a method to learn from prior knowledge using a mixture–of–experts model where gating probabilities are tuned by an adjacency matrix created using side information available during training, like comorbidities, interventions, outcomes, vital signs and laboratory measurements. The adjacency matrix of a nearest neighbor graph is used to discover subgroups of intensive care unit (ICU) patients. Experts are shown to specialize based on how patients are grouped in the adjacency matrix on two real-world decision support tasks: predicting hemodynamic interventions and stratifying patients at risk for developing a sustained period of hypoxemia. The proposed prior knowledge-guided learning (PKL) model discovers clinically meaningful cohorts in patients with respiratory compromise that match well known sub-phenotypes described in the literature.

1. Introduction

A long standing challenge in machine learning in healthcare is in combining the medical knowledge of experts with data-driven insights from electronic health records (EHR). A general methodology to capitalize on prior knowledge is critical in high-stakes decision making domains like medicine where machine learning algorithms are proving increasingly effective in a wide range of applications but are often data inefficient and fail to generalize to new cases. This is largely driven by heterogeneity of the patient population. Individuals within a subgroup share risk factors and have correlated outcomes, however, differences in physiological traits across subgroups manifest as variations in the response to treatment, prevalence of underlying diseases, and long-term outcomes.

Prior knowledge can be derived from retrospective data to uncover a relationship between the observed risk factors (e.g. vital signs and laboratory measurements) and *meta features* that encode the medical knowledge. By meta features, we refer to factors that characterize subgroups of patients, that are correlated with the risk factors but are not available to the model at inference time. A continuous risk prediction algorithm deployed on a bedside patient monitor, for example, may not have access to comorbidities, interventions, or even the unit type – conditions that introduce heterogeneity to the patient population. A heterogenous cohort gives rise to a multifaceted problem: 1) a single-task model trained on a heterogenous cohort may not generalize well to subpopulations, 2) subpopulation-specific models are data inefficient (small data leads to poor representation learning), and 3) the subpopulation a patient belongs-to is typically unknown at inference time, which limits our ability to choose a subpopulation-specific model.

The problem of learning from a diverse clinical population where side information is available during training is often addressed in a multi-task learning (MTL) framework that models sub-tasks along with the primary task to improve generalization (Caruana (1997); Zhang and Yang (2018); Ruder (2017)). The sub-tasks are introduced to improve the representation learning capacity of a model through a shared representation of the data for each task. Sub-tasks are equivalent to the meta features described above and can include future outcomes like ICD-9 diagnosis codes, ICU unit type, and interventions, among other features that are not directly observed at evaluation time and are useful to the primary task. Many current MTL models, however, optimize for objectives such as log-likelihood in a multi-label classification problem, producing useful representations to improve model performance on the primary task only as a side effect.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work offers an alternative to the standard MTL setup by introducing an unsupervised objective based on the adjacency matrix of a nearest neighbor graph that can encode any number of sub-tasks without explicitly modeling each sub-task. Our prior knowledge guided learning (PKL) framework utilizes the well known mixture-of-experts (MoE) model to tune the gating network to assign similar patients to the same expert, which allows the experts to specialize on parts of the data. We describe this general method to encode prior knowledge into a neural network to achieve the following goals:

- Utilize knowledge in EHR that are not available to the model during evaluation.
- Enable cohort discovery by specializing on subgroups of similar patients. In contrast to the Multi-Task Learning (MTL), we condition the expert models to specialize on pre-defined subgroups. The shared data representation is explicitly controlled, and a useful representation is learned by design rather than as a byproduct of solving a multi-task classification problem.

We first use the PKL model to predict hemodynamic interventions like fluids, vasopressors, and packed red blood cell (PRBC) transfusions in ICU patients to validate that the PKL model can learn to cluster similar patients based on the adjacency matrix. We constructed adjacency matrices using the patients ICU unit type and intervention type as

meta features and show that similar patients are grouped to the same expert. We then use the PKL model to discover clinically meaningful phenotypes in patients that develop respiratory compromise as characterized by a sustained period of hypoxemia (oxygenation index ≥ 25). Previous work using unsupervised latent class analysis of clinical trial data had identified at least two distinct sub-phenotypes of acute respiratory distress syndrome characterized by severe inflammation, shock, and metabolic acidosis (Calfee et al. (2018, 2014); Famous et al. (2017)). The PKL model with adjacency matrix derived from observed risk factors like laboratory measurements and vital signs discovers similar phenotypes in patients that develop respiratory compromise.

2. Related Work

This paper touches on several domains, including computational phenotyping with EHR, multi-task learning, and mixture-of-experts (MoE). We briefly highlight relevant work in relation to our proposed method for phenotyping with prior knowledge.

The goal of computational phenotyping is to discover subgroups of similar patients that share an underlying disease mechanism and exhibit similar traits. The task of computational phenotyping is analogous to unsupervised clustering using clinical risk factors to identify subsets of patients that share similar physiological characteristics. One approach to phenotyping is to use the latent representation learned from a neural network to cluster patients and assign meaning to the clusters (Lasko et al. (2013); Kale et al. (2015)). Schulam et al. (2015) clusters time series of clinical markers using a hierarchical probabilistic model to uncover disease subtypes in patients with an autoimmune disease known to be heterogenous. Suresh et al. (2018) proposed a two step solution to discover cohorts of patients with similar traits by using an unsupervised clustering model to group patients then getting predictions from a multi-task model for each subgroup, where the data representation is shared across tasks. Our work starts with the same goal of discovering cohorts of patients with similar phenotypes but our proposed PKL model learns to phenotype in an end-to-end learning framework.

Recently, Harutyunyan et al. (2019) observed there is correlation between many clinical prediction tasks, including mortality, length of stay, decompensation and diagnosis prediction. They used a MTL framework to learn a shared representation for all tasks and predict outcomes using a single forward pass of a neural network by optimizing for an overall loss that was the weighted sum of task-specific losses. MTL introduces inductive bias in the form of sub-tasks which causes the model to prefer hypotheses that explain more than one task. On specific problems, MTL has better performance than single-task learning (STL) and generally leads to solutions that generalize better as a result of the inductive transfer (Ruder (2017)). A common neural architecture for MTL is hard parameter sharing where hidden layers are shared across tasks, while keeping several task-specific output layers (Caruana (1997); Baxter (1997)). We use this particular instantiation of MTL as a benchmark to compare against PKL.

MoE models are ensemble learners where experts specialize on parts of the data and predictions from each expert is combined with mixing probabilities generated by a gating network (Jacobs et al. (1991); Jordan and Jacobs (1994)). MoE appears in many different guises, including conditional computation (Bengio et al. (2016)), and may be unified under

the concept of multiplicative interactions, which can induce inductive bias in practical scenarios such as when multiple streams of information are fused (Jayakumar et al. (2019)). Shazeer et al. (2017) recently described an end-to-end learning framework for a large number of experts where the gating network and experts are learned together using backpropagation. We utilize this learning scheme to train the PKL model.

Another related approach is stratified models, which build distinct predictive models for each setting of one or more categorical variables. For example, Tuck et al. (2019) augment a standard stratified model loss with a regularization term that encourages model parameters to vary smoothly with respect to a similarity derived from the categorical features. These models, however, require an a-priori categorization (clustering) of the data, whereas the proposed PKL model uses softer nearest neighbor information derived from a graph adjacency matrix.

3. Methods

3.1. Mixture-of-Experts

The Mixture-of-Experts (MoE) consists of K "expert networks" $f_1 ... f_K$, and a "gating network" G whose output is a sparse K-dimensional set of probabilities that sum to 1. Figure 1 shows a schematic of the MoE network used to learn from prior knowledge. The experts are neural networks each with their own parameters and described in detail in section 3.3. Experts and the gating network can take different inputs, however, in practice we use the same input $x \in \mathbb{R}^D$ with D numerical features as input to each expert and gating network. Given the output of the gating network $G(x)$ and the output of the k-th expert $f_k(x)$ for an input x, the output of the the MoE model can be written as the following:

$$\hat{y} = \beta_0 + \sum_{k=1}^{K} G(x)_k f_k(x) \tag{1}$$

where $\beta_0 = log\frac{p(y=1)}{p(y=0)}$ is the expected log-odds ratio over the population.

3.2. Learning from prior knowledge

The gating network comprises a fully connected layer, followed by a rectified linear unit (ReLU) non-linearity, batch normalization, and another fully connected layer. A softmax transformation generates gating probabilities over K experts.

Prior knowledge is encoded in an adjacency matrix $A \in \mathbb{R}^{B \times B}$ from a batch of B samples, which is computed from a top-N nearest neighbor graph on M "meta features". For example, in the experiments presented in this paper, meta features included ICU unit type, future hemodynamic interventions, and the raw observations x themselves. For our experiments we selected the 2-nearest neighbors to create the adjacency matrix. Our goal is to have the gating network assign similar patients to the same expert with a high probability. Therefore, given the batch gating probabilities $P \in \mathbb{R}^{B \times K}$, PP^T is an approximation of A since the elements of PP^T are close to 1 for patients assigned to the same expert. Note that PP^T is also the adjacency matrix of a graph on data samples whose edges are weighted by

the inner product between gating probabilities. The similarity regularization term $L_{\text{similarity}}$ pushes the gating probabilities to group patients according to the adjacency matrix:

$$L_{\text{similarity}} = -tr(P^T AP) \tag{2}$$

In this way, two data samples are encouraged to have similar gating probability profiles if they are deemed similar by the prior knowledge graph.

Our two design goals for the gating network are 1) to favor sparse gating probabilities to get experts to specialize on parts of the data and 2) to achieve non-uniform gating probabilities to utilize all experts. We observe that the gating network tends to converge to a state that over utilizes a few experts with high probabilities and therefore introduce regularization terms that push the gating probabilities $P_b = \{p_{b1}, ..., p_{bK}\}$ for sample b to have a desired activation probability κ, similar to Bengio et al. (2016). The two loss terms L_b and L_e, activates each expert with probability κ in expectation over the data and pushes experts to have a desired sparsity for each example. The final regularization term L_v maximizes the variance of gating probabilities of each expert across the data and explicitly discourages uniform gating probabilities:

$$L_b = \frac{1}{K}\sum_{k=1}^{K}((\frac{1}{B}\sum_{b=1}^{B} p_{bk}) - \kappa)^2 \tag{3}$$

$$L_e = \frac{1}{B}\sum_{b=1}^{B}((\frac{1}{K}\sum_{k=1}^{K} p_{bk}) - \kappa)^2 \tag{4}$$

$$L_v = -(\frac{1}{B}\sum_{b=1}^{B}\text{var}_0\{P\} + \frac{1}{K}\sum_{k=1}^{K}\text{var}_1\{P\}) \tag{5}$$

where var_i is taking the variance along the i-th dimension of the matrix.

The gating network and expert models are trained end-to-end with gradient descent using the Adam optimizer, learning rate 0.001, and batch size 512. For classification tasks, we minimize the binary cross entropy along with the regularization terms:

$$\mathcal{L} = L_{\text{BCE}}(y, \hat{y}) + \lambda_p(L_b + L_e) + \lambda_v L_v + \lambda_{\text{similarity}} L_{\text{similarity}} \tag{6}$$

where $\lambda_p = \lambda_v = 1$, and $\lambda_{\text{similarity}} = 0.01$.

3.3. Expert model

Expert models can take any form. In this work, the k–th expert f_k is a form of generalized additive model with interactions (XGAM):

$$f_k = \beta_0 + \sum_{d=1}^{D} u_d(x_d) + \sum_{t=1}^{T} h(E_t) \tag{7}$$

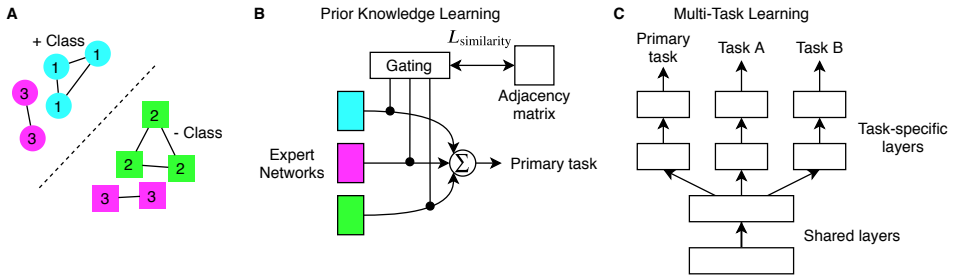


Figure 1: A) Illustrative example of a binary classification problem with three distinct subgroups. Similar samples from the same subgroup are connected by a black line. B) Prior knowledge learning network uses a mixture-of-experts framework. The known subgroups are encoded in an adjacency matrix. The probabilities from the gating network are tuned with an unsupervised loss function to assign similar patients to the same expert according to the adjacency matrix. C) Multi-task learning models have shared hidden layers to transfer knowledge between tasks and has one output network per task.

where x_d is the d-th univariate feature and u_d is a univariate first order linear spline basis function. $E_t \in \mathbb{R}^H$ is an embedding vector of length H representing the leaf node of the t-th decision tree in a gradient boosted decision tree model, like XGBoost. $h(E_t) \in \mathbb{R}$ is a linear transformation mapping the embedded vector representing the leaf node index of the t-th decision tree to a scalar risk. The final prediction of the k-th expert is a sum of the univariate risks and risks assigned to the leaf nodes of a boosted tree.

The univariate linear spline is formed by connecting linear segments and implemented with ReLU functions:

$$u_d(x_d) = \sum_{q=1}^{Q} \left(w_q \text{ReLU}(x_d - \gamma_q) + v_q \text{ReLU}(\gamma_q - x_d) \right) \tag{8}$$

where $\gamma \in \mathbb{R}^Q$ are Q uniformly spaced knots. Inputs x are normalized in the range $[0, 1]$ so knots are equi-spaced in the range $[0, 1]$. $w \in \mathbb{R}^Q$ and $v \in \mathbb{R}^Q$ are free weights to be fitted during training and controls the slope of the line between intervals.

We formulate the embedding matrix E similar to natural language processing where the entities are mapped to a vector representation using a lookup table. Entities are derived from leaf nodes in a trained XGBoost model with each leaf node assigned to a unique token representing the decision path (e.g. [age>65 & lactate>2.5]). The boosting model was trained for 200 rounds and depth–2 to capture pairwise interactions. Trees deeper than two levels are difficult to visualize and we are interested in interpretable models that users can understand the contribution of individual features in the model. The original boosted tree model is not required after deployment as long as the raw feature values can be mapped to tokens in the dictionary of decision paths.

3.4. Multi-task learning

We constructed a multi-task learning network as a baseline. Figure 1 shows the model architecture, which included shared layers and task specific layers, each composed of a the following blocks: a linear layer, followed by a ReLU activation, batch normalization, another linear layer, and a dropout layer. The input to the first shared layer is a linear transformation of the embedded decision paths $(h(E_1), ..., h(E_T))$ from an XGBoost model, as described above.

4. Cohort

The eICU dataset was used for the purposes of training and validating the hemodynamic and respiratory models presented below (Pollard et al. (2018)). The full dataset is comprised of 3.3 million patient encounters from 364 hospitals across the United States. To ensure that charting of hemodynamic intervention data (e.g., vasopressors, inotropes) were accurate in the training and validation cohorts, we restricted our analysis to patients admitted to select hospitals with reliable infusion and ventilation charting data. Specifically, we included ward-years that charted $\geq$ 7 infusion drug entries per patient per day. Included patients with $\geq$0.75 ventilation and airway records per patient per day in the patient care plan, and either $\geq$10 entries per patient per day in respiratory charting tables in eICU database. This filtering step reduced the initial dataset size to 1.4 million patient encounters from 54 hospitals. We selected patients $\geq$ 18 years old and did not have a DNR.

4.1. Cohort & Data Selection for Hemodynamic Instability

Hemodynamic instability is broadly defined as perfusion failure from one or more etiologies including circulatory shock or advanced heart failure. Hemodynamic instability presents with low blood pressure and requires interventions including, vasopressors, fluids, and packed red blood cells (PRBCs) in the case of blood loss. We predict the onset of a significant hemodynamic intervention by dividing patients into stable and unstable cohorts. Stable patients did not receive vasopressors and large volumes of fluids during their ICU stay. Unstable patients received at–least one vasopressor during the ICU stay. These patients ICU stays were further segmented into unstable and intervention periods. An intervention segment started when any of the strong or weak intervention criteria was satisfied (Eshelman et al. (2017), Conroy et al. (2016)).

- Strong interventions:
    1. Administration of any quantity of any of the following inotropic and vasopressor medications: Dobutamine, Dopamine, Epinephrine, Norepinephrine, Phenylephrine, Vasopressin
    2. Administration of Packed Red Blood Cells (PRBCs) in either of the following dosages: 1) 800 cc PRBC over course of 24 hours 2) 500 cc PRBC in two hours followed by fluid therapy
- Weak interventions:

1. Administration of Fluid Therapy (colloid or crystalloid) in the following dosages: 700 cc in one hour, 1500 cc total in four hours, 2400 cc in eight hours, 3000 cc in 12 hours, 500 cc twice in four hours
2. Administration of PRBCs in the following dosage: 500 cc PRBC not followed by fluid therapy within the following 24 hours

The intervention segment continued until there was a gap of more then 12 hours in-between consecutive vasopressors or fluids administrations. The unstable period was a maximum of 24 hours before the start of an intervention but can also be less than 24 hours. A patient can, therefore, have multiple unstable segments during the ICU stay and we use all unstable segments in our model development. The last observation from 1-hour before the intervention was used as a positive class sample and a random time from a stable patient was selected as the negative class sample. We did not include any samples from the first 6 hours of the ICU stay.

The percentage breakdown of hemodynamically unstable patients by intervention category were as follows: 71.9% strong interventions, 36.2% weak interventions, 15.7% PRBCs, 57.9% vasopressors. The cohort selection criteria resulted in 32,896 unstable events and 183,420 stable events (prevalence=18%). A stratified subsample of 20% of the data were held out and reserved for testing of all algorithms, while the remaining 80% were used to train all models.

We selected variables that are routinely acquired in the ICU, including vital signs, laboratory measurements, and blood gas measurements. The full set of variables included: age, heart rate, invasive and noninvasive systolic blood pressure, mean blood pressure, temperature, noninvasive shock index (ratio of heart rate/systolic blood pressure), central venous pressure, base excess, WBC, SaO2, AST, Bands, Basos, BUN, calcium, ionized calcium, CO2, creatinine, EOS, glucose, hematocrit, hemoglobin, lactate, magnesium, PaCO2, potassium, PTT, sodium, bilirubin, FiO2, PIP, and mean airway pressure. Variables were forward filled up to 2 hours for heart rate, systolic blood pressure, and 26 hours for laboratory measurements. We require at–least a heart rate and systolic blood pressure be available for the calculation of a risk score during training.

4.2. Cohort & Data Selection for Respiratory Compromise

Patients with respiratory compromise experience disease progression at different rates largely due to the heterogeneity of respiratory illness. Timely and accurate risk stratification in the ICU can improve outcomes. We used a sustained period of hypoxemia as a marker for respiratory compromise aimed to predict the onset of a sustained period of hypoxemia. Patients with a sustained duration of oxygenation index ≥ 25 for six continuous hours with at least two observations of oxygenation index during the period were considered hypoxemic. Comparator patients were invasively mechanically ventilated but never had oxygenation index ≥ 25 at any time during their ICU stay. Comparator patients did not have an ICD-9 code for Congestive Heart Failure or Acute Respiratory Failure. We applied the same filtering and data processing steps as described in section 4.1. The prevalence of hypoxemia based on the above criteria was 11.8% in the eICU dataset.

We selected variables that are routinely acquired in the ICU, including vital signs, laboratory measurements, and blood gas measurements. Since our task is to distinguish

between respiratory events like hypoxemia onset from comparator patients who may or may not have been on ventilation, we did not include ventilation features in the model since that may be a strong confounder that easily separates the two classes. The full set of variables included: age, heart rate, invasive and noninvasive systolic blood pressure, mean blood pressure, temperature, noninvasive shock index (ratio of heart rate/systolic blood pressure), central venous pressure, base excess, WBC, SaO2, AST, Bands, Basos, BUN, calcium, ionized calcium, CO2, creatinine, EOS, glucose, hematocrit, hemoglobin, lactate, magnesium, PaCO2, potassium, PTT, sodium, and bilirubin. We selected a random sample from the comparator group and a sample 1-hour before the onset of hypoxemia from the respiratory compromise group to train and validate the model.

5. Results

5.1. Evaluation Approach/Study Design

We first present a validation of the effectiveness of including prior knowledge by comparing models trained with ($\lambda_{\text{similarity}} > 0$) and without ($\lambda_{\text{similarity}} = 0$) the similarity loss term. We found that patients in the stepdown unit were more hemodynamically stable than other units - they had higher systolic blood pressure, lower heart rate, and were less likely to be on invasive mechanical ventilation compared to the average patient in a Medical-Surgical or Cardiac Care Unit. Therefore, we developed models with and without prior knowledge about the unit type (Stepdown vs Other). As an additional validation experiment, we repeat this exercise with the specific intervention type with the expectation that patients given similar interventions should be assigned to the same expert. The number of experts and meta features used to create the adjacency matrix are described in Table 1 and results of the validation is presented in section 5.2.

Table 1: Meta features used to create the adjacency matrix for the stepdown and intervention type models.

Condition	Meta features	Experts
Stepdown	In Stepdown Unit, In Other Unit	3
Intervention type	Strong intervention, Weak intervention, PRBC, Is Stable	4

Next we present a comparison between prior knowledge guided learning (PKL), multi-task learning (MTL), and single task learning (STL) in section 5.3. At a high level, PKL and MTL both utilize information tangentially related to the primary objective to improve the performance on the metric we care about, which in the present study is accurately predicting the onset of a hemodynamic intervention. Therefore, in the MTL model we evaluate performance on the primary task of classifying hemodynamic instability by jointly solving auxiliary tasks, which includes predicting the intervention types (Pressors, Fluids, PRBC). In contrast to MTL, which has multiple outputs, the PKL model encodes the intervention type into the gating network and has a single output for the primary task. The STL model uses only the inputs to predict the onset of hemodynamic intervention without any auxiliary information.

Finally, we turn to the task of cohort discovery in section 5.4. Patients with hypoxemia, specifically those who develop acute respiratory distress syndrome (ARDS), exhibit significant heterogeneity in their disease progression and benefit from ventilation settings tailored to the severity of respiratory compromise. Previous work by Calfee et al. (2014) identified subphenotypes of ARDS. We therefore, extracted a cohort of patients who had sustained arterial hypoxemia in the ICU, many of whom go on to develop ARDS, and predicted the onset of a sustained period of hypoxemia. We design an adjacency matrix using the risk factors themselves (x) as meta features to group similar patients. We train a model with the similarity constraint and treat each expert as a subgroup and analyze the physiological characteristics and outcomes within each group to assign clinically meaningful descriptions to each discovered phenotype.

To ensure a fair comparison, we use the same type of classifier and training procedure in all experiments, the XGAM model is used as an expert and trained until validation AUC did not increase for 15 consecutive epochs. Models were allowed to train for a maximum of 100 epochs, although we found models typically converged in less than 50 epochs. All results are presented on a held-out test set representing 20% of the data. Train and test sets were stratified by patient so a single patient was represented in either the train or test, not both. We report the area under the receiver-operator curve (AUC), average precision, precision at the breakeven point (where precision equals recall), and specificity.

5.2. Validation of the Effectiveness of Encoding Prior Knowledge

The prior knowledge learning (PKL) model, if effective, should learn to cluster similar patients in the adjacency matrix to the same expert. We also expect the naive unconstrained mixture-of-experts (MoE) to have a uniform distribution of patients with the condition assigned to each cluster. Indeed that is what we observe in Figure 2. The assignment of stepdown unit patients is concentrated in one expert with the similarity constraint (PKL) but uniform across the experts without a similarity constraint (MoE). Similarly, Figure 3 shows that experts specialize on intervention types when the intervention types are used as meta features to create the adjacency matrix.

Figure 2: Percent of patients in Stepdown units and hemodynamically unstable. Colors indicate experts. PKL groups Stepdown unit patients to the same expert and MoE without the similarity constraints distributes Stepdown unit patients across all experts.

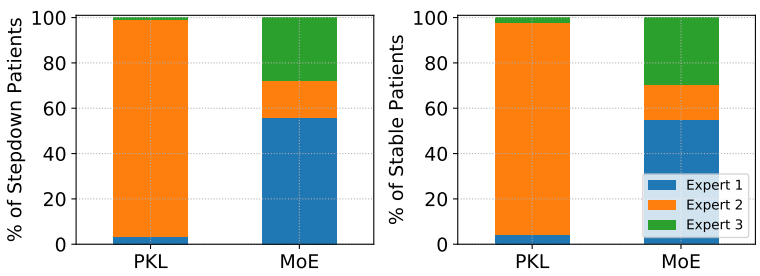
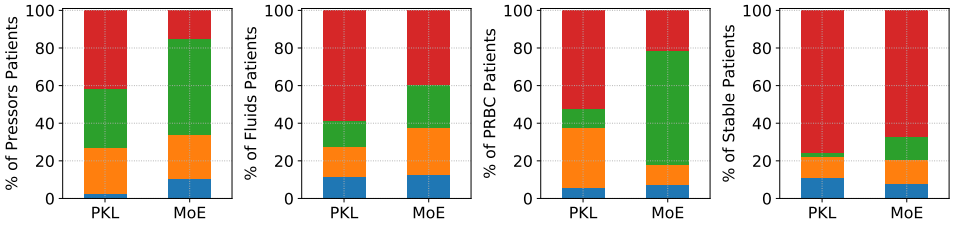


Figure 3: Percent of patients receiving pressors, fluids, or PRBC interventions in each expert. Colors indicate experts.



5.3. Comparison with Multitask Learning

Table 2 compares the PKL model with a full-complexity MTL model (deep non-linear interactions) and shows equivalent AUCs and slightly better positive predictive value in classifying hemodynamic instability in a subgroup of the data: patients in the stepdown unit. The PKL model was trained with stepdown unit as a meta feature (Table 1). The MTL model was trained to classify patients in the stepdown unit as a subtask along with the main task of predicting a hemodynamic intervention. We also find that including prior knowledge into the model, either through the adjacency matrix of the PKL model or through a multitask framework results in significantly better model performance compared to single task model like the naive MoE without constraints.

Table 2: Model performance on stepdown unit subgroup. Prevalence=14%.

Model	AUC	AP	PPV	Sp
PKL	0.746	0.102	0.172	0.979
MTL	0.75	0.09	0.168	0.975
MoE	0.739	0.096	0.198	0.980

On patient subgroups with Pressors, PRBC, or Fluid interventions, Table 3 reveals the need for careful inclusion of prior knowledge into the model. The primary task was to predict the onset of a hemodynamic intervention, which includes the the intervention type in the label. The subtasks are too similar to the main task to yield significantly better performances on the intervention type subgroups. The PKL, MTL, and naive MoE model without constraints all perform similarly well across all subgroups.

Finally, we compare the model performance between single-task (MoE), MTL, and PKL models on the primary task of predicting hemodynamic interventions. Notably, the PKL model performance is equivalent to the MTL model in predicting hemodynamic interventions when using the intervention types as subtasks in the MTL formulation or encoding the intervention types in the adjacency matrix in the PKL formulation (Table 4). The MTL model includes deep non-linear interactions with a hidden state representation of the data that is difficult to interpret. By contrast, the PKL and MoE models are trained with only pairwise interactions to preserve interpretability as described in section 3.3. Therefore, it is also noteworthy that the models perform equally well on this task despite the differences in model complexity.

Table 3: Model performance on intervention type subgroups at the breakeven point.

Model	Intervention type	AUC	AP	PPV	Sp
PKL	Pressors	0.892	0.586	0.551	0.957
	PRBC	0.892	0.227	0.279	0.979
	Fluids	0.723	0.140	0.202	0.960
MTL	Pressors	0.889	0.578	0.55	0.957
	PRBC	0.901	0.245	0.296	0.98
	Fluids	0.721	0.128	0.195	0.959
MoE	Pressors	0.890	0.584	0.551	0.957
	PRBC	0.891	0.227	0.271	0.979
	Fluids	0.725	0.136	0.200	0.960

Table 4: Predicting hemodynamic interventions. MTL and PKL have similar performance.

Model	AUC	AP	PPV	Sp
XGBoost	0.836	0.571	0.538	0.918
XGAM	0.839	0.575	0.543	0.919
MTL	0.843	0.586	0.552	0.921
MoE	0.844	0.586	0.553	0.921
PKL	0.844	0.587	0.555	0.921

5.4. Cohort discovery: Phenotypes of respiratory compromise

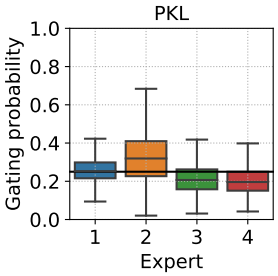
Patients with respiratory compromise experience disease progression at different rates largely due to the heterogeneity of respiratory illness. Timely and accurate risk stratification in the ICU can potentially improve outcomes. We explore phenotyping using the PKL model where the adjacency matrix is designed using the raw feature values. Individuals that share common physiological characteristics based on vital signs and laboratory measurements are grouped into the same expert. This enables the model to discover cohorts within the data that are pertinent to the task.

We trained a model with 4-experts using the raw feature inputs to group similar patients into the same expert. The overall model AUC in predicting sustained hypoxemia was 0.948 (AP: 0.757, PPV: 0.693, Sp: 0.962). Table 5 characterizes the clinical outcomes for patients in each phenotype discovered by the model. We adopt the term phenotype to refer to experts in the PKL model where individual samples are assigned to the expert with the highest gating probability. Phenotype 1 are stable patients with 1.2% of these patients developing sustained hypoxemia. Phenotype 2 are also mostly stable patients (6.2% develop sustained hypoxemia) but but are being measured with an invasive arterial line (69.4%). Phenotype 3 and 4 contain 34.1% and 39.1% hypoxemia patients, respectively, and represent the subgroup with the most severe outcomes including, days on invasive mechanical ventilation, ICU length of stay, and hospital mortality.

We chose a simple method to demonstrate the utility of the PKL model to discover cohorts by assigning patients to experts with the highest gating probability. Figure 4 shows that Expert 2 gating probabilities had higher variability compared to other phenotypes —

these were hard to classify patients with low BP but not respiratory compromise. Gating probabilities for stable patients in Phenotype 1 had higher confidence and were easier to classify.

Figure 4: Distribution of gating probabilities.



We next wanted to understand the biological characteristics that distinguished each phenotype. This type of analysis is particularly meaningful using the PKL model since samples are grouped by the similarity of their vital signs and laboratory measurements. We do this by examining the mean values of the variables used in the model for each phenotype. Figure 5 shows the continuous variables for the four phenotypes discovered by the PKL model, sorted by values in phenotype 1 - the most stable cohort. Phenotype 1 was characterized by high oxygen saturation, normal blood pressure, low lactate levels, and normal pH. Phenotype 2 are patients with low blood pressure but otherwise normal arterial blood gas variables. Phenotype 3 are patients with hemodynamic instability and respiratory compromise. They have shock (low blood pressure, high heart rate, high shock index), low hematocrit and hemoglobin counts, and exhibit signs of respiratory compromise with high respiratory rate, low arterial oxygen saturation, and abnormally low pH and base excess. Lactate and bicarbonate were normal in phenotype 3. Compared to phenotype 3, phenotype 4 had abnormally high PaCO2, high bicarbonate, elevated base excess, and very low arterial oxygen saturation. In contrast to phenotype 3, phenotype 4 patients have normal blood pressure, heart rate, hemoglobin, and hematocrit values suggesting phenotype 4 is not at risk of shock.

To summarize the discovered phenotypes:

- Phenotype 1: Stable patients with normal physiology and low risk of developing hypoxemia
- Phenotype 2: Low blood pressure with invasive arterial line but no signs of respiratory compromise.
- Phenotype 3: Severe subgroup with respiratory compromise and shock.
- Phenotype 4: Respiratory compromise but no-shock.

If we compare the supervised clustering provided by the PKL model to an unsupervised gaussian mixture model (GMM), we clearly see the benefits of the supervised approach. Figure 6 shows the differences in standardized values of each continuous variable by phenotype

Table 5: Clinical outcomes in phenotypes of respiratory compromise. The table shows counts and percentage of samples in each phenotype for categorical variables, otherwise median and interquartile range for numeric variables.

Characteristic	Phenotype 1	Phenotype 2	Phenotype 3	Phenotype 4
Patients	14441	5028	3755	2665
Hypoxemia	175 (1.2)	327 (6.5)	1282 (34.1)	1042 (39.1)
Age (y)	66.0 (56.0, 75.0)	59.0 (50.0, 70.0)	60.0 (51.0, 69.0)	52.0 (40.0, 66.0)
ICU LOS (days)	1.1 (0.7, 2.2)	0.8 (0.5, 1.7)	2.0 (0.8, 4.5)	2.8 (1.1, 6.1)
Hospital Mortality	1343 (9.4)	764 (15.4)	1901 (51.3)	993 (37.7)
APACHE	64.0 (49.0, 85.0)	63.0 (47.0, 88.0)	91.0 (67.0, 118.0)	81.0 (58.0, 104.0)
Invasive ventilation (hours)	20.4 (9.6, 57.6)	17.8 (8.1, 50.7)	74.5 (26.5, 209.6)	112.2 (34.7, 311.8)
Invasive A-line	6318 (43.8)	3487 (69.4)	2056 (54.8)	1533 (57.5)
Sepsis ICD-9	386 (2.7)	177 (3.5)	454 (12.1)	289 (10.9)
Hypoxemia ICD-9	249 (1.7)	132 (2.6)	373 (9.9)	324 (12.2)
Trauma ICD-9	156 (1.1)	93 (1.9)	256 (6.8)	156 (5.9)
Respiratory failure ICD-9	146 (1.0)	212 (4.2)	925 (24.7)	826 (31.0)
Pneumonia ICD-9	949 (6.6)	345 (6.9)	849 (22.6)	727 (27.3)
COPD ICD-9	711 (4.9)	222 (4.4)	334 (8.9)	279 (10.5)
CHF ICD-9	35 (0.2)	33 (0.7)	191 (5.1)	215 (8.1)

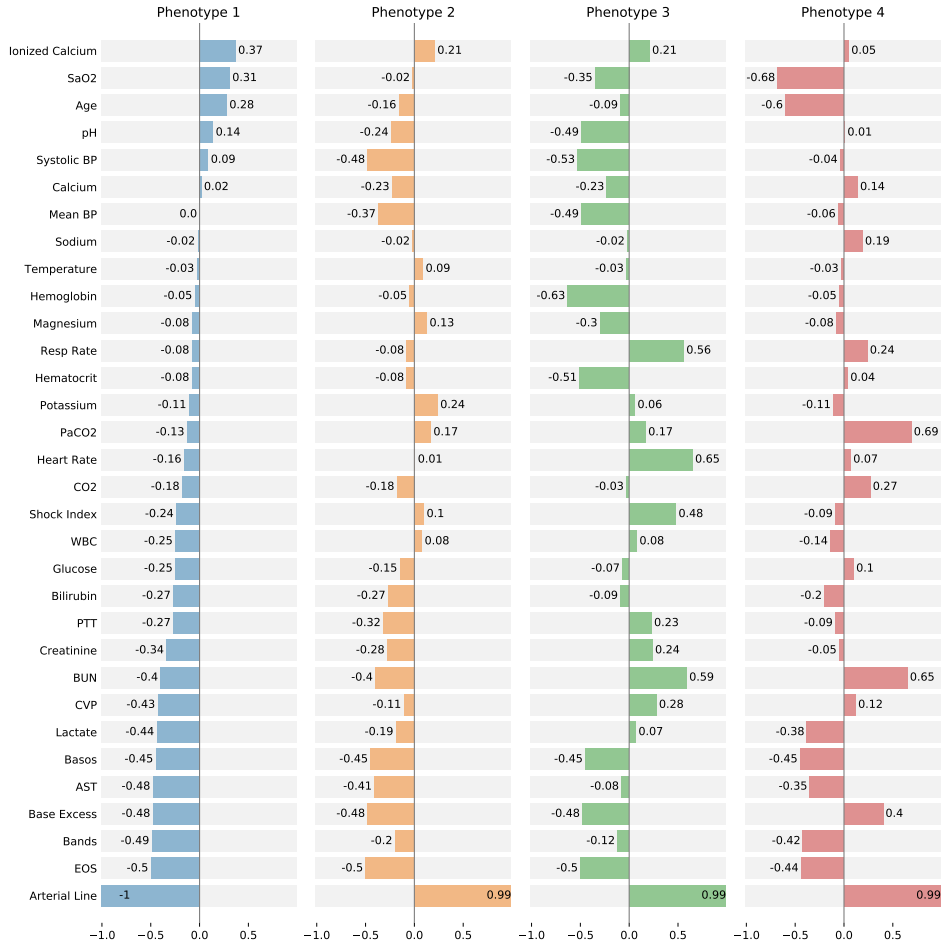
for a GMM model with 4 clusters. Phenotype 3 is possibly the only clinically interesting group because these patients exhibit signs of hypotension and shock with low blood pressure, low hemoglobin and hematocrit, and high heart rate. However, we don't observe a group with clear signs of respiratory compromise like in the PKL model. Comparing Figure 5 with Figure 6 shows that phenotypes discovered with the PKL model are more clinically meaningful with a group of patients that have poor oxygenation exhibited by low SaO2 and high PaCO2. The phenotypes discovered by the PKL model are clearly separated into a stable group (phenotype 1), a group with poor perfusion (phenotype 2), a group with poor oxygenation (phenotype 4), and a severely ill group with both poor perfusion and oxygenation (phenotype 3). The unsupervised approach does not capture these clinically meaningful phenotypes.

6. Discussion

We presented an approach to encode prior knowledge into a neural network and demonstrated that experts specialize according to the meta features used to design the adjacency matrix. Our method is an alternative to MTL with the gating network of a mixture-of-experts model learning clinically meaningful representation of the data through deliberate clustering, in contrast to MTL where useful representations are learned only as a side effect of solving a multi-label classification problem. We assume that only the primary task is relevant in the MTL framework and the sub-tasks are only used to improve the representation learning capacity of the network.

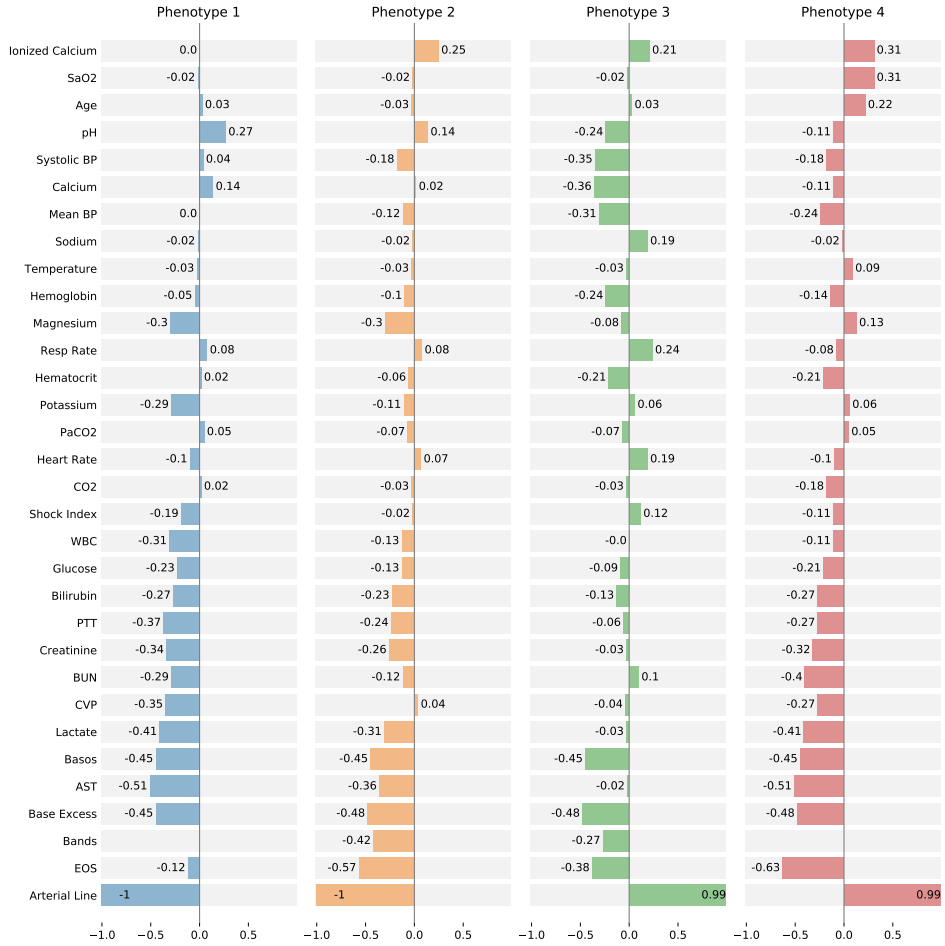
Why does PKL work? To understand why PKL works, we need to examine the underlying mechanisms. PKL is a mixture-of-experts with a regularization on the gating network to approximate a specified adjacency matrix. 1) MoEs with sparse gating probabilities en-

Figure 5: Differences in standardized values of each continuous variable by phenotype in the PKL model. Values are scaled to have mean zero and unit standard deviation.



courage localization by allocating a new case to a small number of experts. The experts are considered local in that the contribution of one expert is decoupled from the other experts, although the weights of all experts are learned through joint optimization – learned globally, acts locally. With the added adjacency matrix regularization, the assignment of samples to experts is further localized based on prior knowledge (see Stepdown unit experiments in Section 5.2). 2) PKL is a form of multilevel modeling that learns risk functions on subgroups of the data (i.e. random effects (Gelman and Hill (2006)), stratified models (Tuck et al. (2019))). A practical example: blood oxygenation of 100% is expected in ventilated patients with high FiO2 and a model would typically learn to assign high risk to SpO2 at 100%. However, non-ventilated patients would be considered healthy with SpO2 at 100% and should be assigned a low risk. When ventilation status is used to stratify patients in the adjacency matrix, experts in the PKL model would learn separate weights for SpO2 on the ventilated and non-ventilated groups, similar to classic multilevel modeling.

Figure 6: Gaussian Mixture Model unsupervised clustering. Differences in standardised values of each continuous variable by phenotype. Values are scaled to have mean zero and unit standard deviation. Compare with PKL model in Figure 5.



Phenotyping The respiratory phenotypes discovered by the PKL model in section 5.4 represent groups of patients that share similar physiological traits. The underlying disease heterogeneity is reflected in the percent of patients with hypoxemia in each phenotype. Table 5 shows that patients with Phenotype 4 are more likely to develop hypoxemia (39%), compared to the less severe Phenotypes 2 (6.5%) and 3 (34.1%). The phenotypes discovered by the PKL model reflect the known subphenotypes of ARDS, which are characterized by shock and metabolic acidosis (Calfee et al. (2018, 2014); Famous et al. (2017)). Comparing ICD-9 codes between Phenotypes 3 and 4 show high prevalence of sepsis patients in Phenotype 3 and higher prevalence of hypoxemia, respiratory failure, and pneumonia in Phenotype 4. Sepsis often manifests with physiological signs of shock. Figure 5 reinforces the distinction between Phenotypes 3 and 4, specifically, Phenotype 3 is characterized by low hemoglobin, hematocrit, high heart rate, and low blood pressure – signs of lack of

perfusion. In contrast Phenotype 4 patients have normal hematocrit, hemoglobin, blood pressure, and heart rate suggesting the underlying disease mechanism may be unrelated to perfusion.

Embedding Decision Paths from Boosted Trees The expert model we used combines gradient boosted decision trees (GBDT) with neural networks. GBDT algorithms, like XGBoost and LightGBM, are ensemble models of decision trees, which are trained in sequence. In each iteration, GBDT learns the decision trees by fitting the negative gradients (also known as residual errors). The decision paths to leaf nodes represent an interpretable and discriminative rule set that we find are very powerful features. For example, $\alpha\mathbb{I}[x_{\text{age}} > \tau_{\text{age}}]\mathbb{I}[x_{\text{lactate}} > \tau_{\text{lactate}}]$ represents a depth–2 tree along the decision path selecting age and lactate, where τ is a learned threshold and α is the risk assigned to the interaction. α is typically the value at the terminal leaf node of a decision tree. There are $2^M T$ potential decision paths leading to a terminal node for a GBDT model trained with T rounds of boosting and depth–M trees. We treat each decision path as a categorical entity and embed these entities in a neural network to re-learn the risks (α) at the terminal nodes. The learning procedure maintains the interpretability of tree based models but uses neural network components to refine the risk values assigned to each leaf node in the original trees.

By treating the learned decision paths as features, α effectively assigns a risk to each leaf node of the decision tree, similar to Wang et al. (2018). This procedure is unlike model stacking, which uses predictions from a base model to train subsequent models (Sill et al. (2009)). Embedding decision paths is also unlike decision tree distillation, which attempts to represent the leaves as nodes on a neural network (Ke et al. (2019)). Also in contrast to He et al. (2014), where leaf indices were treated as categorical variables in a logistic regression, we treat the decision paths as part of a vocabulary to be embedded in a multi–dimensional space. Embedding the decision paths reveals intrinsic continuity of the data by putting similar decision paths close to each other (Guo and Berkhahn (2016)). Maintaining a growing vocabulary of decision paths also enables transfer learning across tasks, which is particularly useful in the small–data setting. In practice, we see consistent gains in model performance over the original GBDT model from which the decision paths were extracted and we find significant improvement over training neural networks with the raw inputs.

Future work A possible future direction would be to use clinicians to derive rules that can be used as meta features to stratify patients for a truly hybrid expert-augmented machine learning model. The gating network would learn to imitate clinician knowledge in subgrouping patients while the expert networks specialize on the task within each subgroup. Another extension of this work is to the time series domain by using a recurrent neural network as the gating network to encode the clinical time series.

Limitations The number of experts in the MoE model is a nuanced hyperparameter and one of the areas this work can be improved upon is a more rigorous method to select the number of experts. Additionally, we take the argmax of the gating probabilities as a pseudo-cluster assignment, which has some limitations. Specifically, on difficult cases where the gating probabilities are uniform and the gating network is uncertain about the cluster assignment. An alternative to the softmax gating is the noisy top-K gating introduced in Shazeer et al. (2017) which can solve this problem. We experimented with the noisy top-K gating and found that it can be easily substituted-in for the softmax gating. By

setting the number of selected experts to 1 will force the gating network to select a single expert for each sample. This extreme specialization through hard partitioning comes at the cost of worse model performance compared to the soft partitioning given by the softmax transformation. Although the lower bound on model performance using XGAM experts is the accuracy given by the XGBoost baseline.

References

Jonathan Baxter. A Bayesian/Information Theoretic Model of Learning to Learn via Multiple Task Sampling. *Machine Learning*, 28(1):7–39, July 1997. ISSN 1573-0565. doi: 10.1023/A:1007327622663.

Emmanuel Bengio, Pierre-Luc Bacon, Joelle Pineau, and Doina Precup. Conditional Computation in Neural Networks for faster models. *arXiv:1511.06297 [cs]*, January 2016.

Carolyn S Calfee, Kevin Delucchi, Polly E Parsons, B Taylor Thompson, Lorraine B Ware, and Michael A Matthay. Subphenotypes in acute respiratory distress syndrome: Latent class analysis of data from two randomised controlled trials. *The Lancet Respiratory Medicine*, 2(8):611–620, August 2014. ISSN 22132600. doi: 10.1016/S2213-2600(14)70097-9.

Carolyn S. Calfee, Kevin L. Delucchi, Pratik Sinha, Michael A. Matthay, Jonathan Hackett, Manu Shankar-Hari, Cliona McDowell, John G. Laffey, Cecilia M. O'Kane, and Daniel F. McAuley. ARDS Subphenotypes and Differential Response to Simvastatin: Secondary Analysis of a Randomized Controlled Trial. *The Lancet. Respiratory medicine*, 6(9):691–698, September 2018. ISSN 2213-2600. doi: 10.1016/S2213-2600(18)30177-2.

Rich Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, July 1997. ISSN 1573-0565. doi: 10.1023/A:1007379606734.

Bryan Conroy, Larry Eshelman, Cristhian Potes, and Minnan Xu-Wilson. A dynamic ensemble approach to robust classification in the presence of missing data. *Machine Learning*, 102(3):443–463, March 2016. ISSN 1573-0565. doi: 10.1007/s10994-015-5530-z.

Larry J. Eshelman, Minnan Xu-Wilson, Abigail A. Flower, Brian Gross, Larry Nielsen, Mohammed Saeed, and Joseph J. Frassica. A Methodology for Evaluating the Performance of Alerting and Detection Algorithms Running on Continuous Patient Data. Preprint, Bioinformatics, September 2017.

Katie R. Famous, Kevin Delucchi, Lorraine B. Ware, Kirsten N. Kangelaris, Kathleen D. Liu, B. Taylor Thompson, and Carolyn S. Calfee. Acute Respiratory Distress Syndrome Subphenotypes Respond Differently to Randomized Fluid Management Strategy. *American Journal of Respiratory and Critical Care Medicine*, 195(3):331–338, February 2017. ISSN 1073-449X. doi: 10.1164/rccm.201603-0645OC.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press, 2006. ISBN 1-139-46093-5.

Cheng Guo and Felix Berkhahn. Entity Embeddings of Categorical Variables. *arXiv:1604.06737 [cs]*, April 2016.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):1–18, June 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9.

Xinran He, Stuart Bowers, Joaquin Quiñonero Candela, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, and Ralf Herbrich. Practical Lessons from Predicting Clicks on Ads at Facebook. In *Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining - ADKDD'14*, pages 1–9, New York, NY, USA, 2014. ACM Press. ISBN 978-1-4503-2999-6. doi: 10.1145/2648584.2648589.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Siddhant M. Jayakumar, Wojciech M. Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. Multiplicative Interactions and Where to Find Them. In *International Conference on Learning Representations*, September 2019.

Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214, 1994.

David C. Kale, Zhengping Che, Mohammad Taha Bahadori, Wenzhe Li, Yan Liu, and Randall Wetzel. Causal Phenotype Discovery via Deep Networks. *AMIA Annual Symposium Proceedings*, 2015:677–686, November 2015. ISSN 1942-597X.

Guolin Ke, Zhenhui Xu, Jia Zhang, Jiang Bian, and Tie-Yan Liu. DeepGBM: A Deep Learning Framework Distilled by GBDT for Online Prediction Tasks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 384–394, Anchorage, AK, USA, July 2019. Association for Computing Machinery. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330858.

Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLoS ONE*, 8(6):e66341, June 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0066341.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):1–13, September 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.178.

Sebastian Ruder. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv:1706.05098 [cs, stat]*, June 2017.

Peter Schulam, Fredrick Wigley, and Suchi Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *arXiv:1701.06538 [cs, stat]*, January 2017.

Joseph Sill, Gabor Takacs, Lester Mackey, and David Lin. Feature-Weighted Linear Stacking. *arXiv:0911.0460 [cs]*, November 2009.

Harini Suresh, Jen J. Gong, and John Guttag. Learning Tasks for Multitask Learning: Heterogenous Patient Populations in the ICU. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 802–810, July 2018. doi: 10.1145/3219819.3219930.

Jonathan Tuck, Shane Barratt, and Stephen Boyd. A Distributed Method for Fitting Laplacian Regularized Stratified Models. *arXiv:1904.12017 [cs, math, stat]*, November 2019.

Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. TEM: Tree-enhanced Embedding Model for Explainable Recommendation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 1543–1552, Lyon, France, 2018. ACM Press. ISBN 978-1-4503-5639-8. doi: 10.1145/3178876.3186066.

Yu Zhang and Qiang Yang. A Survey on Multi-Task Learning. *arXiv:1707.08114 [cs]*, July 2018.