LEARNING TO DIAGNOSE WITH LSTM RECURRENT NEURAL NETWORKS

Zachary C. Lipton *†
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093, USA
zlipton@cs.ucsd.edu

David C. Kale *‡
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
dkale@usc.edu

Charles Elkan
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093, USA
elkan@cs.ucsd.edu

Randall Wetzel
Laura P. and Leland K. Whittier Virtual PICU
Children's Hospital Los Angeles
Los Angeles, CA 90027
rwetzel@chla.usc.edu

ABSTRACT

Clinical medical data, especially in the intensive care unit (ICU), consist of multivariate time series of observations. For each patient visit (or *episode*), sensor data and lab test results are recorded in the patient's Electronic Health Record (EHR). While potentially containing a wealth of insights, the data is difficult to mine effectively, owing to varying length, irregular sampling and missing data. Recurrent Neural Networks (RNNs), particularly those using Long Short-Term Memory (LSTM) hidden units, are powerful and increasingly popular models for learning from sequence data. They effectively model varying length sequences and capture long range dependencies. We present the first study to empirically evaluate the ability of LSTMs to recognize patterns in multivariate time series of clinical measurements. Specifically, we consider multilabel classification of diagnoses, training a model to classify 128 diagnoses given 13 frequently but irregularly sampled clinical measurements. First, we establish the effectiveness of a simple LSTM network for modeling clinical data. Then we demonstrate a straightforward and effective training strategy in which we replicate targets at each sequence step. Trained only on raw time series, our models outperform several strong baselines, including a multilayer perceptron trained on hand-engineered features.

1 INTRODUCTION

Time series data comprised of clinical measurements, as recorded by caregivers in the pediatric intensive care unit (PICU), constitute an abundant and largely untapped source of medical insights. Potential uses of such data include classifying diagnoses accurately, predicting length of stay, predicting future illness, and predicting mortality. However, besides the difficulty of acquiring data, several obstacles stymie machine learning research with clinical time series. Episodes vary in length, with stays ranging from just a few hours to multiple months. Observations, which include sensor data, vital signs, lab test results, and subjective assessments, are sampled irregularly and plagued by missing values (Marlin et al., 2012). Additionally, long-term time dependencies complicate learning with many algorithms. Lab results that, taken together, might imply a particular diagnosis may be separated by days or weeks. Long delays often separate onset of disease from the appearance of symptoms. For example, symptoms of acute respiratory distress syndrome may not appear until 24-48 hours after lung injury (Mason et al., 2010), while symptoms of an asthma attack may present shortly after admission but change or disappear following treatment.

*Equal contributions
†Author website: http://zacklipton.com
‡Author website: http://www-scf.usc.edu/~dkale/

Recurrent Neural Networks (RNNs), in particular those based on Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), model varying-length sequential data, achieving state-of-the-art results for problems spanning natural language processing, image captioning, handwriting recognition, and genomic analysis (Auli et al., 2013; Sutskever et al., 2014; Vinyals et al., 2015; Karpathy & Fei-Fei, 2015; Liwicki et al., 2007; Graves et al., 2009; Pollastri et al., 2002; Vohradský, 2001; Xu et al., 2007). LSTMs can capture long range dependencies and nonlinear dynamics. Some sequence models, such as Markov models, conditional random fields, and Kalman filters, deal with sequential data but are ill-equipped to learn long-range dependencies. Other models require domain knowledge or feature engineering, offering less chance for serendipitous discovery. In contrast, neural networks learn representations and can discover unforeseen structure.

This paper presents the first empirical study using LSTMs to classify diagnoses given multivariate PICU time series. Specifically, we formulate the problem as multilabel classification, since diagnoses are not mutually exclusive. Our examples are clinical episodes, each consisting of 13 frequently but irregularly sampled time series of clinical measurements, including body temperature, heart rate, diastolic and systolic blood pressure, and blood glucose, among others. Associated with each patient are a subset of 429 diagnosis codes. As some are rare, we focus on the 128 most common codes, classifying each episode with one or more diagnoses.

Because LSTMs have never been used in this setting, we first verify their utility and compare their performance to a set of strong baselines, including both a linear classifier and a MultiLayer Perceptron (MLP). We train the baselines on both a fixed window and hand-engineered features. We then test a straightforward *target replication* strategy for recurrent neural networks, inspired by the *deep supervision* technique of Lee et al. (2015) for training convolutional neural networks. We compose our optimization objective as a convex combination of the loss at the final sequence step and the mean of the losses over *all* sequence steps. Additionally, we evaluate the efficacy of using additional information in the patient's chart as auxiliary outputs, a technique previously used with feedforward nets (Caruana et al., 1996), showing that it reduces overfitting. Finally, we apply dropout to non-recurrent connections, which improves the performance further. LSTMs with target replication and dropout surpass the performance of the best baseline, namely an MLP trained on hand-engineered features, even though the LSTM has access only to raw time series.

2 Related Work

Our research sits at the intersection of LSTMs, medical informatics, and multilabel classification, three mature fields, each with a long history and rich body of research. While we cannot do justice to all three, we highlight the most relevant works below.

2.1 LSTM RNNs

LSTMs were originally introduced in Hochreiter & Schmidhuber (1997), following a long line of research into RNNs for sequence learning. Notable earlier work includes Rumelhart et al. (1985), which introduced backpropagation through time, and Elman (1990), which successfully trained RNNs to perform supervised machine learning tasks with sequential inputs and outputs. The design of modern LSTM memory cells has remained close to the original, with the commonly used addition of forget gates (Gers et al., 2000) (which we use), and peep-hole connections (Gers & Schmidhuber, 2000) (which we do not use). The connectivity pattern among multiple LSTM layers in our models follows the architecture described by Graves (2013). Pascanu et al. (2014) explores other mechanisms by which an RNN could be made *deep*. Surveys of the literature include Graves (2012), a thorough dissertation on sequence labeling with RNNs, De Mulder et al. (2015), which surveys natural language applications, and Lipton et al. (2015), which provides a broad overview of RNNs for sequence learning, focusing on modern applications.

2.2 Neural Networks for Medical Data

Neural networks have been applied to medical problems and data for at least 20 years (Caruana et al., 1996; Baxt, 1995), although we know of no work on applying LSTMs to multivariate clinical time series of the type we analyze here. Several papers have applied RNNs to physiologic signals, including electrocardiograms (Silipo & Marchesi, 1998; Amari & Cichocki, 1998; Übeyli, 2009) and

glucose measurements (Tresp & Briegel, 1998). RNNs have also been used for prediction problems in genomics (Pollastri et al., 2002; Xu et al., 2007; Vohradský, 2001). Multiple recent papers apply modern deep learning techniques (but not RNNs) to modeling psychological conditions (Dabek & Caban, 2015), head injuries (Rughani et al., 2010), and Parkinson's disease (Hammerla et al., 2015). Recently, feedforward networks have been applied to medical time series in sliding window fashion to classify cases of gout, leukemia (Lasko et al., 2013), and critical illness (Che et al., 2015).

2.3 Neural Networks for Multilabel Classification

Only a few published papers apply LSTMs to multilabel classification tasks, all of which, to our knowledge, are outside of the medical context. Liu et al. (2014) formulates music composition as a multilabel classification task, using sigmoidal output units. Most recently, Yeung et al. (2015) uses LSTM networks with multilabel outputs to recognize actions in videos. While we could not locate any published papers using LSTMs for multilabel classification in the medical domain, several papers use feedforward nets for this task. One of the earliest papers to investigate multi-task neural networks modeled risk in pneumonia patients (Caruana et al., 1996). More recently, Che et al. (2015) formulated diagnosis as multilabel classification using a sliding window multilayer perceptron.

2.4 Machine Learning for Clinical Time Series

Neural network methodology aside, a growing body of research applies machine learning to temporal clinical data for tasks including artifact removal (Aleks et al., 2009; Quinn et al., 2009), early detection and prediction (Stanculescu et al., 2014a; Henry et al., 2015), and clustering and subtyping (Marlin et al., 2012; Schulam et al., 2015). Many recent papers use models with latent factors to capture nonlinear dynamics in clinical time series and to discover meaningful representations of health and illness. Gaussian processes are popular because they can directly handle irregular sampling and encode prior knowledge via choice of covariance functions between time steps and across variables (Marlin et al., 2012; Ghassemi et al., 2015). Saria et al. (2010) combined a hierarchical dirichlet process with autoregressive models to infer latent disease "topics" in the heart rate signals of premature babies. Quinn et al. (2009) used linear dynamical systems with latent switching variables to model physiologic events like bradycardias. Seeking *deeper* models, Stanculescu et al. (2014b) proposed a second "layer" of latent factors to capture correlations between latent states.

2.5 Target Replication

In this work, we make the task of classifying entire sequences easier by replicating targets at every time step, inspired by Lee et al. (2015), who place an optimization objective after each layer in convolutional neural network. While they have a separate set of weights to learn each intermediate objective, our model is simpler owing to the weight tying in recurrent nets, having only one set of output weights. Additionally, unlike Lee et al. (2015), we place targets at each time step, but not following each layer between input and output in the LSTM. After finishing this manuscript, we learned that target replication strategies similar to ours have also been developed by Ng et al. (2015) and Dai & Le (2015) for the tasks of video classification and character-level document classification respectively. Ng et al. (2015) linearly scale the importance of each intermediate target, emphasizing performance at later sequence steps over those in the beginning of the clip. Dai & Le (2015) also use a target replication strategy with linearly increasing weight for character-level document classification, showing significant improvements in accuracy. They call this technique *linear gain*.

2.6 Regularizing Recurrent Neural Networks

Given the complexity of our models and modest scale of our data, regularization, including judicious use of dropout, is crucial to our performance. Several prior works use dropout to regularize RNNs. Pham et al. (2014), Zaremba et al. (2014), and Dai & Le (2015) all describe an application of dropout to only the non-recurrent weights of a network. The former two papers establish the method and apply it to tasks with sequential outputs, including handwriting recognition, image captioning, and machine translation. The setting studied by Dai & Le (2015) most closely resembles ours as the authors apply it to the task of applying static labels to varying length sequences.

2.7 KEY DIFFERENCES

Our experiments show that LSTMs can accurately classify multivariate time series of clinical measurements, a topic not addressed in any prior work. Additionally, while some papers use LSTMs for multilabel classification, ours is the first to address this problem in the medical context. Moreover, for multilabel classification of sequential clinical data with fixed length output vectors, this paper is the first, to our knowledge, to demonstrate the efficacy of a target replication strategy, achieving both faster training and better generalization.

3 DATA DESCRIPTION

Our experiments use a collection of anonymized clinical time series extracted from the EHR system at Children's Hospital LA (Marlin et al., 2012; Che et al., 2015) as part of an IRB-approved study. The data consists of $10,401$ PICU episodes, each a multivariate time series of 13 variables: diastolic and systolic blood pressure, peripheral capillary refill rate, end-tidal CO_2, fraction of inspired O_2, Glascow coma scale, blood glucose, heart rate, pH, respiratory rate, blood oxygen saturation, body temperature, and urine output. Episodes vary in length from 12 hours to several months.

Each example consists of irregularly sampled multivariate time series with both missing values and, occasionally, missing variables. We resample all time series to an hourly rate, taking the mean measurement within each one hour window. We use forward- and back-filling to fill gaps created by the window-based resampling. When a single variable's time series is missing entirely, we impute a clinically normal value as defined by domain experts. These procedures make reasonable assumptions about clinical practice: many variables are recorded at rates proportional to how quickly they change, and when a variable is absent, it is often because clinicians believed it to be normal and chose not to measure it. Nonetheless, these procedures are not appropriate in all settings. Backfilling, for example, passes information from the future backwards. This is acceptable for classifying entire episodes (as we do) but not for forecasting. Finally, we rescale all variables to $[0, 1]$, using ranges defined by clinical experts. In addition, we use published tables of normal values from large population studies to correct for differences in heart rate, respiratory rate, (Fleming et al., 2011) and blood pressure (NHBPEP Working Group 2004) due to age and gender.

Each episode is associated with zero or more diagnostic codes from an in-house taxonomy used for research and billing, similar to the *Ninth Revision of the International Classification of Diseases* (ICD-9) codes (World Health Organization, 2004). The dataset contains 429 distinct labels indicating a variety of conditions, such as acute respiratory distress, congestive heart failure, seizures, renal failure, and sepsis. Because many of the diagnoses are rare, we focus on the most common 128, each of which occurs more than 50 times in the data. These diagnostic codes are recorded by attending physicians during or shortly after each patient episode and subject to limited review afterwards.

Because the diagnostic codes were assigned by clinicians, our experiments represent a comparison of an LSTM-based diagnostic system to human experts. We note that an attending physician has access to much more data about each patient than our LSTM does, including additional tests, medications, and treatments. Additionally, the physician can access a full medical history including free-text notes, can make visual and physical inspections of the patient, and can ask questions. A more fair comparison might require asking additional clinical experts to assign diagnoses given access only to the 13 time series available to our models. However, this would be prohibitively expensive, even for just the 1000 examples, and difficult to justify to our medical collaborators, as this annotation would provide no immediate benefit to patients. Such a study will prove more feasible in the future when this line of research has matured.

4 METHODS

In this work, we are interested in recognizing diagnoses and, more broadly, the observable physiologic characteristics of patients, a task generally termed *phenotyping* (Oellrich et al., 2015). We cast the problem of phenotyping clinical time series as multilabel classification. Given a series of observations $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(T)}$, we learn a classifier to generate hypotheses $\hat{\boldsymbol{y}}$ of the true labels $\boldsymbol{y}$. Here, t indexes sequence steps, and for any example, T stands for the length of the sequence. Our proposed LSTM RNN uses memory cells with forget gates (Gers et al., 2000) but without peephole

connections (Gers et al., 2003). As output, we use a fully connected layer atop the highest LSTM layer followed by an element-wise sigmoid activation function, because our problem is multilabel. We use *log loss* as the loss function at each output.

The following equations give the update for a layer of memory cells $\boldsymbol{h}_l^{(t)}$ where $\boldsymbol{h}_{l-1}^{(t)}$ stands for the previous layer at the same sequence step (a previous LSTM layer or the input $\boldsymbol{x}^{(t)}$) and $\boldsymbol{h}_l^{(t-1)}$ stands for the same layer at the previous sequence step:

$$\begin{aligned}
\boldsymbol{g}_l^{(t)} &= \phi(W_l^{\mathrm{gx}}\boldsymbol{h}_{l-1}^{(t)} + W_l^{\mathrm{gh}}\boldsymbol{h}_l^{(t-1)} + \boldsymbol{b}_l^{\mathrm{g}}) \\
\boldsymbol{i}_l^{(t)} &= \sigma(W_l^{\mathrm{ix}}\boldsymbol{h}_{l-1}^{(t)} + W_l^{\mathrm{ih}}\boldsymbol{h}_l^{(t-1)} + \boldsymbol{b}_l^{\mathrm{i}}) \\
\boldsymbol{f}_l^{(t)} &= \sigma(W_l^{\mathrm{fx}}\boldsymbol{h}_{l-1}^{(t)} + W_l^{\mathrm{fh}}\boldsymbol{h}_l^{(t-1)} + \boldsymbol{b}_l^{\mathrm{f}}) \\
\boldsymbol{o}_l^{(t)} &= \sigma(W_l^{\mathrm{ox}}\boldsymbol{h}_{l-1}^{(t)} + W_l^{\mathrm{oh}}\boldsymbol{h}_l^{(t-1)} + \boldsymbol{b}_l^{\mathrm{o}}) \\
\boldsymbol{s}_l^{(t)} &= \boldsymbol{g}_l^{(t)} \odot \boldsymbol{i}_l^{(i)} + \boldsymbol{s}_l^{(t-1)} \odot \boldsymbol{f}_l^{(t)} \\
\boldsymbol{h}_l^{(t)} &= \phi(\boldsymbol{s}_l^{(t)}) \odot \boldsymbol{o}_l^{(t)}.
\end{aligned}$$

In these equations, σ stands for an element-wise application of the *sigmoid* (*logistic*) function, ϕ stands for an element-wise application of the $tanh$ function, and $\odot$ is the Hadamard (element-wise) product. The input, output, and forget gates are denoted by $\boldsymbol{i}$, $\boldsymbol{o}$, and $\boldsymbol{f}$ respectively, while $\boldsymbol{g}$ is the input node and has a *tanh* activation.

4.1 LSTM Architectures for Multilabel Classification

We explore several recurrent neural network architectures for multilabel classification of time series. The first and simplest (Figure 1) passes over all inputs in chronological order, generating outputs only at the final sequence step. In this approach, we only have output $\hat{\boldsymbol{y}}$ at the final sequence step, at which our loss function is the average of the losses at each output node. Thus the loss calculated at a single sequence step is the average of *log loss* calculated separately on each label.

$$\text{loss}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{|L|}\sum_{l=1}^{l=|L|} -(y_l \cdot \log(\hat{y}_l) + (1 - y_l) \cdot \log(1 - \hat{y}_l)).$$

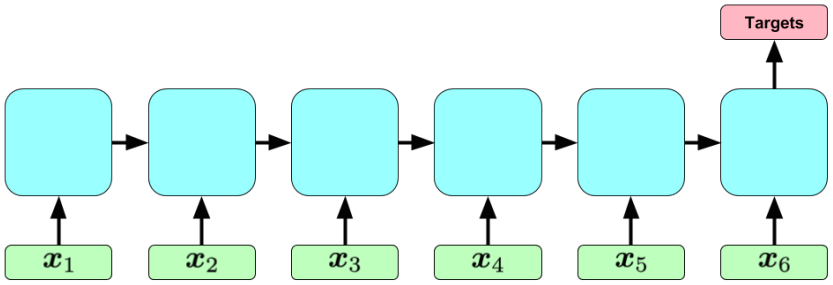


Figure 1: A simple RNN model for multilabel classification. Green rectangles represent inputs. The recurrent hidden layers separating input and output are represented with a single blue rectangle. The red rectangle represents targets.

4.2 Sequential Target Replication

One problem with the simple approach is that the network must learn to pass information across many sequence steps in order to affect the output. We attack this problem by replicating our static targets at each sequence step (Figure 2), providing a local error signal at each step. This approach is inspired by the deep supervision technique that Lee et al. (2015) apply to convolutional nets. This technique is especially sensible in our case because we expect the model to predict accurately even if the sequence were truncated by a small amount. The approach differs from Lee et al. (2015) because

we use the same output weights to calculate $\hat{\boldsymbol{y}}^{(t)}$ for all t. Further, we use this target replication to generate output at each sequence step, but not at each hidden layer.

For the model with target replication, we generate an output $\hat{\boldsymbol{y}}^{(t)}$ at every sequence step. Our loss is then a convex combination of the final loss and the average of the losses over all steps:

$$\alpha \cdot \frac{1}{T} \sum_{t=1}^{T} \text{loss}(\hat{\boldsymbol{y}}^{(t)}, \boldsymbol{y}^{(t)}) + (1 - \alpha) \cdot \text{loss}(\hat{\boldsymbol{y}}^{(T)}, \boldsymbol{y}^{(T)})$$

where T is the total number of sequence steps and $\alpha \in [0, 1]$ is a hyper-parameter which determines the relative importance of hitting these intermediary targets. At prediction time, we take only the output at the final step. In our experiments, networks using target replication outperform those with a loss applied only at the final sequence step.

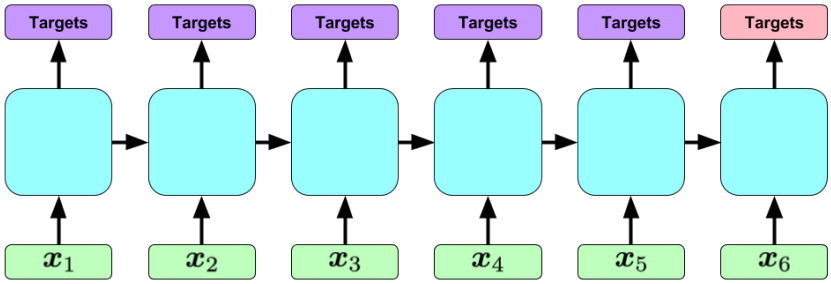


Figure 2: An RNN classification model with *target replication*. The primary target (depicted in red) at the final step is used at prediction time, but during training, the model back-propagates errors from the intermediate targets (purple) at every sequence step.

4.3 Auxiliary Output Training

Recall that our initial data contained 429 diagnostic labels but that our task is to predict only 128. Given the well-documented successes of multitask learning with shared representations and feed-forward networks, we wish to train a stronger model by using the remaining 301 labels or other information in the patient's chart, such as diagnostic categories, as auxiliary targets (Caruana et al., 1996). These additional targets serve reduce overfitting as the model aims to minimize the loss on the labels of interest while also minimizing loss on the auxiliary targets (Figure 3).

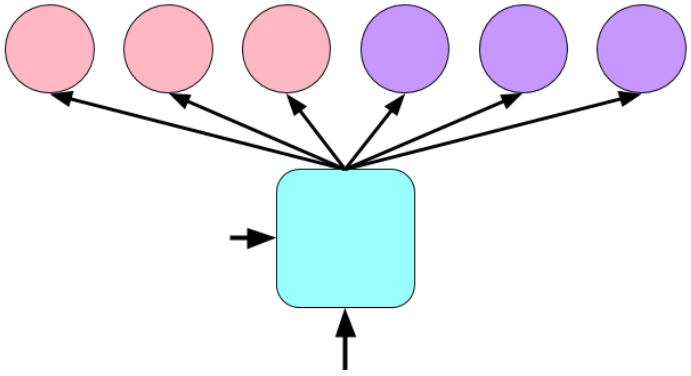


Figure 3: Our dataset contains many labels. For our task, a subset of 128 are of interest (depicted in red). Our *Auiliary Output* neural network makes use of extra labels as additional training targets (depicted in purple). At inference time we generate predictions for only the labels of interest.

4.4 REGULARIZATION

Because we have only $10,401$ examples, overfitting is a considerable obstacle. Our experiments show that both target replication and auxiliary outputs improve performance and reduce overfitting. In addition to these less common techniques we deploy ℓ_2^2 weight decay and dropout. Following the example of Zaremba et al. (2014) and Pham et al. (2014), we apply dropout to the non-recurrent connections only. We first compute each hidden layer's sequence of activations in the left-to-right direction and then apply dropout before computing the next layer's activations. In our experiments, we find that dropout decreases overfitting, enabling us to double the size of each hidden layer.

5 EXPERIMENTS

All models are trained on 80% of the data and tested on 10%. The remaining 10% is used as a validation set. We train each LSTM for 100 epochs using stochastic gradient descent (SGD) with momentum. To combat exploding gradients, we scale the norm of the gradient and use ℓ_2^2 weight decay of 10^{-6}, both hyperparameters chosen using validation data. Our final networks use 2 hidden layers and either 64 memory cells per layer with no dropout or 128 cells per layer with dropout of 0.5. These architectures are also chosen based on validation performance. Throughout training, we save the model and compute three performance metrics (micro AUC, micro F1, and precision at 10) on the validation set for each epoch. We then test the model that scores best on at least two of the three validation metrics. To break ties, we choose the earlier epoch.

We evaluate a number of baselines as well as LSTMs with various combinations of target replication (TR), dropout (DO), and auxiliary outputs (AO), using either the additional 301 diagnostic labels or 12 diagnostic categories. To explore the regularization effects of each strategy, we record and plot both training and validation performance after each epoch. Additionally, we report performance of a target replication model (Linear Gain) that scales the weight of each intermediate target linearly as opposed our proposed approach. Finally, to show that our LSTM learns a model complementary to the baselines, we evaluate an ensemble of the best LSTM with the best baseline.

5.1 MULTILABEL EVALUATION METHODOLOGY

We report micro- and macro-averaged versions of Area Under the ROC Curve (AUC). By micro AUC, we mean a single AUC computed on flattened $\hat{Y}$ and Y matrices, whereas we calculate macro AUC by averaging each per-label AUC. The blind classifier achieves 0.5 macro AUC but can exceed 0.5 on micro AUC by predicting labels in descending order by base rate. Additionally, we report micro- and macro-averaged F1 score, computed in similar fashion to the respective micro and macro AUCs. F1 metrics require a thresholding strategy, and here we select thresholds based upon validation set performance. We refer to Lipton et al. (2014) for an analysis of the strengths and weaknesses of each type of multilabel F-score and a characterization of optimal thresholds.

Finally, we report *precision at* 10, which captures the fraction of true diagnoses among the model's top 10 predictions, with a best possible score of 0.2281 on the test split of this data set because there are on average 2.281 diagnoses per patient. While F1 and AUC are both useful for determining the relative quality of a classifier's predictions, neither is tailored to a real-world application. Thus, we consider a medically plausible use case to motivate this more interpretable metric: generating a short list of the 10 most probable diagnoses. If we could create a high recall, moderate precision list of 10 likely diagnoses, it could be a valuable hint-generation tool for differential diagnosis. Testing for only the 10 most probable conditions is much more realistic than testing for all conditions.

5.2 BASELINE CLASSIFIERS

We provide results for a *base rate* model that predicts diagnoses in descending order by incidence to provide a minimum performance baseline for micro-averaged metrics. We also report the performance of logistic regression, which is widely used in clinical research. We train a separate classifier for each diagnosis but choose an overall ℓ_2^2 penalty for all individual classifiers based on validation performance. For a much stronger baseline, we train a multilabel MLP with 3 hidden layers of 300 hidden units each, rectified linear activations, and dropout of 0.5. All MLPs were trained for 1000 epochs, with hyperparameters chosen based on validation set performance. Each baseline is tested

with two sets of inputs: raw time series and hand-engineered features. For raw time series, we use the first and last six hours. This provides classifiers with temporal information about changes in patient state from admission to discharge within a fixed-size input, as required by all baselines. We find this works better than providing the first or last 12 hours alone.

Our hand-engineered features are inspired by those used in state-of-the-art severity of illness scores (Pollack et al., 1996): for each variable, we compute the first and last measurements and their difference scaled by episode length, mean and standard deviation, median and quartiles, minimum and maximum, and slope of a line fit with least squares. These 143 features capture many of the indicators that clinicians look for in critical illness, including admission and discharge state, extremes, central tendencies, variability, and trends. They previously have been shown to be effective for these data (Marlin et al., 2012; Che et al., 2015). Our strongest baseline is an MLP using these features.

5.3 RESULTS

Our best performing LSTM (LSTM-DO-TR) used two layers of 128 memory cells, dropout of probability 0.5 between layers, and target replication, and outperformed the MLP with hand-engineered features. Moreover simple ensembles of the best LSTM and MLP outperformed both on all metrics. Table 1 shows summary results for all models. Table 2 shows the LSTM's predictive performance for six diagnoses with the highest F1 scores. Full per-diagnosis results can be found in Appendix C.

Target replication improves performance on all metrics, accelerating learning and reducing overfitting (Figure 4). We also find that the LSTM with target replication learns to output correct diagnoses earlier in the time series, a virtue that we explore qualitatively in Appendix A. As a comparison, we trained a LSTM-DO-TR variant using the linear gain strategy of Ng et al. (2015); Dai & Le (2015). In general, this model did not perform as well as our simpler target replication strategy, but it did achieve the highest macro F1 score among the LSTM models.

Classification performance for 128 ICU phenotypes

Model	Micro AUC	Macro AUC	Micro F1	Macro F1	Prec. at 10
Base Rate	0.7128	0.5	0.1346	0.0343	0.0788
Log. Reg., First 6 + Last 6	0.8122	0.7404	0.2324	0.1081	0.1016
Log. Reg., Expert features	0.8285	0.7644	0.2502	0.1373	0.1087
MLP, First 6 + Last 6	0.8375	0.7770	0.2698	0.1286	0.1096
MLP, Expert features	**0.8551**	**0.8030**	**0.2930**	**0.1475**	**0.1170**
LSTM Models with two 64-cell hidden layers					
LSTM	0.8241	0.7573	0.2450	0.1170	0.1047
LSTM, AuxOut (Diagnoses)	0.8351	0.7746	0.2627	0.1309	0.1110
LSTM-AO (Categories)	0.8382	0.7748	0.2651	0.1351	0.1099
LSTM-TR	0.8429	0.7870	0.2702	0.1348	0.1115
LSTM-TR-AO (Diagnoses)	0.8391	0.7866	0.2599	0.1317	0.1085
LSTM-TR-AO (Categories)	0.8439	0.7860	0.2774	0.1330	0.1138
LSTM Models with Dropout (probability 0.5) and two 128-cell hidden layers					
LSTM-DO	0.8377	0.7741	0.2748	0.1371	0.1110
LSTM-DO-AO (Diagnoses)	0.8365	0.7785	0.2581	0.1366	0.1104
LSTM-DO-AO (Categories)	0.8399	0.7783	0.2804	0.1361	0.1123
LSTM-DO-TR	**0.8560**	**0.8075**	**0.2938**	0.1485	0.**1172**
LSTM-DO-TR-AO (Diagnoses)	0.8470	0.7929	0.2735	0.1488	0.1149
LSTM-DO-TR-AO (Categories)	0.8543	0.8015	0.2887	0.1446	0.1161
LSTM-DO-TR (Linear Gain)	0.8480	0.7986	0.2896	**0.1530**	0.1160
Ensembles of Best MLP and Best LSTM					
Mean of LSTM-DO-TR & MLP	0.8611	0.8143	0.2981	0.1553	0.1201
Max of LSTM-DO-TR & MLP	**0.8643**	**0.8194**	**0.3035**	**0.1571**	**0.1218**

Table 1: Results on performance metrics calculated across all labels. *DO*, *TR*, and *AO* indicate dropout, target replication, and *auxiliary outputs*, respectively. *AO (Diagnoses)* uses the extra diagnosis codes and *AO (Categories)* uses diagnostic categories as additional targets during training.

Auxiliary outputs improved performance for most metrics and reduced overfitting. While the performance improvement is not as dramatic as that conferred by target replication, the regularizing effect is greater. These gains came at the cost of slower training: the auxiliary output models required more epochs (Figure 4 and Appendix B), especially when using the 301 remaining diagnoses. This may be due in part to severe class imbalance in the extra labels. For many of these labels it may take an entire epoch just to learn that they are occasionally nonzero.

Top 6 diagnoses measured by F1 score

Label	*F1*	AUC	Precision	Recall
Diabetes mellitus with ketoacidosis	0.8571	0.9966	1.0000	0.7500
Scoliosis, idiopathic	0.6809	0.8543	0.6957	0.6667
Asthma, unspecified with status asthmaticus	0.5641	0.9232	0.7857	0.4400
Neoplasm, brain, unspecified	0.5430	0.8522	0.4317	0.7315
Delayed milestones	0.4751	0.8178	0.4057	0.5733
Acute Respiratory Distress Syndrome (ARDS)	0.4688	0.9595	0.3409	0.7500

Table 2: LSTM-DO-TR performance on the 6 diagnoses with highest F1 scores.

The LSTMs appear to learn models complementary to the MLP trained on hand-engineered features. Supporting this claim, simple ensembles of the LSTM-DO-TR and MLP (taking the *mean* or *maximum* of their predictions) outperform the constituent models significantly on all metrics (Table 1). Further, there are many diseases for which one model substantially outperforms the other, e.g., intracranial hypertension for the LSTM, septic shock for the MLP (Appendix C).

6 Discussion

Our results indicate that LSTM RNNs, especially with target replication, can successfully classify diagnoses of critical care patients given clinical time series data. The best LSTM beat a strong MLP baseline using hand-engineered features as input, and an ensemble combining the MLP and LSTM improves upon both. The success of target replication accords with results by both Ng et al. (2015) and Dai & Le (2015), who observed similar benefits on their respective tasks. However, while they saw improvement using a linearly increasing weight on each target from start to end, this strategy performed worse in our diagnostic classification task than our uniform weighting of intermediate targets. We believe this may owe to the peculiar nature of our data. Linear gain emphasizes evidence from later in the sequence, an assumption which often does not match the progression of symptoms in critical illnesses. Asthma patients, for example, are often admitted to the ICU severely symptomatic, but once treatment begins, patient physiology stabilizes and observable signs of disease may abate or change. Further supporting this idea, we observed that when training fixed-window baselines, using the first 6 and last 6 hours outperformed using the last 12 hours only.

While our data is of large scale by clinical standards, it is small relative to datasets found in deep learning tasks like vision and speech recognition. At this scale, regularization is critical. Our experiments demonstrate that target replication, auxiliary outputs, and dropout all work to reduce the generalization gap. as shown in Figure 4 and Appendix B. However, some of these techniques are complementary while others seem to cancel each other out. For example, our best model combined target replication with dropout. This combination significantly improved upon the performance using target replication alone, and enabled the effective use of larger capacity models. In contrast, the benefits of dropout and auxiliary output training appear to wash each other out. This

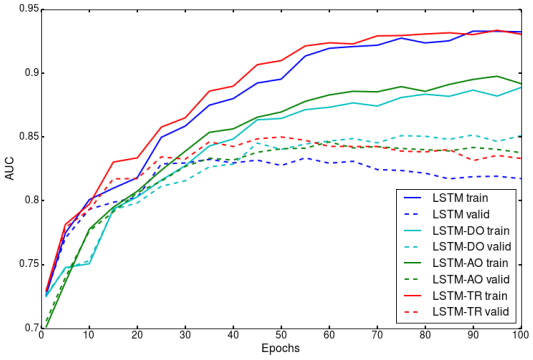


Figure 4: Training curves showing the impact of the *DO*, *AO*, and *TR* strategies on overfitting.

may be because target replication confers more than regularization, mitigating the difficulty of learning long range dependencies by providing local objectives.

7 CONCLUSION

While our LSTMs produce promising results, this is only a first step in this line of research. Recognizing diagnoses given full time series of sensor data demonstrates that LSTMs can capture meaningful signal, but ultimately we would like to predict developing conditions and events, outcomes such as mortality, and treatment responses. In this paper we used diagnostic labels without timestamps, but we are obtaining timestamped diagnoses, which will enable us to train models to perform early diagnosis by predicting future conditions. In addition, we are extending this work to a larger PICU data set with 50% more patients and hundreds of variables, including treatments and medications.

On the methodological side, we would like to both better exploit and improve the capabilities of LSTMs. Results from speech recognition have shown that LSTMs shine in comparison to other models using raw features, minimizing need for preprocessing and feature engineering. In contrast, our current data preparation pipeline removes valuable structure and information from clinical time series that could be exploited by an LSTM. For example, our forward- and back-filling imputation strategies discard useful information about when each observation is recorded. Imputing normal values for missing time series ignores the meaningful distinction between truly normal and missing measurements. Also, our window-based resampling procedure reduces the variability of more frequently measured vital signs (e.g., heart rate).

In future work, we plan to introduce indicator variables to allow the LSTM to distinguish actual from missing or imputed measurements. Additionally, the flexibility of the LSTM architecture should enable us to eliminate age-based corrections and to incorporate non-sequential inputs, such as age, weight, and height (or even hand-engineered features), into predictions. Other next steps in this direction include developing LSTM architectures to directly handle missing values and irregular sampling. We also are encouraged by the success of target replication and plan to explore other variants of this technique and to apply it to other domains and tasks. Additionally, we acknowledge that there remains a debate about the interpretability of neural networks when applied to complex medical problems. We are developing methods to interpret the representations learned by LSTMs in order to better expose patterns of health and illness to clinical users. We also hope to make practical use of the distributed representations of patients for tasks such as patient similarity search.

8 ACKNOWLEDGEMENTS

Zachary C. Lipton was supported by the Division of Biomedical Informatics at the University of California, San Diego, via training grant (T15LM011271) from the NIH/NLM. David Kale was supported by the Alfred E. Mann Innovation in Engineering Doctoral Fellowship. The VPICU was supported by grants from the Laura P. and Leland K. Whittier Foundation. We acknowledge NVIDIA Corporation for Tesla K40 GPU hardware donation and Professors Julian McAuley and Greg Ver Steeg for their support and advice. Finally, we thank the anonymous ICLR reviewers for their feedback, which helped us to make significant improvements to this work and manuscript.

REFERENCES

Aleks, Norm, Russell, Stuart J, Madden, Michael G, Morabito, Diane, Staudenmayer, Kristan, Cohen, Mitchell, and Manley, Geoffrey T. Probabilistic detection of short events, with application to critical care monitoring. In *Advances in Neural Information Processing Systems (NIPS) 21*, pp. 49–56, 2009.

Amari, Shun-ichi and Cichocki, Andrzej. Adaptive blind signal processing-neural network approaches. *Proceedings of the IEEE*, 86(10):2026–2048, 1998.

Auli, Michael, Galley, Michel, Quirk, Chris, and Zweig, Geoffrey. Joint language and translation modeling with recurrent neural networks. In *Empirical Methods in Natural Language Proessing (EMNPL)*, volume 3, 2013.

Baxt, W.G. Application of artificial neural networks to clinical medicine. *The Lancet*, 346(8983): 1135–1138, 1995.

Caruana, Rich, Baluja, Shumeet, Mitchell, Tom, et al. Using the future to "sort out" the present: Rankprop and multitask learning for medical risk evaluation. In *Advances in Neural Information Processing Systems (NIPS) 8*, pp. 959–965, 1996.

Che, Zhengping, Kale, David C., Li, Wenzhe, Bahadori, Mohammad Taha, and Liu, Yan. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 507–516. ACM, 2015.

Dabek, Filip and Caban, Jesus J. A neural network based model for predicting psychological conditions. In *Brain Informatics and Health*, pp. 252–261. Springer, 2015.

Dai, Andrew M and Le, Quoc V. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems (NIPS) 28*, pp. 3061–3069, 2015.

De Mulder, Wim, Bethard, Steven, and Moens, Marie-Francine. A survey on the application of recurrent neural networks to statistical language modeling. *Computer Speech & Language*, 30(1): 61–98, 2015.

Elman, Jeffrey L. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Fleming, Susannah, Thompson, Matthew, Stevens, Richard, Heneghan, Carl, Plddemann, Annette, Maconochie, Ian, Tarassenko, Lionel, and Mant, David. Normal ranges of heart rate and respiratory rate in children from birth to 18 years: A systematic review of observational studies. *The Lancet*, pp. 1011–1018, 2011.

Gers, Felix and Schmidhuber, Jürgen. Recurrent nets that time and count. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN)*, volume 3, pp. 189–194. IEEE, 2000.

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

Gers, Felix A., Schraudolph, Nicol N., and Schmidhuber, Jürgen. Learning precise timing with LSTM recurrent networks. *The Journal of Machine Learning Research*, 3:115–143, 2003.

Ghassemi, Marzyeh, Pimentel, Marco AF, Naumann, Tristan, Brennan, Thomas, Clifton, David A, Szolovits, Peter, and Feng, Mengling. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Graves, Alex. *Supervised Sequence Labelling with Recurrent Neural Networks*, volume 385 of *Studies in Computational Intelligence*. Springer-Verlag Berlin Heidelberg, 2012.

Graves, Alex. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Graves, Alex, Liwicki, Marcus, Fernández, Santiago, Bertolami, Roman, Bunke, Horst, and Schmidhuber, Jürgen. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, 2009.

Hammerla, Nils Y, Fisher, James M, Andras, Peter, Rochester, Lynn, Walker, Richard, and Plötz, Thomas. PD disease state assessment in naturalistic environments using deep learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Henry, Katharine E, Hager, David N, Pronovost, Peter J, and Saria, Suchi. A targeted real-time early warning score (trewscore) for septic shock. *Science Translational Medicine*, 7(299 299ra122): 1–9, 2015.

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Karpathy, Andrej and Fei-Fei, Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128–3137, June 2015.

Lasko, Thomas A., Denny, Joshua C., and Levy, Mia A. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS ONE*, 8(6): e66341, 06 2013.

Lee, Chen-Yu, Xie, Saining, Gallagher, Patrick, Zhang, Zhengyou, and Tu, Zhuowen. Deeply-supervised nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Lipton, Zachary C, Elkan, Charles, and Naryanaswamy, Balakrishnan. Optimal thresholding of classifiers to maximize F1 measure. In *Machine Learning and Knowledge Discovery in Databases*, pp. 225–239. Springer, 2014.

Lipton, Zachary C., Berkowitz, John, and Elkan, Charles. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

Liu, I, Ramakrishnan, Bhiksha, et al. Bach in 2014: Music composition with recurrent neural network. *arXiv preprint arXiv:1412.3191*, 2014.

Liwicki, Marcus, Graves, Alex, Bunke, Horst, and Schmidhuber, Jürgen. A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, volume 1, pp. 367–371, 2007.

Marlin, Ben M., Kale, David C., Khemani, Robinder G., and Wetzel, Randall C. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI)*, 2012.

Mason, Robert J., Broaddus, V. Courtney, Martin, Thomas, King Jr., Talmadge E., Schraufnagel, Dean, Murray, John F., and Nadel, Jay A. *Murray and Nadel's textbook of respiratory medicine: 2-volume set*. Elsevier Health Sciences, 2010.

National High Blood Pressure Education Program Working Group on Children and Adolescents. The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents. *Pediatrics*, 114:555–576, 2004.

Ng, Joe Yue-Hei, Hausknecht, Matthew, Vijayanarasimhan, Sudheendra, Vinyals, Oriol, Monga, Rajat, and Toderici, George. Beyond short snippets: Deep networks for video classification. *arXiv preprint arXiv:1503.08909*, 2015.

Oellrich, Anika, Collier, Nigel, Groza, Tudor, Rebholz-Schuhmann, Dietrich, Shah, Nigam, Bodenreider, Olivier, Boland, Mary Regina, Georgiev, Ivo, Liu, Hongfang, Livingston, Kevin, Luna, Augustin, Mallon, Ann-Marie, Manda, Prashanti, Robinson, Peter N., Rustici, Gabriella, Simon, Michelle, Wang, Liqin, Winnenburg, Rainer, and Dumontier, Michel. The digital revolution in phenotyping. *Briefings in Bioinformatics*, 2015.

Pascanu, Razvan, Gulcehre, Caglar, Cho, Kyunghyun, and Bengio, Yoshua. How to construct deep recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Pham, Vu, Bluche, Théodore, Kermorvant, Christopher, and Louradour, Jérôme. Dropout improves recurrent neural networks for handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pp. 285–290. IEEE, 2014.

Pollack, M. M., Patel, K. M., and Ruttimann, U. E. *PRISM III: an updated Pediatric Risk of Mortality score. Critical Care Medicine*, 24(5):743–752, 1996.

Pollastri, Gianluca, Przybylski, Darisz, Rost, Burkhard, and Baldi, Pierre. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Structure, Function, and Bioinformatics*, 47(2):228–235, 2002.

Quinn, John, Williams, Christopher KI, McIntosh, Neil, et al. Factorial switching linear dynamical systems applied to physiological condition monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1537–1551, 2009.

Rughani, Anand I., Dumont, Travis M., Lu, Zhenyu, Bongard, Josh, Horgan, Michael A., Penar, Paul L., and Tranmer, Bruce I. Use of an artificial neural network to predict head injury outcome: clinical article. *Journal of Neurosurgery*, 113(3):585–590, 2010.

Rumelhart, David E, Hinton, Geoffrey E, and Williams, Ronald J. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

Saria, Suchi, Koller, Daphne, and Penn, Anna. Learning individual and population level traits from clinical temporal data. In *Proc. Neural Information Processing Systems (NIPS), Predictive Models in Personalized Medicine Workshop*. Citeseer, 2010.

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

Silipo, Rosaria and Marchesi, Carlo. Artificial neural networks for automatic ecg analysis. *IEEE Transactions on Signal Processing*, 46(5):1417–1425, 1998.

Stanculescu, Ioan, Williams, Christopher K, Freer, Yvonne, et al. Autoregressive hidden markov models for the early detection of neonatal sepsis. *Biomedical and Health Informatics, IEEE Journal of*, 18(5):1560–1570, 2014a.

Stanculescu, Ioan, Williams, Christopher KI, and Freer, Yvonne. A hierarchical switching linear dynamical system applied to the detection of sepsis in neonatal condition monitoring. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014b.

Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc VV. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS) 27*, pp. 3104–3112, 2014.

Tresp, Volker and Briegel, Thomas. A solution for missing data in recurrent neural networks with an application to blood glucose prediction. In *Advances in Neural Information Processing Systems (NIPS) 10*, pp. 971–977. 1998.

Übeyli, Elif Derya. Combining recurrent neural networks with eigenvector methods for classification of ecg beats. *Digital Signal Processing*, 19(2):320–329, 2009.

Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, Dumitru. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164, June 2015.

Vohradský, Jiří. Neural network model of gene expression. *The FASEB Journal*, 15(3):846–854, 2001.

World Health Organization. *International statistical classification of diseases and related health problems*, volume 1. World Health Organization, 2004.

Xu, Rui, Wunsch II, Donald, and Frank, Ronald. Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 4(4):681–692, 2007.

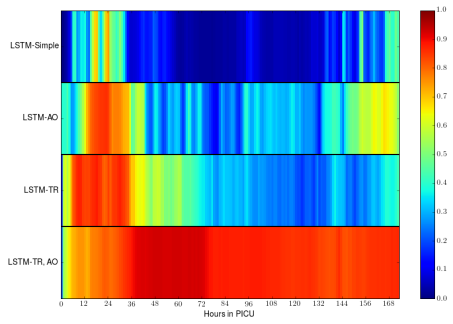
Yeung, Serena, Russakovsky, Olga, Jin, Ning, Andriluka, Mykhaylo, Mori, Greg, and Fei-Fei, Li. Every moment counts: Dense detailed labeling of actions in complex videos. *arXiv preprint arXiv:1507.05738*, 2015.

Zaremba, Wojciech, Sutskever, Ilya, and Vinyals, Oriol. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

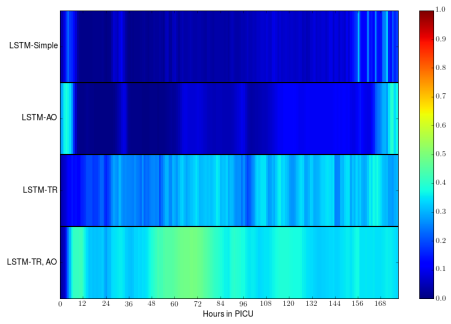
Appendices

A Hourly Diagnostic Predictions

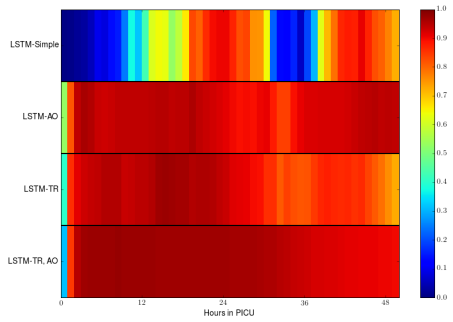
Our LSTM networks predict 128 diagnoses given sequences of clinical measurements. Because each network is connected left-to-right, i.e., in chronological order, we can output predictions at each sequence step. Ultimately, we imagine that this capability could be used to make continuously updated real-time alerts and diagnoses. Below, we explore this capability qualitatively. We choose examples of patients with a correctly classified diagnosis and visualize the probabilities assigned by each LSTM model at each sequence step. In addition to improving the quality of the final output, the LSTMs with target replication (LSTM-TR) arrive at correct diagnoses quickly compared to the simple multilabel LSTM model (LSTM-Simple). When auxiliary outputs are also used (LSTM-TR,AO), the diagnoses appear to be generally more confident.



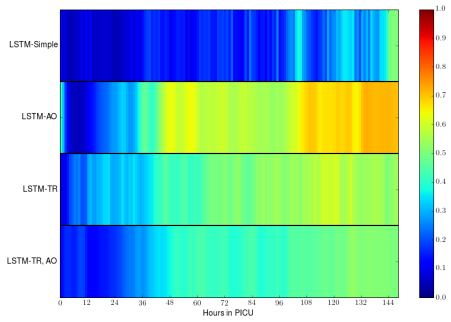
(a) Asthma with Status Asthmaticus



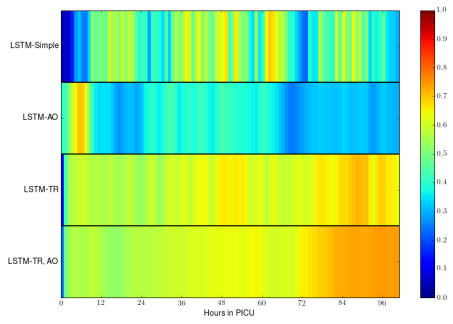
(b) Acute Respiratory Distress Syndrome



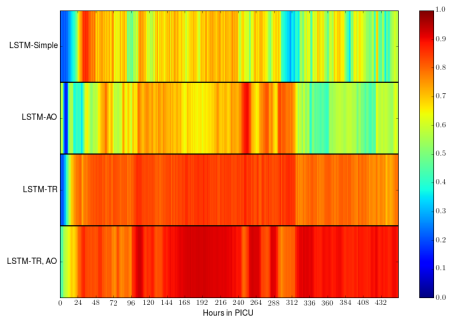
(c) Diabetic Ketoacidosis



(d) Brain Neoplasm, Unspecified Nature



(e) Septic Shock



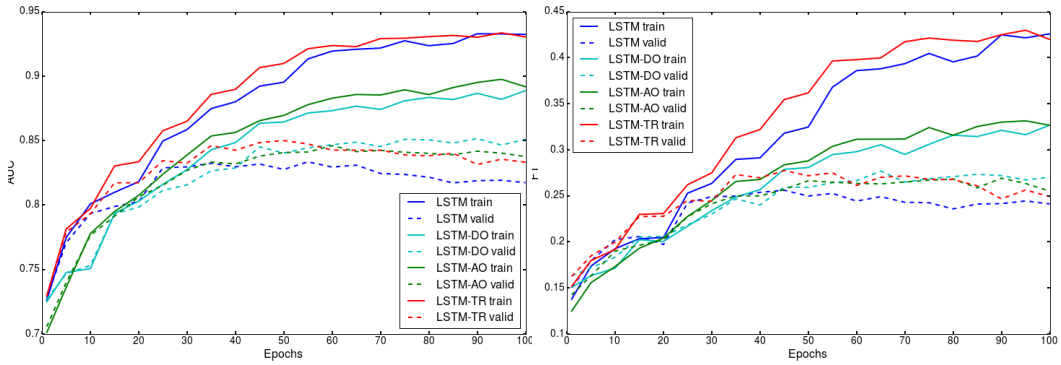
(f) Scoliosis

Figure 5: Each chart depicts the probabilities assigned by each of four models at each (hourly resampled) time step. LSTM-Simple uses only targets at the final time step. LSTM-TR uses target replication. LSTM-AO uses auxiliary outputs (diagnoses), and LSTM-TR,AO uses both techniques. LSTMs with target replication learn to make accurate diagnoses earlier.

Our LSTM-TR,AO effectively predicts status asthmaticus and acute respiratory distress syndrome, likely owing to the several measures of pulmonary function among our inputs. Diabetic ketoacidosis also proved easy to diagnose, likely because glucose and pH are included among our clinical measurements. We were surprised to see that the network classified scoliosis reliably, but a deeper look into the medical literature suggests that scoliosis often results in respiratory symptoms. This analysis of step-by-step predictions is preliminary and informal, and we note that for a small number of examples our data preprocessing introduces a target leak by back-filling missing values. In future work, when we explore this capability in greater depth, we will reprocess the data.

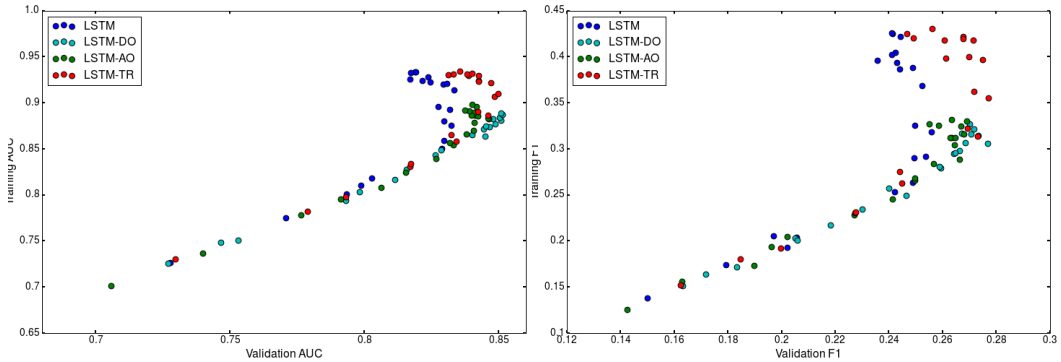
B LEARNING CURVES

We present visualizations of the performance of LSTM, LSTM-DO (with dropout probability 0.5), LSTM-AO (using the 301 additional diagnoses), and LSTM-TR (with $\alpha = 0.5$), during training. These charts are useful for examining the effects of dropout, auxiliary outputs, and target replication on both the speed of learning and the regularization they confer. Specifically, for each of the four models, we plot the training and validation micro AUC and F1 score every five epochs in Figure 6. Additionally, we plot a scatter of the performance on the training set vs. the performance on the validation set. The LSTM with target replication learns more quickly than a simple LSTM and also suffers less overfitting. With both dropout and auxiliary outputs, the LSTM trains more slowly than a simple LSTM but suffers considerably less overfitting.



(a) AUC learning curves

(b) F1 learning curves



(c) AUC training *vs.* validation

(d) F1 training *vs.* validation

Figure 6: Training and validation performance plotted for the simple multilabel network (LSTM-Simple), LSTM with target replication (LSTM-TR), and LSTM with auxiliary outputs (LSTM-AO). Target replication appears to increase the speed of learning and confers a small regularizing effect. Auxiliary outputs slow down the speed of learning but impart a strong regularizing effect.

C PER DIAGNOSIS RESULTS

While averaged statistics provide an efficient way to check the relative quality of various models, considerable information is lost by reducing performance to a single scalar quantity. For some labels, our classifier makes classifications with surprisingly high accuracy while for others, our features are uninformative and thus the classifier would not be practically useful. To facilitate a more granular investigation of our model's predictive power, we present individual test set F1 and AUC scores for each individual diagnostic label in Table 3. We compare the performance our best LSTM, which combines two 128-cell hidden layers with *dropout* of probability 0.5 and *target replication*, against the strongest baseline, an MLP trained on the hand-engineered features, and an ensemble predicts the maximum probability of the two. The results are sorted in descending order using the F1 performance of the LSTM, providing insights into the types of conditions that the LSTM can successfully classify.

Classifier Performance on Each Diagnostic Code, Sorted by F1

	LSTM-DO-TR		**MLP, Expert features**		**Max Ensemble**	
Condition	*F1*	AUC	F1	AUC	F1	AUC
Diabetes mellitus with ketoacidosis	0.8571	0.9966	0.8571	0.9966	0.8571	0.9966
Scoliosis, idiopathic	0.6809	0.8543	0.6169	0.8467	0.6689	0.8591
Asthma, unspecified with status asthmaticus	0.5641	0.9232	0.6296	0.9544	0.6667	0.9490
Neoplasm, brain, unspecified nature	0.5430	0.8522	0.5263	0.8463	0.5616	0.8618
Developmental delay	0.4751	0.8178	0.4023	0.8294	0.4434	0.8344
Acute respiratory distress syndrome (ARDS)	0.4688	0.9595	0.3913	0.9645	0.4211	0.9650
Hypertension, unspecified	0.4118	0.8593	0.3704	0.8637	0.3636	0.8652
Arteriovenous malformation of brain	0.4000	0.8620	0.3750	0.8633	0.3600	0.8684
End stage renal disease on dialysis	0.3889	0.8436	0.3810	0.8419	0.3902	0.8464
Acute respiratory failure	0.3864	0.7960	0.4128	0.7990	0.4155	0.8016
Renal transplant status post	0.3846	0.9692	0.4828	0.9693	0.4800	0.9713
Epilepsy, unspecified, not intractable	0.3740	0.7577	0.3145	0.7265	0.3795	0.7477
Septic shock	0.3721	0.8182	0.3210	0.8640	0.3519	0.8546
Other respiratory symptom	0.3690	0.8088	0.3642	0.7898	0.3955	0.8114
Biliary atresia	0.3636	0.9528	0.5000	0.9338	0.4444	0.9541
Acute lymphoid leukemia, without remission	0.3486	0.8601	0.3288	0.8293	0.3175	0.8441
Congenital hereditary muscular dystrophy	0.3478	0.8233	0.0000	0.8337	0.2727	0.8778
Liver transplant status post	0.3448	0.8431	0.3333	0.8104	0.3846	0.8349
Respiratory complications, prodecure status post	0.3143	0.8545	0.2133	0.8614	0.3438	0.8672
Grand mal status	0.3067	0.8003	0.3883	0.7917	0.3529	0.8088
Intracranial injury, closed	0.3048	0.8589	0.3095	0.8621	0.3297	0.8820
Diabetes insipidus	0.2963	0.9455	0.3774	0.9372	0.4068	0.9578
Acute renal failure, unspecified	0.2553	0.8806	0.2472	0.8698	0.2951	0.8821
Other diseases of the respiratory system	0.2529	0.7999	0.1864	0.7920	0.2400	0.8131
Croup syndrome	0.2500	0.9171	0.1538	0.9183	0.0000	0.9263
Bronchiolitis due to other infectious organism	0.2466	0.9386	0.2353	0.9315	0.2712	0.9425
Congestive heart failure	0.2439	0.8857	0.0000	0.8797	0.0000	0.8872
Infantile cerebral palsy, unspecified	0.2400	0.8538	0.1569	0.8492	0.2083	0.8515
Congenital hydrocephalus	0.2393	0.7280	0.2247	0.7337	0.1875	0.7444
Cerebral edema	0.2222	0.8823	0.2105	0.9143	0.2500	0.9190
Craniosynostosis	0.2222	0.8305	0.5333	0.8521	0.6154	0.8658
Anoxic brain damage	0.2222	0.8108	0.1333	0.8134	0.2500	0.8193
Pneumonitis due to inhalation of food or vomitus	0.2222	0.6547	0.0326	0.6776	0.0462	0.6905
Acute and subacute necrosis of the liver	0.2182	0.8674	0.2778	0.9039	0.2381	0.8964
Respiratory syncytial virus	0.2154	0.9118	0.1143	0.8694	0.1622	0.9031
Unspecified disorder of kidney and ureter	0.2069	0.8367	0.1667	0.8496	0.1667	0.8559
Craniofacial malformation	0.2059	0.8688	0.4444	0.8633	0.3158	0.8866
Pulmonary hypertension, secondary	0.2000	0.9377	0.0870	0.8969	0.2105	0.9343
Bronchopulmonary dysplasia	0.1905	0.8427	0.1404	0.8438	0.1333	0.8617
Drowning and non-fatal submersion	0.1905	0.8341	0.1538	0.8905	0.1429	0.8792
Genetic abnormality	0.1828	0.6727	0.1077	0.6343	0.1111	0.6745
Other and unspecified coagulation defects	0.1818	0.7081	0.0000	0.7507	0.1600	0.7328
Vehicular trauma	0.1778	0.8655	0.2642	0.8505	0.2295	0.8723

Table 3: F1 and AUC scores for individual diagnoses.

Classifier Performance on Each Diagnostic Code, Sorted by F1

	LSTM-DO-TR		MLP, Expert features		Max Ensemble	
Condition	*F1*	AUC	F1	AUC	F1	AUC
Other specified cardiac dysrhythmia	0.1667	0.7698	0.1250	0.8411	0.0800	0.8179
Acute pancreatitis	0.1622	0.8286	0.1053	0.8087	0.1379	0.8440
Esophageal reflux	0.1515	0.8236	0.0000	0.7774	0.1739	0.8090
Cardiac arrest, outside hospital	0.1500	0.8562	0.1333	0.9004	0.1765	0.8964
Unspecified pleural effusion	0.1458	0.8777	0.1194	0.8190	0.1250	0.8656
Mycoplasma pneumoniae	0.1429	0.8978	0.1067	0.8852	0.1505	0.8955
Unspecified immunologic disorder	0.1429	0.8481	0.1000	0.8692	0.1111	0.8692
Congenital alveolar hypoventilation syndrome	0.1429	0.6381	0.0000	0.7609	0.0000	0.7246
Septicemia, unspecified	0.1395	0.8595	0.1695	0.8640	0.1905	0.8663
Pneumonia due to adenovirus	0.1379	0.8467	0.0690	0.9121	0.1277	0.8947
Insomnia with sleep apnea	0.1359	0.7892	0.0752	0.7211	0.0899	0.8089
Defibrination syndrome	0.1333	0.9339	0.1935	0.9461	0.2500	0.9460
Unspecified injury, unspecified site	0.1333	0.8749	0.0000	0.7673	0.1250	0.8314
Pneumococcal pneumonia	0.1290	0.8706	0.1149	0.8664	0.1461	0.8727
Genetic or other unspecified anomaly	0.1277	0.7830	0.0870	0.7812	0.1429	0.7905
Other spontaneous pneumothorax	0.1212	0.8029	0.0972	0.8058	0.1156	0.8122
Bone marrow transplant status	0.1176	0.8136	0.0000	0.8854	0.2353	0.8638
Other primary cardiomyopathies	0.1176	0.6862	0.0000	0.6371	0.1212	0.6635
Intracranial hemorrhage	0.1071	0.7498	0.1458	0.7306	0.1587	0.7540
Benign intracranial hypertension	0.1053	0.9118	0.0909	0.7613	0.1379	0.8829
Encephalopathy, unspecified	0.1053	0.8466	0.0909	0.7886	0.0000	0.8300
Ventricular septal defect	0.1053	0.6781	0.0741	0.6534	0.0833	0.6667
Crushing injury, unspecified	0.1017	0.9183	0.0952	0.8742	0.1200	0.9111
Malignant neoplasm, disseminated	0.0984	0.7639	0.0588	0.7635	0.0667	0.7812
Orthopaedic surgery, post status	0.0976	0.7605	0.1290	0.8234	0.0845	0.8106
Thoracic surgery, post status	0.0930	0.9160	0.0432	0.7401	0.0463	0.9137
Ostium secundum type atrial septal defect	0.0923	0.7876	0.1538	0.8068	0.1154	0.7998
Malignant neoplasm, in gastrointestinal organs	0.0853	0.8067	0.1111	0.7226	0.1412	0.7991
Coma	0.0833	0.7255	0.1111	0.6542	0.1250	0.7224
Pneumonia due to inhalation of food or vomitus	0.0800	0.8282	0.0923	0.8090	0.0952	0.8422
Extradural hemorrage from injury, no open wound	0.0769	0.7829	0.0000	0.8339	0.0988	0.8246
Prematurity (less than 37 weeks gestation)	0.0759	0.7542	0.1628	0.7345	0.1316	0.7530
Asthma, unspecified, without status asthmaticus	0.0734	0.6679	0.0784	0.6914	0.0678	0.6867
Gastrointestinal surgery, post status	0.0714	0.7183	0.0984	0.6999	0.0851	0.7069
Nervous disorder, not elsewhere classified	0.0708	0.7127	0.1374	0.7589	0.1404	0.7429
Unspecified gastrointestinal disorder	0.0702	0.6372	0.0348	0.6831	0.0317	0.6713
Pulmonary congestion and hypostasis	0.0678	0.8359	0.0000	0.8633	0.0000	0.8687
Thrombocytopenia, unspecified	0.0660	0.7652	0.0000	0.7185	0.0000	0.7360
Lung contusion, no open wound	0.0639	0.9237	0.0000	0.9129	0.2222	0.9359
Acute pericarditis, unspecified	0.0625	0.8601	0.0000	0.9132	0.0000	0.9089
Nervous system complications from implant	0.0597	0.6727	0.0368	0.7082	0.0419	0.7129
Heart disease, unspecified	0.0588	0.8372	0.0000	0.8020	0.0000	0.8264
Suspected infection in newborn or infant	0.0588	0.6593	0.0000	0.7090	0.0606	0.6954

Classifier Performance on Each Diagnostic Code, Sorted by F1

	LSTM-DO-TR		MLP, Expert features		Max Ensemble	
Condition	*F1*	AUC	F1	AUC	F1	AUC
Anemia, unspecified	0.0541	0.7782	0.0488	0.7019	0.0727	0.7380
Muscular disorder, not elsewhere classified	0.0536	0.6996	0.0000	0.7354	0.1000	0.7276
Malignant neoplasm, adrenal gland	0.0472	0.6960	0.0727	0.6682	0.0548	0.6846
Hematologic disorder, unspecified	0.0465	0.7315	0.1194	0.7404	0.0714	0.7446
Hematemesis	0.0455	0.8116	0.0674	0.7887	0.0588	0.8103
Dehydration	0.0435	0.7317	0.1739	0.7287	0.0870	0.7552
Unspecified disease of spinal cord	0.0432	0.7153	0.0571	0.7481	0.0537	0.7388
Neurofibromatosis, unspecified	0.0403	0.7494	0.0516	0.7458	0.0613	0.7671
Intra-abdominal injury, no open wound	0.0333	0.7682	0.1569	0.8602	0.0690	0.8220
Thyroid disorder, unspecified	0.0293	0.5969	0.0548	0.5653	0.0336	0.6062
Hereditary hemolytic anemia, unspecifed	0.0290	0.7474	0.0000	0.6182	0.0000	0.6962
Subdural hemorrage, no open wound	0.0263	0.7620	0.1132	0.7353	0.0444	0.7731
Unspecified intestinal obstruction	0.0260	0.6210	0.2041	0.7684	0.0606	0.7277
Hyposmolality and/or hyponatremia	0.0234	0.6999	0.0000	0.7565	0.0000	0.7502
Primary malignant neoplasm, thorax	0.0233	0.6154	0.0364	0.6086	0.0323	0.5996
Supraventricular premature beats	0.0185	0.8278	0.0190	0.7577	0.0299	0.8146
Injury to intrathoracic organs, no open wound	0.0115	0.8354	0.0000	0.8681	0.0000	0.8604
Child abuse, unspecified	0.0000	0.9273	0.3158	0.9417	0.1818	0.9406
Acidosis	0.0000	0.9191	0.1176	0.9260	0.0000	0.9306
Infantile spinal muscular atrophy	0.0000	0.9158	0.0000	0.8511	0.0000	0.9641
Fracture, femoral shaft	0.0000	0.9116	0.0000	0.9372	0.0513	0.9233
Cystic fibrosis with pulmonary manifestations	0.0000	0.8927	0.0000	0.8086	0.0571	0.8852
Panhypopituitarism	0.0000	0.8799	0.2222	0.8799	0.0500	0.8872
Blood in stool	0.0000	0.8424	0.0000	0.8443	0.0000	0.8872
Sickle-cell anemia, unspecified	0.0000	0.8268	0.0000	0.7317	0.0000	0.7867
Cardiac dysrhythmia, unspecified	0.0000	0.8202	0.0702	0.8372	0.0000	0.8523
Agranulocytosis	0.0000	0.8157	0.1818	0.8011	0.1667	0.8028
Malignancy of bone, no site specified	0.0000	0.8128	0.0870	0.7763	0.0667	0.8318
Pneumonia, organism unspecified	0.0000	0.8008	0.0952	0.8146	0.0000	0.8171
Unspecified metabolic disorder	0.0000	0.7914	0.0000	0.6719	0.0000	0.7283
Urinary tract infection, no site specified	0.0000	0.7867	0.0840	0.7719	0.2286	0.7890
Obesity, unspecified	0.0000	0.7826	0.0556	0.7550	0.0000	0.7872
Apnea	0.0000	0.7822	0.2703	0.8189	0.0000	0.8083
Respiratory arrest	0.0000	0.7729	0.0000	0.8592	0.0000	0.8346
Hypovolemic shock	0.0000	0.7686	0.0000	0.8293	0.0000	0.8296
Hemophilus meningitis	0.0000	0.7649	0.0000	0.7877	0.0000	0.7721
Diabetes mellitus, type I, stable	0.0000	0.7329	0.0667	0.7435	0.0833	0.7410
Tetralogy of fallot	0.0000	0.7326	0.0000	0.6134	0.0000	0.6738
Congenital heart disease, unspecified	0.0000	0.7270	0.1333	0.7251	0.0000	0.7319
Mechanical complication of V-P shunt	0.0000	0.7173	0.0000	0.7308	0.0000	0.7205
Respiratory complications due to procedure	0.0000	0.7024	0.0000	0.7244	0.0000	0.7323
Teenage cerebral artery occlusion and infarction	0.0000	0.6377	0.0000	0.5982	0.0000	0.6507