Prediction of Cardiac Arrest from Physiological Signals in the Pediatric ICU

Sana Tonekaboni[1,2] STONEKABONI@CS.TORONTO.EDU
Mjaye Mazwi[3] MJAYE.MAZWI@SICKKIDS.CA
Peter Laussen[3] PETER.LAUSSEN@SICKKIDS.CA
Danny Eytan[3] DANNY.EYTAN@SICKKIDS.CA
Robert Greer[3] ROBERT.GREER@SICKKIDS.CA
Sebastian D. Goodfellow[2] SEBASTIAN.GOODFELLOW@SICKKIDS.CA
Andrew Goodwin[3] ANDREW.GOODWIN@SICKKIDS.CA
Michael Brudno[1] BRUDNO@CS.TORONTO.EDU
Anna Goldenberg[1,2] ANNA.GOLDENBERG@VECTORINSTITUTE.AI

[1] *Department of Computer Science*
University of Toronto
Toronto, Ontario, Canada
[2] *Vector institute*
Toronto, Ontario, Canada
[3] *Department of Critical Care Medicine*
The Hospital for Sick Children
Toronto, Ontario, Canada

Abstract

Cardiac arrest is a rare but devastating event in critically ill children associated with death, disability and significant healthcare costs. When a cardiac arrest occurs, the limited interventions available to save patient lives are associated with poor patient outcomes. The most effective way of improving patient outcomes and decreasing the associated healthcare costs would be to prevent cardiac arrest from occurring. This observation highlights the importance of prediction models that consistently identify high risk individuals and assist health care providers in providing targeted care to the right patient at the right time. In this paper, we took advantage of the power of convolutional neural networks (CNN) to extract information from high resolution temporal data, and combine this with a recurrent network (LSTM) to model time dependencies that exist in these temporal signals. We trained this CNN+LSTM model on high-frequency physiological measurements that are recorded in the ICU to facilitate early detection of a potential cardiac arrest at the level of the individual patient. Our model results in an F1 value of .61 to .83 across six different physiological signals, the most predictive single signal being the heart rate. To address the issue of instances of missing data in the recorded physiological signals, we have also

implemented an ensemble model that combines predictors for the signals that were collected for a given patient. The ensemble achieves .83 average F1 score on a held-out test set, on par with the best performing signal, even in the absence of a number of signals. The results of our model are clinically relevant. We intend to explore implementation of this model at the point of care as a means of providing precise, personalized, predictive care to an at-risk cohort of patients.

1. Introduction

Critically ill patients in the Intensive Care Unit (ICU) undergo dynamic changes in state and associated risk, requiring continuous, close monitoring to identify deterioration. Patient deterioration or clinical instability requires aggressive life-saving treatments or interventions (Vincent, 2013). Early identification of patient instability allows proactive interventions aimed at stabilization of the patients condition. To facilitate this earlier recognition monitoring devices are utilized that record and keep track of patient vital sign behavior at all times. This data is sampled at rates of up to 500Hz, creating massive amounts of information that can be overwhelming for physicians attempting to evaluate patients. However, this large and complex data is an ideal substrate for Machine Learning methods that have the potential to learn patterns and trends that exist within these data (Henry et al., 2015; Soleimani et al., 2017; Suresh et al., 2017).

Improved survival in critically ill patients relies on consistent identification and prediction of states that precede instability and patient injury, as this allows proactive measures to improve patient condition (Sandroni et al., 2006). This has motivated researchers to develop methods that predict destabilizing insults like Septic Shock (Henry et al., 2015), need for clinical interventions (Suresh et al., 2017), and disease severity (Ghassemi et al., 2015) using data generated by patients in the process of care. Cardiac arrest is a devastating complication of terminal clinical instability. It is classically defined as a requirement for chest compressions or cardiac defibrillation. In-hospital cardiac arrest in pediatric patients is strongly associated with mortality (survival to discharge in only 28 - 37 percent of patients), disability, cost and health care dependence (Ortmann et al., 2011). Effective models of prediction would improve recognition of pediatric patients at risk of cardiac arrest and enable medical and surgical interventions to avert impending arrest in select groups of patients. This would improve patient disability-free survival. However, a wide variety of physiological disturbances and clinical states can precede cardiac arrest making consistent, reproducible prediction challenging. Earlier work focused on predicting patient conditions of this sort have traditionally relied upon data elements available in the Electronic Health Record (EHR)(Verplancke et al., 2008; Choi et al., 2017; Wellner et al., 2017). More recent work has leveraged the density and continuity of physiological time series data for prediction of this complex patient behavior. This data also has the temporal signature required to support clinical decisions that arise in the dynamic process of care for the critically ill child. For higher frequency physiological recordings, previous authors have taken advantage of deep networks for learning relevant patterns in the signals that reflect on patient condition (Rajpurkar et al. (2017), (Razavian et al., 2016), Aczon et al. (2017)). Recurrent structures are also commonly used to incorporate the time-dependencies that exist in these signals (Suresh et al., 2017; Razavian et al., 2016; Lipton et al., 2016; Choi et al., 2017) as are probabilistic approaches. Ghassemi et al. (2015) Joseph Futoma (2017) that take

advantage of Gaussian Processes to estimate temporal clinical data. Survival analysis and time to event estimation are also used for prediction tasks using health data (Soleimani et al., 2017; Ranganath et al., 2016).

In this work, we designed a model that detects impending cardiac arrest in pediatric patients by examining patient's high-resolution longitudinal physiological recordings. This is done by associating a cardiac arrest risk score to windows of recording, in order to identify high-risk individuals. Taking advantage of the temporal nature of these signals, we use a long short-term memory network (LSTM) to learn the existing trend in the signals (Hochreiter and Schmidhuber, 1997).

Technical Significance Our method uses a combination of Convolutional and Recurrent network for the analysis of longitudinal physiological data. Combining these two structures allows us to take advantage of CNN's power in extracting informative features from time-series data and RNN's capability of learning time dependencies in temporal data. Since physiological signals can be densely packed with information, we use CNN to find a compact latent representation, and use that representation for a recurrent network to learn the existing pattern in signals prior to a cardiac arrest. We use an ensemble of models trained on different signal types to assign a cardiac arrest risk score to windows of recordings. Also, in order to tackle the problem of missing measurements within recordings, we use a low-pass filtering method for sample imputation.

Clinical Relevance Even in experienced centers, in-hospital cardiac arrest in critically ill pediatric patients is associated with survival to discharge in only 28 to 37% of patients. Disability free survival is rare with the most relevant category of disability being neurological injury (Ortmann et al., 2011). Limited treatments are available when cardiac arrest occurs and include cardiopulmonary resuscitation (CPR) with or without surgical cannulation to an extracorporeal pump. These both represent high-risk, necessary interventions aimed at rescuing patients from certain death, but both contribute to the injury associated with cardiac arrest. The most effective way of improving the dismal outlook for pediatric patients at risk of cardiac arrest is proactive identification of the patient at risk and prevention of cardiac arrest with the need for resuscitation. In addition to saving lives and decreasing long-term disability, this would significantly reduce health care costs associated with the complex needs of survivors (Dimick et al., 2004).

2. Cohort

For this study, we utilized high-resolution physiological signals collected and archived from the Pediatric ICU at the Hospital for Sick Children in Toronto Ontario to predict the onset of a cardiac arrest. These recordings are sampled at a frequency of every 5 seconds pervasively for all patients throughout their ICU stay, and include measured vitals such as heart rate, blood pressure, and respiratory rate, as well as other derived metrics created by mathematical modification of measured vital signs, such as SDNN (a heart rate variability metric). Inclusion criteria were any critically ill child admitted to the critical care unit during the period of review who experienced a cardiac arrest. This subset of patients experienced either a single event or multiple cardiac arrests during their ICU course. All data recordings are labeled with a verified time of the cardiac arrest event. For each subject,

we extracted the physiological signals that were collected in their medical record in the 24 hours prior to the arrest. The distribution of number of available signals for each individual can be seen in Figure 1. In total, we had 228 instances of cardiac arrests across 208 patients.

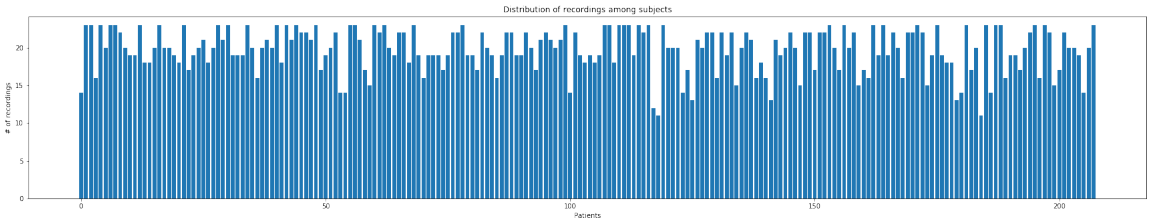


Figure 1: The total count of unique physiological signals that are available for each patient in the cohort.

2.1. Cohort Selection

Despite the large number of physiological signals that are able to be recorded for patients in the ICU, not every individual has a comprehensive set of all of the measurements recorded in the database due to variation in monitoring strategy based upon clinician preference or patient characteristics. Typical ICU practice is to record vital signs based on perceived patient need and presence of the relevant sensor as part of the patients care array. As an example, no End Tidal Capnogram is recorded if the patient is not intubated at the time of the arrest. As a result, we had to select a subset of the signals and patients for our analysis. As shown in figure 2, among the 110 different types of signals available, many were only present for a few patients. Therefore, our inclusion criteria for the signal types and subjects were as follows:

1. Select vitals that are recorded for at least 100 subjects in the database
2. Select subject who have at least one of the selected vitals on their record.

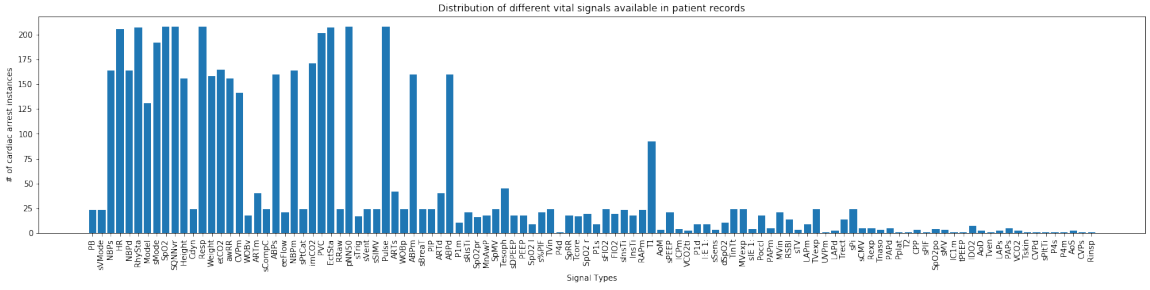


Figure 2: Number of different signal types present in the entire dataset. As shown here, some signals like heart rate are commonly collected for patients while others like end tidal CO2(etCO2) are only collected for a subset of patients

Six physiological signal types were selected for the analysis of cardiac arrest. Table 1 lists the name of those signals along with the number of subjects in the database that had these measures recorded. Also, for patients who experienced multiple arrests, we removed the second arrest instance if this occurred too close to the first one. This was to avoid having predictive windows that overlap with arrests.

Table 1: List of physiological signals used for the prediction and the number of cardiac arrest instances for each recording

Signal	Description	Number of arrest instances
HR	Heart rate	203
Resp	Respiratory rate	203
Pulse	Pulse rate	203
SpO2	Oxygen saturation level	203
ABP	Arterial Blood Pressure	155
SDNN	Standard deviation of R-R intervals	203

2.2. Data Extraction and Preprocessing

After the inclusion criteria were applied there were 203 cardiac arrest instances from 171 unique patients. For each instance, we had a different number and combination of recorded physiological variables available for review. There are two main sources of sparsity in the data used for this study:
1) Missing measurements within a specific physiological signal and
2) absence of an entire type of signal for some individuals if that was not part of the monitoring strategy being employed at that bedside.

To deal with missing values that exist within signals, we have implemented a combination of low-pass filtering and linear interpolation. The filter averages existing signal values over a length l window of time ($l = 3$ minutes in our implementation). As shown in Figure 3 in the first step, the filter is able to smooth over missing values in blocks $< len(l)$, and it also helps to remove the high-frequency noise. For any remaining discontinuity, we used linear interpolation for signal imputation. One advantage of the low-pass filtering method is that it can easily be implemented for real-time processing of the signals. As new samples are generated, if an individual value is missing, it will be replaced by the average of the other samples of the filter window.

To handle the problem of missing entire signals for certain patients, we have implemented an ensemble model, that is further explained in the Method section.

2.3. Feature Choices

Along with the pre-processed signals, we also included patient's age in the analysis, as it is assumed to have a confounding effect on the arrest prediction task. No manual feature selection was done on the temporal signals, as we used convolutional neural networks to

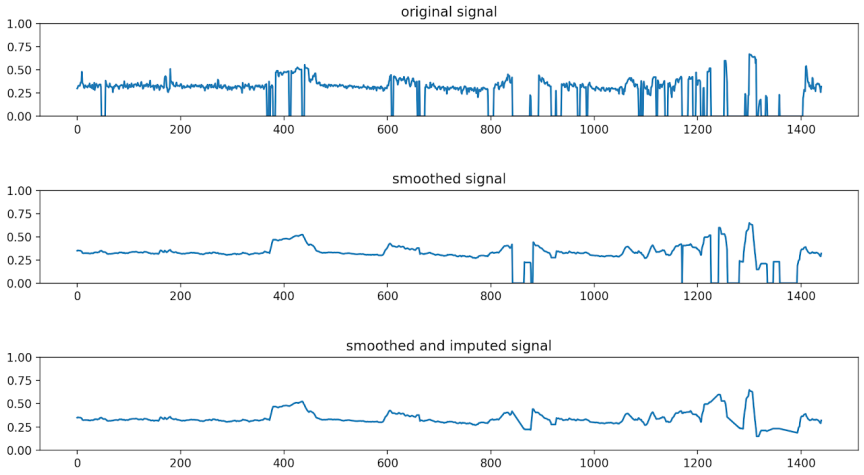


Figure 3: Preprocessing steps of the physiological signals. The first plot on the top shows the normalized recording with the missing values replaced with zeros. The second plot is the signal after low-pass filtering. The third plot is the final signal after linear imputation of missing values that weren't resolved with filtering

extract the hidden information that exists within each recording. The implementation details of the CNN are further discussed in the Methods section.

3. Methods

Our goal was to learn the underlying pattern that exists in patient's physiological recordings, prior to a cardiac arrest. The risk predictor model takes in a length T minutes window of a physiological signal, and assigns a risk score to it, based on how similar it is to pre-arrest patterns. Looking into the trend of these scores during a patient's stay at the hospital can help clinicians identify individuals at high risk for a cardiac arrest by defining a patient trajectory.

In the following section we describe the building blocks of the risk predictor model: 1)A convolutional neural network that was used for feature extraction and for finding a compact latent representation of the signal; 2) An LSTM network that captures the time-dependency in the recordings; 3) The ensemble predictor that combines risk scores are evaluated from different physiological signals. Following the section on network structure, we present our method for generating training samples that makes the training process more robust and prevents overfitting.

3.1. Convolutional Neural Network(CNN)

We use 1-dimensional CNN, in order to learn and extract features from the time series signals. Without a fully-connected layer, the output of the network can be considered as a

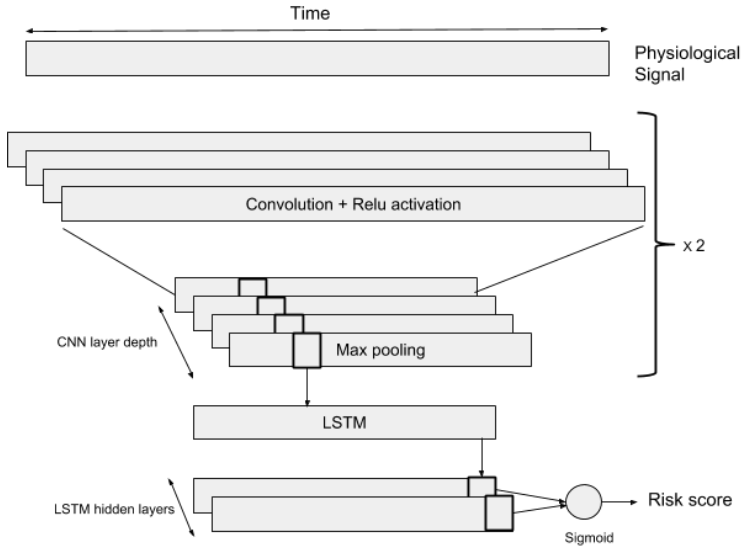


Figure 4: The CNN+LSTM network structure. A window of physiological signal is feed into the convolutional network and the outputs of the CNN are used by the LSTM to generate the Risk score

latent representation of the original signal in time. This method is used instead of hand-picking features as it is expected to be more robust and to generalize well to different types of recordings (as was also previously shown in (Razavian et al., 2016)). As demonstrated in Figure 4, the network is composed of 2 convolutional layers. Each layer followed by *tanh* activation and max-pooling. Also, dropout is applied to the elements of the final layer, to prevent overfitting. All hyper-parameters are set by k-fold cross-validation. The output of the network, $f_{1:T'}$, is a condensed latent representation ($f \in \mathcal{R}_{T'}$), of the original signal ($x \in \mathcal{R}_T$), where $T' < T$, and is the input to the LSTM that will generate the final risk score.

3.2. Long Short-Term Memory Network(LSTM)

The extracted representation $f_{1:T'}$ goes into an LSTM network with $n = 16$ hidden units, that generates 16 hidden states and 16 output states for every time step. We call the output states $h \in \mathcal{R}_T^n$ and used the value of the hidden units of the final outputs to generate a primary risk probability r (Figure 5). This score is calculated by feeding the final output to a simple one layer network with *Sigmoid* activation (As shown in Figure 5). The recurrent structure of the LSTMs encodes relative information of measurements across time and provides a nonlinear improvement in model generalization.

$$h_1^0, ..., h_{t'}^0, ..., h_1^n, ..., h_{t'}^n = LSTM(f_1 ... f_{t'}) \tag{1}$$

$$r = Sigmoid(h_{t'} W + b) \tag{2}$$

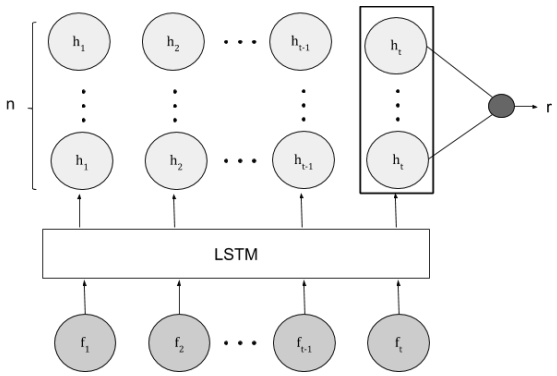


Figure 5: The LSTM structure that accepts the latent representation of the physiological signal and generates the primary risk score using the hidden layers of its output

In order to generate the final risk score, we took into account another confounding factor: patient's age. Specifically, we use patient's age to generate a scaling factor for the final score. As shown in Eq. 3 this scale is calculated using a small neural network with weights W_{age} and bias b_{age}. The final risk score ($risk$) is then evaluated using Eq. 4 by scaling the primary risk probability r.

$$agescale = RELU(age * W_{age} + b_{age}) \tag{3}$$

$$risk = Sigmoid(r * agescale) \tag{4}$$

3.3. The Ensemble

There is a substantial discontinuity in our data where certain types of recordings for patients during their stay at the hospital are not available due to interruption of a recording device, movement of the patient within the unit or to a procedure or absence of a physiological signal altogether as a result of that signal not being part of that patient's monitoring strategy. As a result, we trained separate risk predictor models for each physiological signal independently. The final risk score is then the weighted aggregate of the risk predictions generated from existing signals for an individual. The performance of the models across signals vary, depending on the number of samples available for training and also the amount of information each signal carries by nature. We assigned different weights to each of the models W_i (where i is a given physiological signal), based on their prediction performance during the training and validation process. This weight is the average F1 score calculated on K different folds during the training. The ensemble generates the final risk score as the weighted average of the predicted risks from the existing signals in a patient's record (Eq. 5). The performance is then tested on held out samples that were not used at any point during the training of any of the models.

$$Riskscore = \frac{\Sigma_{i=1}^{n}(W_i * risk_i)\mathbb{1}_i}{\Sigma_{i=1}^{n}(W_i)\mathbb{1}_i} \tag{5}$$

Where $\mathbb{1}_i = 0$ if signal i is missing and $\mathbb{1}_i = 1$ otherwise

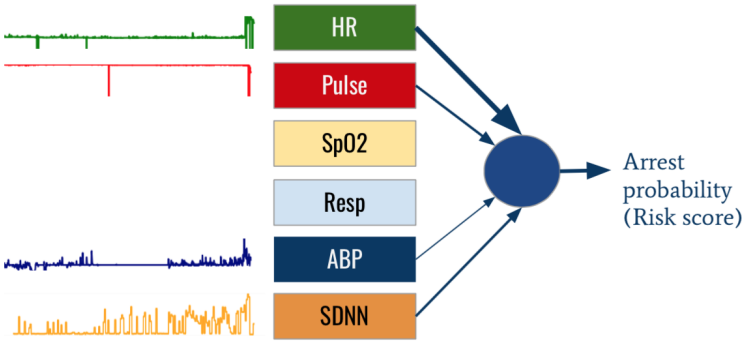


Figure 6: Risk score prediction using the ensemble model with existing physiological signals

3.4. Building Training Samples

For training, we needed to learn the pattern from windows of physiological recordings. We extracted windows of length T from the D selected types of vital signals. We considered recordings that are earlier than Δ hours prior to the arrest (e.g. we used 5 hours prior to the arrest), as normal signals, and recordings anywhere between 2 hours and 5 minutes prior to the arrest as representing periods of risk. The learning task involved being able to distinguish patterns that exist in normal signal versus signals in the "pre-arrest range" in order to flag the high risk patients earlier. To prevent overfitting to small sample size for certain types of recordings and also to avoid dependance on reference time points, we randomly pick windows of length T from the pre-arrest and normal ranges every few epochs during training (as shown in Figure 7) as we continue optimizing the network parameters.

4. Results

4.1. Evaluation Approach/Study Design

4.2. Evaluation Approach

We evaluated how well we could distinguish normal windows of physiological signals from pre-arrest ones. Given the predicted risk scores and using a threshold of 0.5, predictions are labeled and classified as 'pre-arrest' and 'normal'. To evaluate the performance of our classification based on this threshold, we use the F1 score (defined in Eq. 6) that combines information about the sensitivity and specificity of the prediction. Both Recall and Precision values are valuable in our model, as Recall determines the number of pre-arrest windows we were able to identify, and Precision shows the accuracy of our positive predictions.

$$F1 = \frac{2 * Recall * Precision}{Recall + Precision} \tag{6}$$

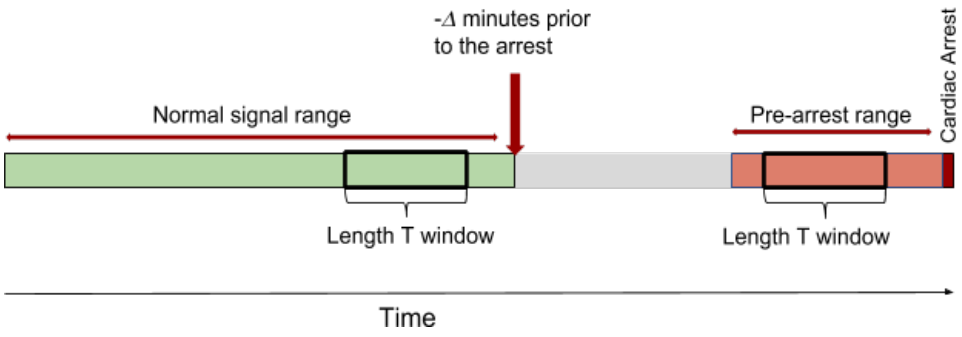


Figure 7: Random selection of time windows from the pre-arrest and normal region of a physiological recording. The normal region(colored in green), is part of the signal that is far from the cardiac arrest, in our case more Δ minutes earlier. The pre-arrest range (colored in red) includes any window of recording that is closer than 5 minutes to the cardiac arrest.

Signal Type	CNN	LSTM	CNN+LSTM
Heart rate (HR)	0.753 ± 0.075	0.666 ± 0.118	0.816 ± 0.034
Respiratory rate (Resp)	0.739 ± 0.043	N/A	0.67 ± 0.05
Pulse rate (Pulse)	0.720 ± 0.093	0.698 ± 0.06	0.746 ± 0.092
Oxygen Saturation level (SpO2)	0.714 ± 0.078	0.617 ± 0.101	0.723 ± 0.055
Ambulatory Blood Pressure (ABP)	0.675 ± 0.17	0.647 ± 0.17	0.76 ± 0.13
SDNN Heart Rate Variability Metric (SDNN)	0.663 ± 0.080	N/A	0.692 ± 0.07

Table 2: F1 scores for prediction, averaged over 5 cross-validation folds. Results for LSTM training are not reported as they took longer than 48 hours to converge

4.3. Results for individual Models

Table 2 shows the F1 prediction scores for models trained on each of the physiological signals independently. The values are evaluated using a 5-fold cross validation and the scores are compared across 3 different network structures: 1) CNN , 2) LSTM trained on raw signals and 3) CNN+LSTM structure that uses the latent representation of the signals as input to the LSTM network as described above.

From Table 2 we observe that the LSTM model, encoding the time dependence of observations from raw data, performs considerably worse than the other two structures. We attribute this performance to the noise intrinsic to the physiological data, illustrating that identifying robust features from the raw data is of great importance in training LSTM networks. We further show that the performance of CNN+LSTM improves on the CNN performance across 5 out of 6 signals. This is not unexpected, given that our CNN+LSTM

	HR	Resp	Pulse	SpO2	ABP	SDNN
Recall	0.769	0.692	0.714	0.444	0.857	0.909
Precision	0.90	0.642	0.8333	1.0	0.6	0.714
F1 score	0.833	0.666	0.769	0.6153	0.705	0.8

Table 3: Performance of the models on the unobserved test set that includes 20 individuals with one incident of cardiac arrest

architecture takes advantage of both robust feature selection of CNN and time dependence encoding of LSTM.

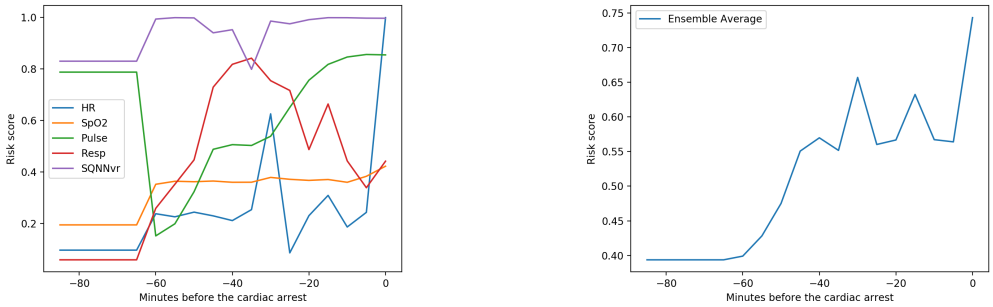
Further, we tested the performance of CNN+LSTM models trained for each of the signals, on a previously unseen test cohort, consisting of 20 patients (with at least one cardiac arrest each). Our results shown in Table 3 are comparable and even slightly better than our cross validation performance in Table 2. We attribute the slight improvement to the fact that we had more samples available for training of the final model since we used the whole training set and not folds.

4.4. Results for the Ensemble Model

We also looked into how the ensemble can benefit the prediction results. Using subjects from our test cohort, in Table 4 we showed how single models perform on their own compared to the ensemble. Moving windows of signals were selected from these patients' physiological recordings, and for every window, each model generated a risk score. Note that in addition to instances of signals missing entirely, even for some signals, a number of windows did not have any observations. As seen in the scores from single models, missing windows can dramatically impact the performance. However, for all of the individuals that we have observed (4 are captured in Table 4), the ensemble is able to improve on the performance over models based on the individual measurements. An interesting observation is how some models perform better than others for different subjects. This is mainly because of the fact explained in previous sections that there is no single pathway that leads to a cardiac arrest. The arrest can be caused by respiratory failure, cardiac failure, or many other reasons, and different models may be able to capture these signs.

To evaluate the performance of the model on non-cardiac arrest patients, we have used another cohort of 20 subjects who have been in the ICU for at least 24 hours, but had never experienced a cardiac arrest. In the test setting, risk scores were generated on sliding windows in time, every 10 minutes, for 24 hours, and similar to previous sections, scores higher than 0.5 generated an alarm. The false positive rate of our classifier was 0.012 on these samples.

In Figure 8, we show the predicted risk score as a function of the difference between the right boundary of the sliding window and the event for a random patient from the test set. Figure 8(a) plots the generated score from models based on a single type of signal, and we can see that it is highly volatile and very sensitive to the quality of the signal in that window of time. Figure 8(b) shows the ensemble risk predictions that are less volatile.



(a) Predicted risk score from individual models (b) Final predicted risk score from the ensemble

Figure 8: Risk score prediction for a sliding window in time

Subject	Ensemble	HR	Resp	Pulse	SpO2	ABP	SDNN
1	0.857	0.8	0.5	0.3636	0.6	0.75	0.66
2	0.66	0.5	0.33	0.54	0.6	missing	0.33
3	0.857	0.66	0.33	0.85	0.22	missing	0.3
4	0.666	0.66	0.6	1.	0.57	0.2	0.3636

Table 4: F1 scores of single models and the ensemble model for subjects from the test cohort with missing values

Looking into the trend of risk scores, signs of an arrest building up can be observed as early as 30 minutes prior to the event.

5. Discussion and Related Work

In this work, we presented a general architecture for early prediction of cardiac arrest, from high-resolution physiological signals that is clinically relevant at the point of care. Previous work has shown promising results using recurrent structures for learning and prediction from time series health signals (Lipton et al. (2016)). In their work, (Razavian et al. (2016)) used LSTM to model physiological signal at a much lower resolution, for prediction of disease onset from patient lab results. Their work showed that using recurrent neural structure to model low resolution physiological signal leads to good performance. Choi et al. (2017) also uses an extension of the RNN framework, called gated recurrent units (GRU), for predicting heart failure from EHR data such as diagnostic codes and medication prescription. These EHR measures are also recorded less frequently than our 5-sec resolution physiological measures and over long periods of time, which distinguishes their prediction task from ours, as they look into months of observations to identify individuals with high risk of heart failure. In temporal signals with more frequent measurements, such as the ones presented in this work, the performance of the LSTM can be sensitive to the high-frequency noise

that might be present in the signal which results in lower performance as is shown in Table 2 in the Results section of the paper.

Suresh et al. (2017) used physiological recordings from the ICU to predict different clinical interventions. The authors have demonstrated that representing these signals with discretized words (using a Z score measure) improves the performance of the LSTMs considerably compared to training the LSTM on raw data. In general, training recurrent structures on time series signal with high noise level, has a high risk for the network to only learn the naive solution Giles et al. (1999) and for the performance to be sensitive to small fluctuations in input signal values. A number of work in speech recognition and image processing, combine deep network structures with recurrent networks, in order to take advantage of both the power of DNNs in reducing frequency variation and temporal modeling in LSTMs (Sainath et al., 2015; Xu et al., 2016) as we have done in our paper.

In our work, we were able to improve the performance of LSTMs by using a latent representation of the original signal as the input to the network. We use convolution networks to extract a compact latent representation of the time series data, prior to using an LSTM for the prediction task. This new structure not only improved prediction performance compared to an LSTM network trained on raw recordings but also reduced the training time considerably (by order 10). Looking into the results from Table 2, the 'CNN+LSTM' model outperforms LSTM on prediction for all physiological signals.

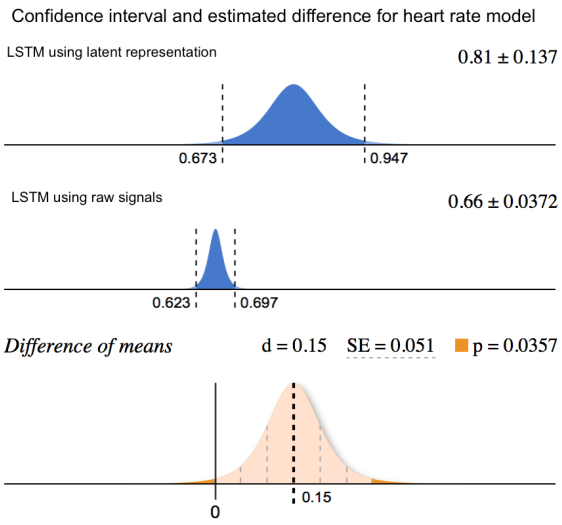
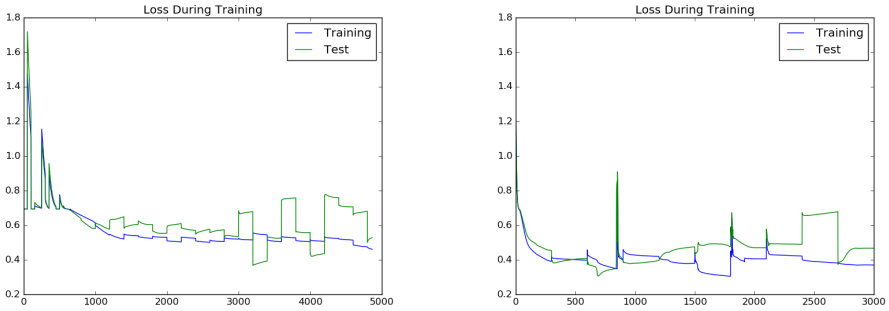


Figure 9: Evaluating the significance in the difference in F1 results for the Heart Rate model using LSTM on raw signals and LSTM+CNN

In addition, using latent representation adds more stability to the model in most cases, as can be inferred from standard deviation values reported in 2. We can also see that in Figure 10, training LSTM on raw signals can be unstable and sensitive to variation in the signal. Looking at Figure 10, we see sudden changes in the loss at regular intervals during the training. These are epochs where we randomly select new windows of recordings. Using



(a) Training LSTM

(b) Training LSTM+CNN

Figure 10: Comparing model performance during training for the LSTM model using ran physiological data and LSTM+CNN structure that uses the latent representation of the signal. Plots show the value of loss function on both test and training set.

the latent representation as inputs to the LSTM 10(b), gives the training process more stability and robustness.

In this work, we were able to tackle two main challenges common to clinical data: 1) sparsity and 2) complexity(Alistair E. W. Johnson, 2016). Our low-pass filtering method handles short timescale discontinuities in temporal signals, while longer discontinuities were addressed using linear interpolation. We showed that the ensemble method improves the prediction performance by aggregating risk predictions from different models. Using a mixture of experts makes the model robust to cases where certain physiological signals are highly corrupted (due to issues with recording hardware) or missing entirely (not recorded at all). This means our model can generate a risk score for a window of recordings even with a minimum of one existing signal, but as more recordings are available, the results will be more confident and not sensitive to the noise likely present in a single recording.

A next step for this model will involve exploring how to deploy a model like this at the point of care, as a predictive tool. As patients' physiological signals are being recorded, our model generates a cardiac arrest risk score for a sliding window in time. This score, is independent of previous predictions. For calculating the personalized risk we will rely on the output of the predictor model, as well as the trend of a patient's clinical status during their stay in the ICU. Ultimately, we expect our predictive model to be implemented in the ICU and are exploring a translational framework. Translation will require better characterization of the patient populations in whom the model performs particularly well as this cohort would be the group who would most benefit from model deployment (i.e. maximization of true positives). Implementation is a complex task for models that make frequent predictions as an implementation plan has to define acceptable thresholds for false positives and true negatives associated with model performance at the point of care. This is essential in order to minimize the likelihood of compounding the well document problem of monitor alarm fatigue amongst clinicians (Keller, 2012).

Approaches of the sort described here offer promising avenues for predictive rather than reactive medicine provided in a highly personalized framework where clinical decision support can be provided at the level of the individual patient.

References

Melissa Aczon, David Ledbetter, Long Van Ho, Alec M. Gunny, Albert Flynn, J. Williams, and Randall C. Wetzel. Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks. *CoRR*, abs/1701.06675, 2017.

Shamim Nemati Katherine E. Niehaus David A. Clifton Gari D. Clifford Alistair E. W. Johnson, Mohammad M. Ghassemi. Machine learning and decision support in critical care. In *Proceedings of the IEEE*, pages 444 – 466, 2016.

Edward Choi, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. In *JAMIA*, 2017.

Justin B. Dimick, Steven L. Chen, Paul A. Taheri, William G. Henderson, Shukri F. Khuri, and Darrell A. Campbell. Hospital costs associated with surgical complications: A report from the private-sector national surgical quality improvement program. *Journal of the American College of Surgeons*, 199(4):531537, 2004. doi: 10.1016/j.jamcollsurg.2004.05.276.

Marzyeh Ghassemi, Marco A. F. Pimentel, Tristan Naumann, Thomas Brennan, David A. Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015:446–453, 2015.

C. Lee Giles, Steve Lawrence, A. C. Tsoi, and Ah Chung Tsoi. Noisy time series prediction using a recurrent neural network and grammatical inference. 1999.

Katharine Henry, D. Hager, Peter J. Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7 299: 299ra122, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 8:1735–80, 1997.

Katherine Heller Mark Sendak Nathan Brajer Meredith Clement Armando Bedoya Cara OBrien Joseph Futoma, Sanjay Hariharan. An improved multi-output gaussian process rnn with real-time validation for early sepsis detection. In *Machine Learning for Healthcare*, 2017.

James P Keller. Clinical alarm hazards: a "top ten" health technology safety concern. *Journal of electrocardiology*, 45 6:588–91, 2012.

Zachary Chase Lipton, David C. Kale, Charles Elkan, and Randall C. Wetzel. Learning to diagnose with lstm recurrent neural networks. *ICLR*, abs/1511.03677, 2016.

L. Ortmann, P. Prodhan, J. Gossett, S. Schexnayder, R. Berg, V. Nadkarni, and A. Bhutta. Outcomes after in-hospital cardiac arrest in children with cardiac disease: A report from get with the guidelines-resuscitation. *Circulation*, 124(21):23292337, 2011. doi: 10.1161/circulationaha.110.013466.

Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. *CoRR*, abs/1707.01836, 2017.

Rajesh Ranganath, Adler J. Perotte, Noémie Elhadad, and David M. Blei. Deep survival analysis. In *MLHC*, 2016.

Narges Razavian, Jake Marcus, and David Sontag. Multi-task prediction of disease onsets from longitudinal lab tests. *CoRR*, abs/1608.00647, 2016.

Tara N. Sainath, Oriol Vinyals, Andrew W. Senior, and Hasim Sak. Convolutional, long short-term memory, fully connected deep neural networks. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4580–4584, 2015.

Claudio Sandroni, Jerry Nolan, Fabio Cavallaro, and Massimo Antonelli. In-hospital cardiac arrest: incidence, prognosis and possible measures to improve survival. *Intensive Care Medicine*, 33:237–245, 2006.

Hossein Soleimani, James Hensman, and Suchi Saria. Scalable joint models for reliable uncertainty-aware event prediction. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

Harini Suresh, Nathan Hunt, Alistair Edward William Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. *CoRR*, abs/1705.08498, 2017.

Thierry Verplancke, Stijn Van Looy, Dominique D Benoit, Stijn Vansteelandt, Pieter Depuydt, Filip De Turck, and Johan Decruyenaere. Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. *BMC Medical Informatics and Decision Making*, 8:56 – 56, 2008.

Jean-Louis Vincent. Critical care - where have we been and where are we going? In *Critical care*, 2013.

Ben Wellner, Joan Grand, Elizabeth Canzone, Matt Coarr, Patrick W Brady, Jeffrey Simmons, Eric S. Kirkendall, Nathan Dean, Monica Kleinman, and Peter Sylvester. Predicting unplanned transfers to the intensive care unit: A machine learning approach leveraging diverse clinical elements. In *JMIR medical informatics*, 2017.

Zhenqi Xu, Shan Shan Li, and Weihong Deng. Learning temporal features using lstm-cnn architecture for face anti-spoofing. *IEEE Xplore*, 2016.