PHYSIOLOGICAL SIGNAL EMBEDDINGS (PHASE) VIA INTERPRETABLE STACKED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In health, machine learning is increasingly common, yet neural network embedding (representation) learning is arguably under-utilized for physiological signals. This inadequacy stands out in stark contrast to more traditional computer science domains, such as computer vision (CV), and natural language processing (NLP). For physiological signals, learning feature embeddings is a natural solution to data insufficiency caused by patient privacy concerns – rather than share data, researchers may share informative *embedding models* (i.e., representation models), which map patient data to an output embedding. Here, we present the PHASE (PHysiologicAl Signal Embeddings) framework, which consists of three components: i) *learning neural network embeddings* of physiological signals, ii) *predicting outcomes* based on the learned embedding, and iii) *interpreting the prediction results* by estimating feature attributions in the "stacked" models (i.e., feature embedding model followed by prediction model). PHASE is novel in three ways: 1) To our knowledge, PHASE is the first instance of transferal of neural networks to create physiological signal embeddings. 2) We present a tractable method to obtain feature attributions through stacked models. We prove that our stacked model attributions can approximate Shapley values – attributions known to have desirable properties – for arbitrary sets of models. 3) PHASE was extensively tested in a cross-hospital setting including publicly available data. In our experiments, we show that PHASE *significantly outperforms* alternative embeddings – such as raw, exponential moving average/variance, and autoencoder – currently in use. Furthermore, we provide evidence that *transferring neural network embedding/representation learners* between distinct hospitals still yields performant embeddings.

1 INTRODUCTION

Learning embeddings (i.e., representation learning) (Bengio et al., 2013) has been applied to medical images and clinical text (Tajbakhsh et al., 2016; Ravishankar et al., 2016; Lv et al., 2014) but has been under-explored for time series physiological signals in electronic health records. This paper introduces the PHASE (PHysiologicAl Signal Embeddings) framework to learn embeddings of physiological signals (Figure 1a), which can be used for various prediction tasks (Figure 1b), helps interpretation by computing feature attributions of the original features (i.e., not embeddings) for a prediction result in a tricky "stacked" model situation (i.e., embedding model followed by prediction model) (Figure 1c), and has been extensively tested in terms of its transferability using data from multiple hospitals (Figure 1d).

Based on computer vision (CV) and natural language processing (NLP), exemplars of representation learning, physiological signals are well suited to embeddings. In particular, CV and NLP share two notable traits with physiological signals. The first is *consistency*. For CV, the domain has consistent features: edges, colors, and other visual attributes. For NLP, the domain is a particular language with semantic relationships consistent across bodies of text. For sequential signals, physiological patterns are arguably consistent across individuals. The second attribute is *complexity*. Across these three domains, each particular domain is sufficiently complex such that learning embeddings is non-trivial. Together, consistency and complexity suggest that for a particular domain, every research group independently spends a significant time to learn embeddings that may ultimately be

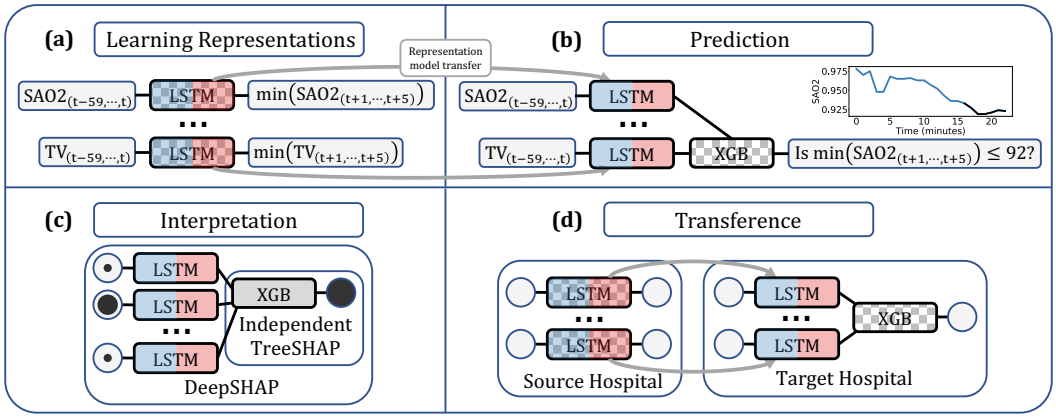

Figure 1: The PHASE framework, which consists of embedding learning, prediction, interpretation, and transference. The checkered patterns denote that a model is being trained in the corresponding stage, whereas solid colors denote fixed weights/models. The red side of the LSTM denotes the hidden layer we will use to generate embeddings. In (c), the size of the black circles on the left represent the feature attributions being assigned to the original input features.

quite similar. In order to avoid this negative externality, NLP and CV have made great progress on standardizing their embeddings; in health, physiological signals are a natural next step.

Furthermore, physiological signals have unique properties that make them arguably better suited to representation learning than traditional CV and NLP applications. First, physiological signals are typically generated in the health domain, which is constrained by patient *privacy concerns*. These concerns make sharing data between hospitals next to impossible; however, sharing models between hospitals is intuitively safer and generally accepted. Second, a key component to successful transfer learning is a *community* of researchers that work on related problems. According to Faust et al. (2018), there were at least fifty-three research publications using deep learning methods for physiological signals in the past ten years. Additionally, we discuss particular examples of neural networks for physiological signals in Section 2.2. These varied applications of neural networks imply that there is a large community of machine learning research scientists working on physiological signals, a community that could one day work collaboratively to help patients by sharing models.

Although embedding learning has many aforementioned advantages, it makes interpretation more difficult. Existing interpretation methods (Shrikumar et al., 2016; Sundararajan et al., 2017; Lundberg & Lee, 2017; Lundberg et al., 2018) do not work for models trained using learned embeddings, because they will assign attributions to the embeddings. Feature attributions assigned to embeddings will be meaningless, because the embeddings do not map to any particular input feature. Instead, each embedding is a complicated, potentially non-linear combination of the original raw physiological signals. In a health domain, inability to meaningfully interpret your model is unsatisfactory. Healthcare providers and patients alike generally want to know the reasoning behind predictions/diagnoses. Interpretability can enhance both scientific discovery as well as provide credibility to predictive models. In order to provide a principled methodology for mapping embedding attributions back into physiological signal attributions, we provide a proof that justifies PHASE's Shapley value framework in Section 3.3. This framework generalizes across arbitrary stacked models and currently encompasses differentiable models (e.g., linear models, neural networks) and tree-based models (e.g., gradient boosting machines and random forests).

In the following sections, we discuss previous related work (Section 2) and describe the PHASE framework (Section 3). In Section 4, we first evaluate how well our neural network embeddings make accurate predictions (Section 4.3.1). Second, we evaluate whether transferring these embedding learners still enables accurate predictions across three different hospitals separated by location and across hospital departments (Section 4.3.2). Lastly, we present a visualization of our methodology for providing Shapley value feature attributions through stacked models in Section 4.3.3.

2 RELATED WORK

2.1 REPRESENTATION LEARNING IN THE HEALTH DOMAIN

Representation learning (embedding learning) in health is growing more popular. One particularly natural subdomain is medical image analysis, e.g., mammography analysis, kidney detection in ultrasound images, optical coherence tomography image analysis, diagnosing pneumonia using chest X-ray images, lung pattern analysis, otitis media image analysis, and more (Arevalo et al., 2016; Ravishankar et al., 2016; Kermany et al., 2018; Liao et al., 2013; Christodoulidis et al., 2016; Shie et al., 2015). Outside of image analysis, additional examples of transfer learning in the medical domain include Lv et al. (2014), Wiens et al. (2014), Brisimi et al. (2018), Choi et al. (2017), Choi et al. (2016), and Che et al. (2016). Even within physiological signals, some examples of embedding learning are beginning to sprout up, including Wu et al. (2013), who utilize kNNs to perform transfer learning for brain-computer interaction. Comparatively, PHASE transfers neural networks as embedding functions learned in an unsupervised manner, where the embeddings provide a basis for training a model on any prediction task (as opposed to being tied to the prediction they were trained on).

2.2 NEURAL NETWORKS FOR PHYSIOLOGICAL SIGNALS

To our knowledge, our work is the first to transfer unsupervised deep neural networks for embedding sequential physiological signals, embeddings which are not tied to a particular prediction problem. One caveat is that supervised deep learning can be said to inherently learn embeddings. In physiological signals, there are several examples of particular supervised learning tasks with neural networks. Srinivasan et al. (2007) and Guo et al. (2010) both detect epilepsy from signals, Wilson & Russell (2003) utilize psycho-physiological measurements to assess mental workload, Wagner et al. (2005) and Chanel et al. (2006) utilize physiological signals to classify emotions, Koike & Kawato (1995) reconstruct human arm movement from EMG signals, Sullivan et al. (2010) reconstruct missing physiological signals, and Yang & Hsieh (2016) use acoustic signals to detect anomalies in heart sound. Based on this substantive community of research scientists working on physiological signals, there is a clear opportunity to unify independent research by appropriately using *unsupervised* feature embedding learning.

In the vein of embedding learning, Martinez et al. (2013) applied autoencoders to blood volume pulse and skin conductance measured from 36 people playing a video game and used the encodings to predict affective state. In their paper, the sample size is fairly small and reflects that their primary objective was to perform feature extraction and feature selection. In contrast, PHASE evaluates transferring embedding learners (i.e., feature extractors) across multiple hospitals (Table 1).

2.3 FORECASTING FOR OPERATING ROOM DATA

Lundberg et al. (2017) proposed an approach, namely Prescience, which achieved state-of-the-art hypoxemia predictions using operating room data, the same data we used to evaluate PHASE. Prescience utilizes gradient boosting machines (GBM) applied to features extracted by using traditional time series feature extraction methods – exponential moving average/variance embeddings. Prescience compares prediction models including gradient boosting machines, linear lasso, linear SVM, and a parzen window with the objective of forecasting low blood oxygen in the future. Ultimately, Lundberg et al. (2017) find the highest performing method to be GBM trees. With samples drawn from the same data set, PHASE seeks to substitute the feature extraction used in Prescience with a deep learning approach, which resulted in a better average precision compared to Prescience without the clinical text features (4 is Prescience and 12 is PHASE in Figure 3).

2.4 FEATURE ATTRIBUTIONS FOR INTERPRETABILITY

Interpretability of models has been addressed in numerous independent pieces of work (Shrikumar et al., 2016; Sundararajan et al., 2017; Lundberg & Lee, 2017; Lundberg et al., 2018). For our evaluation of interpretability, we choose to focus on Shapley values introduced by Lloyd Shapley, originally in the context of game theory (Shapley, 1953). Lundberg & Lee (2017) identify Shapley values as the only additive feature attribution method that satisfies the properties of local accuracy,

missingness, and consistency. For PHASE, our pipeline includes multiple models – GBM trees and a LSTM networks. Methods exist for obtaining Shapley values for GBM trees (TreeSHAP) and for neural networks (DeepLIFT/DeepSHAP) (Lundberg et al., 2018; Shrikumar et al., 2016; Lundberg, 2018). However, the default version of these methodologies do not beget a theoretically justified approach for propagating attributions through multiple models. In fact, to the authors' knowledge, a tractable method for obtaining Shapley value attributions for stacked models does not exist. In this paper, we utilize versions of TreeSHAP and DeepLIFT that create single reference attributions that can be composed to address stacked models. At the end of obtaining many attributions, the average of these attributions approximates the Shapley value attributions (more details in Section 3.3).

3 OUR APPROACH: PHYSIOLOGICAL SIGNAL EMBEDDINGS (PHASE)

Taking inspiration from sinusoidal waveforms, we name our methodology PHASE. In the PHASE framework, the first step is to learn neural network embeddings for physiological signals (Figure 1a). The second step is to predict outcomes based on the learned feature embedding (as in Figure 1b), potentially across multiple hospitals (as in Figure 1d). Finally, the last step is to interpret the prediction results by estimating feature attributions through the models trained in the first two steps (as in Figure 1c).

3.1 LEARNING EMBEDDINGS - LONG SHORT TERM MEMORY (LSTM) NETWORKS

PHASE uses LSTM networks to learn feature embeddings from time series physiological data. LSTMs are a popular variant on recurrent neural networks introduced by Hochreiter & Schmidhuber (1997). They have the capacity to model long term dependencies, while avoiding the vanishing gradient problem (Pascanu et al., 2012). We utilize LSTMs with forget gates, introduced by Gers et al. (2000), implemented in the Keras library with a Tensorflow back-end. We train our networks with either regression or classification objectives. For regression, we optimize using Adam with an MSE loss function. For classification we optimize using RMSProp with a binary cross-entropy loss function. In both cases we utilize dropout and recurrent dropout for regularization. Our models consist of two hidden layers, each with 200 LSTM cells with dense connections between all layers.

For PHASE, we first train a univariate LSTM on each physiological signal $\mathcal{P}$, predicting the minimum of $\mathcal{P}$ in the future five minutes (Figure 1a). Note that we choose the minimum of the next five minutes because we care about forecasting adverse outcomes. Then, we obtain hidden embeddings of the original physiological signals by passing them through to the hidden layer (the red layer in Figure 1a). These embeddings are unsupervised in the sense that training them simply requires the same feature (albeit at different time steps). We find that the completely unsupervised alternative of an LSTM autoencoder is significantly less performant than the LSTM trained to predict the minimum of the next five minutes (8 is autoencoder and 12 is PHASE in Figure 3).

One reason behind having *univariate* neural networks is for transference. By using univariate networks, the input to the final prediction model may be any set of physiological signals with existing embedding learners. This is especially useful because hospital departments have substantial variation in the signals they may choose to collect for features. Another reason for univariate networks is that data in a single hospital is often collected at different points in time, or new measurement devices may be introduced to data collection systems. For traditional pipelines, it may be necessary to re-train entire machine learning pipelines when new features are introduced. With univariate networks, the flexibility would mean pre-existing embedding learners would not necessarily need to be re-trained.

3.2 PREDICTION - GRADIENT BOOSTING MACHINES (GBM)

PHASE can use any prediction model. In this paper, we focus on gradient boosting machine trees because the Prescience method found that they outperform several other models in the operating room data (Lundberg et al., 2017). Gradient boosting machines were introduced by Friedman (2001). This technique creates an ensemble of weak prediction models in order to perform classification/regression tasks in an iterative fashion. In particular, we utilize XGBoost, a popular implementation of gradient boosting machines that uses additive regression trees (Chen & Guestrin, 2016). XGBoost often dominates in Kaggle, a platform for predictive modeling competitions. In

particular, seventeen out of twenty nine challenge winning solutions used XGBoost in 2015 (Chen & Guestrin, 2016). For PHASE, we postulate that utilizing embeddings of time series signals provides stronger features for the ultimate prediction with XGB (as visualized in Figure 1b).

3.3 INTERPRETATION - SHAPLEY VALUES THROUGH STACKED MODELS

PHASE addresses an inherent challenge in the *interpretation* of an embedding model (or feature representation model). Estimating feature attributions is a common way to make a prediction result interpretable. At a high level, the goal is to explain how much each feature matters for a particular model's prediction. However, this goal is only meaningful if the model being explained uses features with a natural human interpretation. For example, if we interpret PHASE's GBM model, which takes the embeddings as input and outputs a prediction, our feature attributions will be assigned to the embeddings, which are not meaningful to doctors or patients.

The answer is to extend the prediction model (here, a GBM) by combining the feature embedding model (here, a LSTM network), which makes a "stacked" model (Figure 1c). Since the original features in the embedding stage are meaningful, one solution is to utilize a model agnostic feature attribution method over the "stacked" model. For our attributions, we aim to provide Shapley values, but unfortunately the exact model agnostic computation has an exponential computational complexity ($O(N2^M)$, where N is the sample size and M is the number of features) (Shapley, 1953). In response, one might want to use a model-specific method of computing approximate Shapley values – which often gains speed by using knowledge of the model. However, to the authors' knowledge, there was previously no known model-specific method to estimate Shapley values for a stack comprised of LSTMs and a GBM.

Single reference Shapley values: Our new method for estimating Shapley values for the aforementioned stacked model (i.e., LSTMs and GBM), requires adaptations on two existing feature attributions methods. First is DeepSHAP, a variant on DeepLIFT – a feature attribution method for neural networks (Lundberg, 2018; Shrikumar et al., 2016). DeepSHAP differs from DeepLIFT in that it can find attributions for single references. Both methods can be written as modifications to a traditional backward pass through a neural network (Ancona et al., 2018). Since the computational complexity of a backward pass is the same as a forward pass through the network, we can consider this cost "low". The second method we utilize is "Independent TreeSHAP". This method is a variation on normal TreeSHAP (Lundberg et al., 2018), but it can be computed for single references. Independent TreeSHAP has a computational complexity of $O(MLT)$, where L is the maximum number of leaves in any given tree and T is the number of trees in the GBM.

Combining "stacked" model Shapley values: Combining these two methods amounts to treating the "stacked" model (Figure 1c) as a larger neural network and applying DeepSHAP to pass back attributions as gradients at each layer (Ancona et al., 2018). However, at the GBM layer we obtain the appropriate gradients by dividing the Independent TreeSHAP Shapley values by the difference between the sample and the references. According to Theorem 1, we can then average over these single reference attributions for an approximation to the Shapley values.

Generalizability: Note that Theorem 1 also implies that for any arbitrary set of models in a stack, if single reference Shapley values are obtainable for each model, the Shapley values for the entire stack can be obtained. Because the single reference Shapley value methods are known for neural networks and for trees, any "stacked" model composed of these two methods can be explained. Worth noting is that many embedding/prediction models can be represented as neural networks, making our framework to attribute "stacked" models fairly general.

Theorem 1. *Computing the average over single reference Shapley values approaches the true Shapley values.*

Proof. Starting with the definition of Shapley values:

$$\begin{aligned}\phi_i &= \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \big(f_{S \cup \{i\}}(x) - f_S(x)\big) \\ &= \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \big(\mathbb{E}_{\mathcal{D}}[f(x)|x_{S \cup \{i\}}] - \mathbb{E}_{\mathcal{D}}[f(x)|x_S]\big)\end{aligned}$$

where $\mathcal{D}$ is the data distribution, F is the set of all features, and f is our model. Rewriting the sum over all permutations of F, rather than over all combinations, the weighting term becomes one:

$$\begin{aligned}\phi_i &= \sum_{S_p \subseteq F \setminus \{i\}} \mathbb{E}_{\mathcal{D}}[f(x)|x_{S_p \cup \{i\}}] - \mathbb{E}_{\mathcal{D}}[f(x)|x_{S_p}] \\ &= \mathbb{E}_F[\mathbb{E}_{\mathcal{D}}[f(x)|x_{S_p \cup \{i\}}] - \mathbb{E}_{\mathcal{D}}[f(x)|x_{S_p}]] \\ &= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_F[f(x)|x_{S_p \cup \{i\}}] - \mathbb{E}_{N_p}[f(x)|x_{S_p}]]\end{aligned}$$

where the last step depends on independence between the permutations and the data generating mechanism. □

4 EXPERIMENTS

We first describe our data sets (Section 4.1), evaluation metric (Section 4.2), and the results of comparisons between PHASE and alternative approaches in various testing scenarios (Section 4.3).

4.1 DATA DESCRIPTION

Hospital 0/1 data was collected via the Anesthesia Information Management System (AIMS), which records all data measured in the operating room during surgery. Both medical centers are within the same city (within 10 miles of each other). Hospital P is a sub-sampled version of the publicly available MIMIC data set from PhysioNet, which contains data obtained from an intensive care unit in Boston, Massachusetts (Johnson et al., 2016). Hospital P data was collected several thousands of miles from the medical centers associated with hospital 0/1 data. More details about these hospitals are in Table 1.

Model	Department	# Procedures	# Samples	Base Rate
Hospital 0	OR	29,035	3,528,507	1.09%
Hospital 1	OR	28,136	3,751,163	2.18%
Hospital P	ICU	1,669	5,080,864	3.93%

Table 1: Statistics of the different data sources. Hospital P is a public data set (PhysioNet).

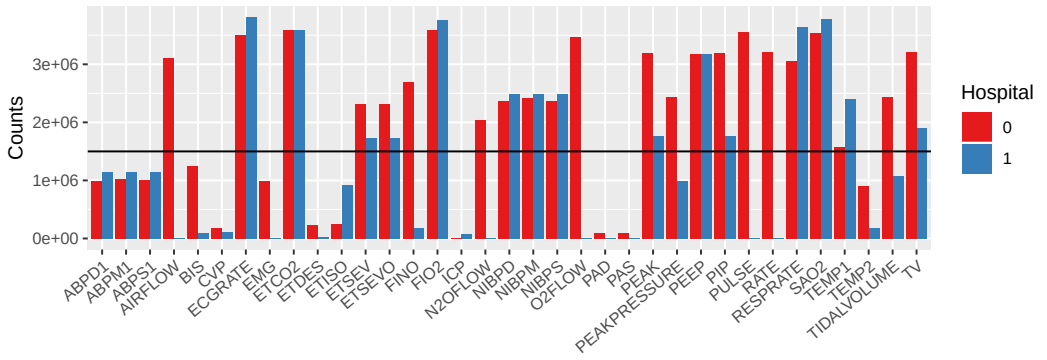

Figure 2: Counts of each feature across both AIMS hospitals. Fifteen features have more than 1.5 million counts for both hospitals (ECGRATE, ETCO2, ETSEV, ETSEVO, FIO2, NIBPD, NIBPM, NIBPS, PEAK, PEEP, PIP, RESPRATE, SAO2, TEMP1, TV).

The hospital 0/1 data includes static information (height, weight, age, sex, procedure codes), as well as real-time measurements of thirty-five physiological signals (e.g., SaO_2, FiO_2, $ETCO_2$, etc.) sampled minute by minute. Although the hospital P data contains several physiological signals sampled at a high frequency, we solely use a minute by minute SaO_2 signal. Any missing values in the data are imputed by the mean and each feature is normalized to have unit mean and variance.

4.2 EVALUATION METHODOLOGY

PHASE and alternative approaches are evaluated based on real-time prediction tasks, for example, whether a certain condition will occur in the next 5 minutes (details in Section 4.3). Our evaluation metric for prediction performance in binary classification is area under the precision-recall curve, otherwise known as *average precision* (AP). Rather than ROC curves, PR curves often better highlight imbalanced labels. Precision is defined as $\frac{tp}{tp+fp}$ and recall is $\frac{tp}{tp+fn}$, where tp is true positives, fp false positives, and fn false negatives. The area under the curve provides a summary statistic that balances both precision and recall.

4.3 RESULTS

In this section, we compare different embeddings of physiological signals as discussed in Table 2 (where Min^{h} represents PHASE). Based on these comparisons, we see the overall performance of predicting three adverse clinical events in Section 4.3.1 as well as a discussion of how well the embedding learners transfer between hospitals in Section 4.3.2. Lastly, in Section 4.3.3, we depict the attributions from our new model stacking method for obtaining Shapley values.

In the operating room, there are a few physiological signals that stand out as indicators of adverse outcomes. Three of these signals include SaO_2 (blood oxygen), $ETCO_2$ (end tidal CO2), and NIBPM (non-invasive blood pressure measurement), which are linked to our three adverse outcomes: hypoxemia, hypocapnia, and hypotension, respectively. Forecasting these outcomes is particularly important because deviations from the norm could spell disaster. For hypoxemia, the label at each time point is one if SaO_2 is 92 percent or lower in the future five minutes. Points where SaO_2 is currently below 92 percent are not considered. For hypocapnia, the label is one if $ETCO_2$ is less than or equal to 35 mm Hg in the future five minutes. Points where $ETCO_2$ is currently below 35 mm Hg are not considered. For hypotension, the label is one if NIBPM is less than or equal to 60 mm Hg in the future five minutes. Points where NIBPM is currently below 60 mm Hg are not considered.

Table 2: Notation for different embeddings.

Raw	Sixty minutes of raw signal.
EMA (Prescience)	Exponential moving averages as well as variances of each signal. Computed where weights decay with a half-life of 6 seconds, 1 minute, or 5 minutes.
Min^{h} (PHASE)	Hidden layer embedding from an LSTM trained to predict the minimum of the current signal five minutes into the future on hospital h's data.
Auto^{h}	Hidden layer embedding from an LSTM trained to predict an output signal identical to the input signal on hospital h's data.
Hypox^{h}	Hidden layer embedding from an LSTM trained to predict hypoxemia on hospital h's data.

4.3.1 PREDICTION PERFORMANCE

In Figure 3, we analyze the performance of XGB with different embeddings of the same signals across our three prediction tasks. In terms of pre-training for this experiment, there is none for the Raw and EMA embeddings. However for Min^{h} and Auto^{h}, we have trained fifteen univariate LSTM networks for both objectives across both hospitals (a total of sixty networks). We fix these same LSTM networks to generate hidden embeddings of the original signals across the three final prediction tasks of hypoxemia, hypocapnia, and hypotension.

In terms of performance, the first observation in Figure 3 is that average precisions points using all signals (in blue) are almost always significantly above their associated average precision points using only a single signal (SaO_2 for hypoxemia, $ETCO_2$ for hypocapnia, NIBPM for hypotension) (in red). These outcomes derived from forecasting the single signal (in red) are complex and benefit from having access to more signals. This suggests that the LSTMs in PHASE's embedding stage, despite being unsupervised and agnostic of the final prediction task, are still capable of learning meaningful information for prediction tasks that depend on other features.

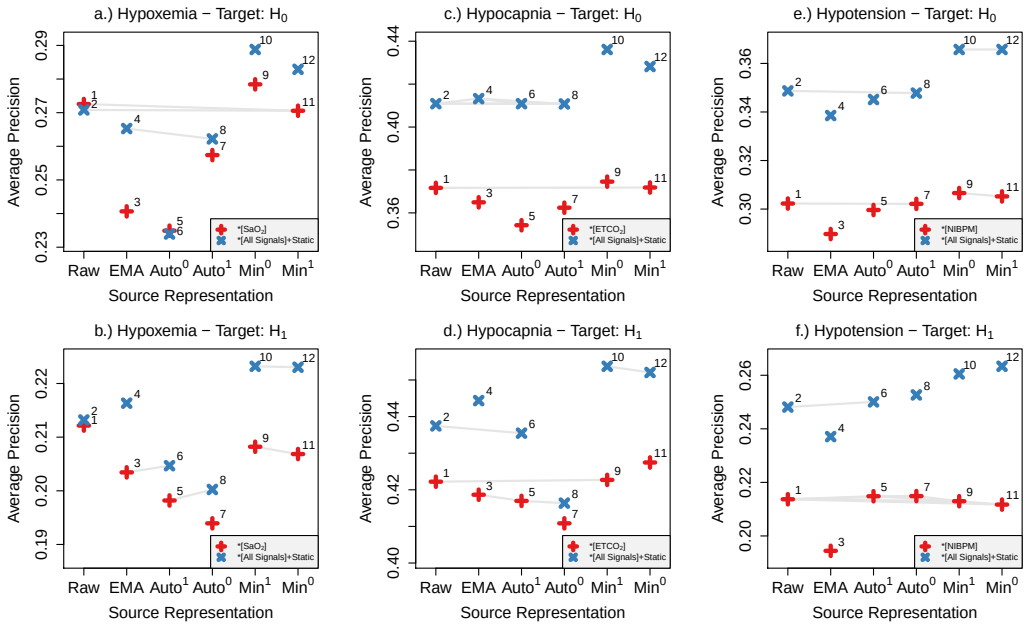

Figure 3: GBMs with different embeddings of physiological signals. Gray lines signify insignificant differences (all others pairs are significant at a p-value of 0.01) based on one hundred bootstraps of the test set with adjusted pairwise comparisons via ANOVA with Tukey's HSD test. For hypoxemia, we train XGB with a learning rate of 0.02, whereas for hypocapnia and hypotension, we found a learning rate of 0.1 to be sufficient for comparing embeddings. For all, we utilize the 15 features above the line in both hospitals (Figure 2). Notation described in Table 2, where Min^h represents PHASE. Note that $*[SaO_2]$ denotes that we have a Raw, EMA, or Min embedding of SaO_2 and $*$[All Signals]+Static denotes a Raw, EMA, or Min embedding of all the signals plus static variables.

Most importantly, the Min^h (PHASE) models (10 and 12) consistently outperform all other models in Figure 3 by a significant margin. Promisingly, we often see no significant difference between Min^0 and Min^1 irrespective of which hospital we are currently training on. This indicates that the version of PHASE that transfers it's LSTM embedding functions from a separate hospital does as well as the version of PHASE that purely learns embeddings without attempting to transfer.

4.3.2 Transference between Hospitals

First, we can look back to Figure 3. The feature embeddings learned in a source hospital that differs to the target hospital (12) performs significantly better than the EMA (Prescience presented in Lundberg et al. (2017)) and Raw embeddings (2 and 4) and generally on par with a matching source and target hospital (10). This is promising, because it suggests that the domain shift between hospitals 0 and 1 does not prevent physiological signal embeddings from transferring well.

In Figure 4, we can see that the Min^P embeddings obtained from the publicly available PhysioNet data creates a decent embedding for SaO_2 (2_p) with a similar AP to the EMA embedding. Furthermore, we can see that the Hypox^P embeddings (4_p) are on par with the regression embeddings using the target hospital as a source (10).

The Hypox^P embeddings are similarly obtained from the PhysioNet data with a classification objective rather than a regression one. One potential reason for the difference between the Min^P and Hypox^P classification results is that for binary classification we up-sample the hypoxemia examples which may modify the parameter space in comparison to the regression. Another potential reason is that for binary classification we select a model based on the validation average precision (whereas for the regression we select based on the validation loss). Loss is not scale invariant, therefore it is likely leads to higher variance in ultimate average precision evaluation. Because this effect was minimal in the case of hospital 0 and hospital 1, it may imply that the LSTM's prediction task may

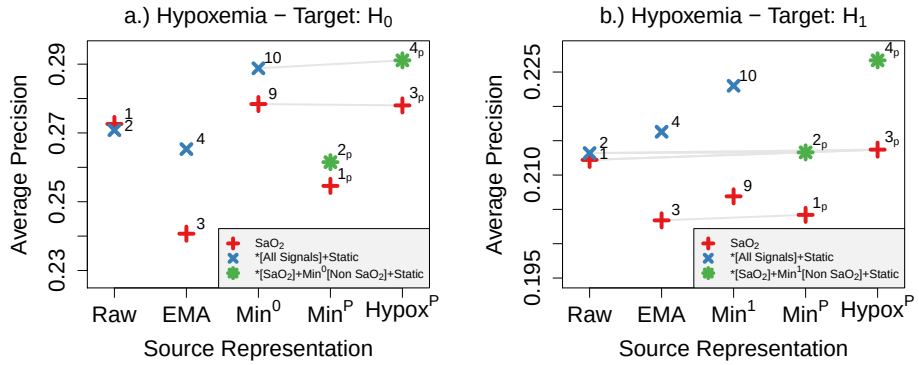

Figure 4: GBMs with different embeddings of physiological signals. Gray lines signify insignificant differences (all others are significant at a p-value of 0.01) based on one hundred bootstraps of the test set with adjusted pairwise comparisons via ANOVA with Tukey's HSD test. We train XGB with a learning rate of 0.1 using the 15 features above the line in both hospitals (Figure 2). Notation described in Table 2. The PhysioNet embeddings borrow signals from the target hospital's embeddings so $*[\text{SaO}_2] + \text{Min}^T[\text{Non SaO}_2] + \text{Static}$ denotes that we have a Min^P embedding of SaO_2, a Min^T embedding of the remaining 14 variables, where T is the target hospital, and static variables.

be more important for pairs of hospitals with greater degrees of domain shift. Ultimately, it appears that even across a domain shift that spans thousands of miles and distinct departments (operating room versus intensive care unit), we are able to obtain meaningful embedding functions via LSTM networks.

4.3.3 Interpretation for Stacked Models

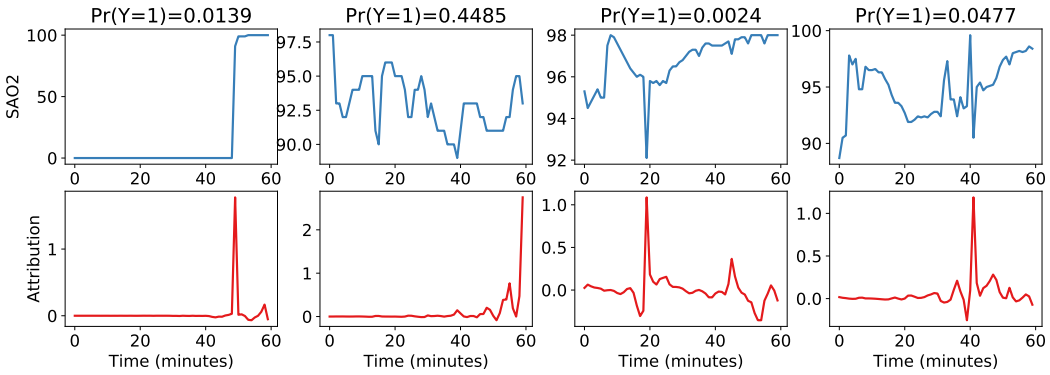

Figure 5: Attributions from stacked models composed of a univariate LSTM without it's final layer and a gradient boosting machine trained to predict hypoxemia (as in Figure 1c). The LSTM and the GBM were both trained in hospital 0. We averaged over 100 randomly sampled single reference attributions.

We address interpretation in Figure 5; based on visual inspection, these attributions appear to make sense. One rational property of these attributions is that time points with high local variance are correlated with large attribution values. In other words, variability in the signal indicates potential adverse outcomes. Furthermore, we can see that if the signal is low at the time steps closest to the prediction, the importance of the other minutes is comparatively washed out. Finally, although signals are typically zero at the beginning of surgery, these are not truly a cause for concern. These spurious signals are not picked up in our stacked model attributions, as we would hope. All in all, the attributions appear to be reasonably accurate to the contributions we might expect from the original time points.

5 CONCLUSION

This paper presents PHASE: a new approach to machine learning with physiological signals based on transferring embedding learners. PHASE has potentially far-reaching impacts, because neural networks inherently create an embedding before the final output layer. As discussed in Section 2.2, many research scientists are working on neural networks for physiological signals for supervised tasks. Based on our results, these researchers could train unsupervised versions of neural networks that could potentially become semi-private ways of sharing meaningful signals picked up in data sets. Utilizing univariate, unsupervised neural networks as embeddings not only allows the final prediction task to be any task, it allows for modularity in the embedding stage. With PHASE, modularity means that any set of inputs with available embedding functions can easily be composed for prediction. Furthermore, in our experiments we show that embeddings in physiological signals actually transfer with a great deal of success under three significant adverse clinical outcomes. Perhaps most importantly, this paper introduced a framework for passing feature attributions through these stacked models. By showing that Shapley values may be computed as the mean over single reference Shapley values, we ensure that this model stacking framework generalizes to all models with single reference Shapley value methods.

In the direction of future work, it is important to carefully consider representation learning in health – particularly in light of model inversion attacks as discussed in Fredrikson et al. (2015). To this end, future work in making precise statements about the privacy of models deserves attention, for which one potential avenue may be differential privacy (Dwork, 2008). Other important areas to explore include extending these results to higher sampling frequencies. Our data was sampled once per minute, but higher resolution data may beget different neural network architectures. Lastly, further work may include quantifying the relationship between domain shifts in hospitals and PHASE.

REFERENCES

Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations (ICLR 2018)*, 2018.

John Arevalo, Fabio A González, Raúl Ramos-Pollán, Jose L Oliveira, and Miguel Angel Guevara Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer methods and programs in biomedicine*, 127:248–257, 2016.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Theodora S Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International journal of medical informatics*, 112:59–67, 2018.

Guillaume Chanel, Julien Kronegg, Didier Grandjean, and Thierry Pun. Emotion assessment: Arousal evaluation using eegs and peripheral physiological signals. In *International workshop on multimedia content representation, classification and security*, pp. 530–537. Springer, 2006.

Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable deep models for icu outcome prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, pp. 371. American Medical Informatics Association, 2016.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1495–1504. ACM, 2016.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 787–795. ACM, 2017.

Stergios Christodoulidis, Marios Anthimopoulos, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Multi-source transfer learning with convolutional neural networks for lung pattern analysis. *arXiv preprint arXiv:1612.02589*, 2016.

Cynthia Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pp. 1–19. Springer, 2008.

Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: a review. *Computer methods and programs in biomedicine*, 2018.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333. ACM, 2015.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232, 10 2001. doi: 10.1214/aos/1013203451. URL https://doi.org/10.1214/aos/1013203451.

Felix A. Gers, Jurgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10):2451–2471, 2000.

Ling Guo, Daniel Rivero, and Alejandro Pazos. Epileptic seizure detection using multiwavelet transform based approximate entropy and artificial neural networks. *Journal of neuroscience methods*, 193(1):156–163, 2010.

Sepp Hochreiter and Jrgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 17351780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Alistair E.w. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, Roger G. Mark, and et al. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

Yasuharu Koike and Mitsuo Kawato. Estimation of dynamic joint torques and trajectory formation from surface electromyography signals using a neural network model. *Biological cybernetics*, 73 (4):291–300, 1995.

Shu Liao, Yaozong Gao, Aytekin Oto, and Dinggang Shen. Representation learning: a unified deep learning framework for automatic prostate mr segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 254–261. Springer, 2013.

Scott Lundberg. Shap. https://github.com/slundberg/shap, 2018.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. *bioRxiv*, 2017. doi: 10.1101/206540. URL https://www.biorxiv.org/content/early/2017/10/21/206540.

Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *CoRR*, abs/1802.03888, 2018. URL http://arxiv.org/abs/1802.03888.

Xinbo Lv, Yi Guan, and Benyang Deng. Transfer learning based clinical concept extraction on data from multiple sources. *Journal of Biomedical Informatics*, 52:55 – 64, 2014. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2014.05.006. URL http://www.sciencedirect.com/science/article/pii/S1532046414001233. Special Section: Methods in Clinical Research Informatics.

Hector P Martinez, Yoshua Bengio, and Georgios N Yannakakis. Learning deep physiological models of affect. *IEEE Computational Intelligence Magazine*, 8(2):20–33, 2013.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. Understanding the exploding gradient problem. *CoRR*, abs/1211.5063, 2012. URL http://arxiv.org/abs/1211.5063.

Hariharan Ravishankar, Prasad Sudhakar, Rahul Venkataramani, Sheshadri Thiruvenkadam, Pavan Annangi, Narayanan Babu, and Vivek Vaidya. Understanding the mechanisms of deep transfer learning for medical images. In *Deep Learning and Data Labeling for Medical Applications*, pp. 188–196. Springer, 2016.

Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

Chuen-Kai Shie, Chung-Hisang Chuang, Chun-Nan Chou, Meng-Hsi Wu, and Edward Y Chang. Transfer representation learning for medical image analysis. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pp. 711–714. IEEE, 2015.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *CoRR*, abs/1605.01713, 2016. URL http://arxiv.org/abs/1605.01713.

Vairavan Srinivasan, Chikkannan Eswaran, and Natarajan Sriraam. Approximate entropy-based epileptic eeg detection using artificial neural networks. *IEEE Transactions on information Technology in Biomedicine*, 11(3):288–295, 2007.

AM Sullivan, Henian Xia, JC McBride, and Xiaopeng Zhao. Reconstruction of missing physiological signals using artificial neural networks. In *Computing in Cardiology, 2010*, pp. 317–320. IEEE, 2010.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, May 2016. ISSN 0278-0062. doi: 10.1109/TMI.2016.2535302.

Johannes Wagner, Jonghwa Kim, and Elisabeth André. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 940–943. IEEE, 2005.

Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4):699–706, 2014.

Glenn F Wilson and Christopher A Russell. Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human factors*, 45(4):635–644, 2003.

Dongrui Wu, Brent J Lance, and Thomas D Parsons. Collaborative filtering for brain-computer interaction using transfer learning and active class selection. *PloS one*, 8(2):e56624, 2013.

Te-chung Issac Yang and Haowei Hsieh. Classification of acoustic physiological signals based on deep learning neural networks with augmented features. In *Computing in Cardiology Conference (CinC), 2016*, pp. 569–572. IEEE, 2016.