Research Report - Sophia Sun
JHU CSMR REU 2017
Advisor: Suchi Saria Ph.D.

Abstract

This summer we laid some groundwork for Saria Lab's acute cardiac determination early warning system. Specifically, a ETL (extract, transform, load) data pipelines for both routinely available physiological/laboratory data and high frequency bedside monitor data were built, feature engineering was explored, and some baseline prediction models were developed.

The baseline models were trained to predict onsets of cardiogenic shock, cardiac arrest, and acute heart failure. Through these models, we proved that cardiac deteriorations can be feasibly modeled using machine learning algorithms, especially XGgboost and LSTM neural networks. Early warning and prediction from real-time data are achievable. The system in the future may allow clinicians to identify patients at risk for cardiac deterioration and provide earlier interventions that would prevent or mitigate the associated morbidity and mortality.

Introduction

Cardiovascular diseases are the number one cause of death globally.[1] In the recent decades, there is increasing prevalence of cardiogenic shock in intensive care units, resulting in overcrowded resources and increased cost. [2]

Studies have shown that more than half of cardiac arrests can be prevented based on clinical evidence of deterioration 8 hours prior to the event.[3,4] If a cardiac arrest has occurred already, a quick response can decrease the mortality rate by 25%. [4, 5] However, the inability to correctly identify patients with sufficient intervention time limits the effectiveness of emergency response teams.[6] Therefore, an automatic risk assessment and early warning system can be critical to improving the survival rate and lowering both length of stay for patients and cost to the hospital.

Existing warning systems mostly consist of early warning scores, such as the Modified Early Warning Score (MEWS) [7] and England's National Early Warning Score (NEWS) [8]. These systems are based primarily on expert opinion thus have limited scientific validation [6], and are typically not effective at warning about specific cardiac deteriorations [8], which is essential for informing physicians to take effective early intervention.

The increasing prevalence of electronic health records (EHR) in health care has enabled development of more sophisticated and specialized prediction systems, as more data is collected with higher frequency and quality. Previous research on EHR data have saw effective prediction of adverse events such as septic shock [9], post-procedure complications [10], resuscitation events, and death [11].

My work during the REU program this summer is part of a larger project of the lab, in collaboration with cardiologists from JHMI. The project's goal is to develop a machine learning system to predict, identify, and warn patients' risk of developing cardiac deteriorations, as well as identify the top factors for such risk. My work has focused on literature review, construction of the ETL (extract, transform, load) data pipeline, and development of several proof-of-concept baseline prediction models. This report will elaborate on each of these three aspects.

Literature Review

General-purpose illness severity scoring systems are well-adapted and recommended regarding the early recognition and response to patient deterioration. Efforts have been made to incorporate time-aware data into these scores [12]. Other specially designed scoring systems such as the National Early Warning Score (NEWS) has been tested to discriminate patient in danger of cardiac arrest, death, and ICU transfers with a high AUROC score of 0.87 [8]. However, these physiological scoring systems suffers from sensitivity and specificity issues when it comes to warning specific conditions. [8, 12] This study takes all features used in these scores into modeling account.

Electronic Heath Records (EHR) data enables researchers to construct more sophisticated prediction and early warning systems. Recent research work has explored utilizing EHR data and machine learning for adverse event predictions, risk assessment, and diagnosis.

A common method for predicting deteriorations or adverse events utilizes general linear models such as multivariable logistic regression. Alvarez et. al. used a lasso technique on EHR data to construct a logistic model that predicts out of intensive care unit onset of resuscitation events, including several acute heart conditions, and death.[11] Cao et. al. predicted hemodynamic stability in ICUs using logistic regression on short-term trends.[13] Mortazavi et. al. also used general linear models as baselines in developing their prediction system of adverse postoperative complications.[10] Other classic machine learning classification algorithms such as SVMs and Boosting are also featured in researches as primary or supporting prediction methods. [10, 14] A major drawback for classic classification models is that there is no native representation of time. Trends of values are inputted either as separated features[15] or as engineered features such as maximum or standard deviation over time[12].

Other paradigms of modeling have also been explored in the context of EHR adverse event prediction. Henry et al. developed an advanced model for real-time estimation of septic shock (TREWScore) based on Cox proportional hazards and lasso regularization. [9] Artificial neural network have also been proven to be able to diagnose[15, 16] and identify abnormal events[17] from EHR-related data streams.

The data pipeline and design of baseline models in this study are set up similarly to [9] and [10], to explore three of the aforementioned modeling paradigms – classification models, survival analysis regression, and neural networks.

Procedures

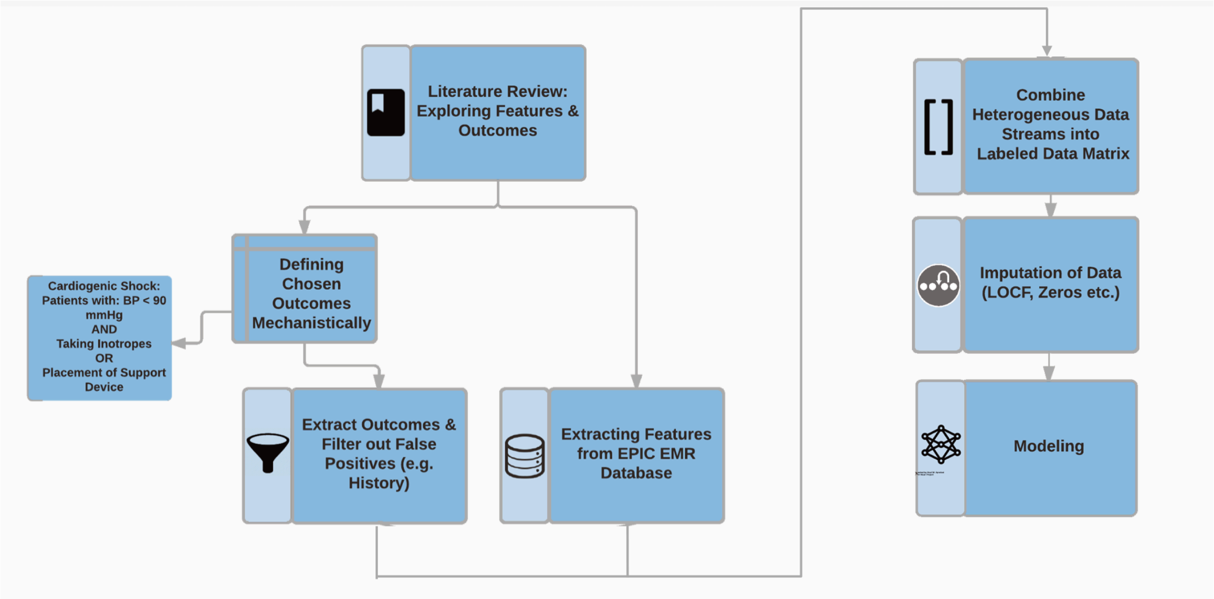
Literature Review: Exploring Features & Outcomes
Defining Chosen Outcomes Mechanistically
Cardiogenic Shock: Patients with: BP < 90 mmHg AND Taking Inotropes OR Placement of Support Device
Extract Outcomes & Filter out False Positives (e.g. History)
Extracting Features from EPIC EMR Database
Combine Heterogeneous Data Streams into Labeled Data Matrix
Imputation of Data (LOCF, Zeros etc.)
Modeling

Fig. 1 System diagram for data analytic engine

Setting and patient population

The cohort used to develop and train the models were patients admitted to Johns Hopkins Hospital (JHH) in Baltimore, MD from January to July 2017. EHR data including patient information, history, and medical information was extracted from JHH's Clarity electronic health record system and stored into an encrypted database. The cohort consists of 243,325 separate hospital stays, or encounters, from 92,209 patients. For each encounter, all data from admission to either discharge or the onset of an outcome event is considered. Each encounter was treated as a separate data series in the current stage of the modeling process. A possible improvement is to incorporate data from a patient's previous visits as a feature.

Outcomes

The outcome variables specified by our cardiologists were acute heart failure, cardiogenic shock, and cardiac arrest. Based on the data available through the EHR system, these three outcomes are defined as follows:

- Acute heart failure: Patients diagnosed with relevant ICD-9 Clinical Modification Codes (Appendix B) **and** use of furosemide intravenously 2 or more times within 48 hours. The onset of the event is assumed to be the timestamp of the first furosemide IV.
- Cardiogenic shock: Patients with systolic blood pressure <90 mmHg **and** either had an initiation of inotropes (dopamine, dobutamine, milrinone, norepinephrine) **or** had a placement of mechanical support devices (Appendix B). The onset of event is assumed to be the later timestamp of the two necessary conditions.
- Cardiac arrest: Evidence of cardiac arrest in physician's procedure notes. The onset of event, if not recorded in the notes, are assumed to be the timestamp of the note that contains the evidence.

In our patient cohort, 293 out of 243,325 encounters satisfied the definitions of outcome specified above (.1%). These encounters are described as positive encounters, and patients who did not have any of the three specified conditions in their stay are referred to as negative cases.

Feature Extraction

The design and definitions of potential predictive features were drawn from literature and refined with expert clinical opinion. As a result, 80 features were extracted from the EHR database (Appendix B), including:

- Patient demographic information: Age, gender, weight, ...
- Patient history: history of arrhythmia, immunodeficiency, ...
- Vital signs: heart rate, blood pressures, temperature, ...
- Medications: dosage of atenolol, isosorbide dinitrate, troponin, ...
- Orders of procedures: electroencephalogram, telemetry....
- Lab test values: hematocrit, PaO2, PaCO2, ...

The extracted features are preprocessed such that they are all numerical values — the medications are parsed into the dosage of prescription, and all categorical variables were created into distinct binary yes/no variables for each factor. As all our features except for demographic information and patient history has a timestamp available, a full outer join on timestamps was applied to all features found for each encounter. After this step, we have a varying-length time series with 80 features at each time-step for each encounter. As a lot of extracted features are sparse and different tests are conducted at different times, the full outer join results in significant amount of empty fields in the table.

Imputation

Other than as the result of a full outer join, data might be missing for a variety of reasons, from tests only ordered when a certain pathology was observed (missing not at random), to laboratory results that were normal did not set the flag variables (missing at random), to technical errors occurring when filling out forms and transmitting that data to the backend databases (missing completely at random).

To address the missing values in the design matrix, three imputation strategies were used. Firstly, binary indicator variables (such as procedures or history) and medicine doses were imputed with zeros if not present for a given time-step, indicating that the field is either missing or not prescribed. Secondly, a last-one-carry-forward strategy was implemented for vital signs and lab measurements that cannot be imputed with zeros (since, for example, a 0 heart rate would indicate a severe condition). The missing values are filled in with the most recent available value until a new measurement is encountered. Thirdly, if there are no available previous measurements (in the case of the first few time-steps of an encounter), missing values are imputed with the global mean of the population.

The application of these three strategies are arbitrary. We hope to revise the strategies with rigor in future renditions of the pipeline. Some possible improvements include using regression model to predict the missing values based on the previous measurements, and adding a binary flag to each field to indicate if measurement was reported or imputed.

Training, testing, and validation data

The nature of Statistical Classifiers requires a balanced training set with similar numbers of positive and negative encounters. to construct such balanced dataset, we randomly sampled 293 encounters from the set of negative encounters. The dataset was split 70/30 into testing and training sets. Both the testing and training sets are balanced. To keep our evaluation criterion consistent across the model families, the same set of encounters from the Statistical Classifiers were used for training and testing for the other two families. However, due to the different constructions of our baseline models, data were prepared separately for each model family.

Prediction Model	Data Preparation
Statistical Classifiers logistic regression, support vector machine, and XGboost	The design matrix transformed into a pure numerical matrix for SciKitLearn compatibility.
Survival Analysis Models Cox's Proportional Hazard Aalen's Additive Model	A new column "time to event" was calculated and added. All features were standardized, centered to the mean and component wise scaled to unit variance.
Recurrent Neural networks LSTM networks	For each encounter, the rows were binned into hourly windows. The binned fields consist of either the latest value or the sum of values over the past hour, based on their imputation strategy. These hourly time series are again imputed accordingly. For positive cases, data for 48 hours before onset was taken. For negative cases, the first 48 hours since their admission was taken.

Model development

Due to the time constraint of the program, we implemented the baseline models in their canonical form using existing python packages.

Prediction Model	Implementation Package
Statistical Classifiers	logistic regression, support vector machine: *ScikitLearn* XGBoost: *XGBoost*
Survival Analysis Models	Cox's Proportional Hazard and Aalen's Additive Model: *lifelines*
Recurrent Neural networks	LSTM: *Keras* and *TensorFlow*

For statistical classifiers and survival analysis models, each time-step (row in the design matrix) was treated as a separate data point for prediction (i.e. no continuity of time was taken into consideration). For recurrent neural networks, however, time series of 24 hours in length are fed into the network structurally.

Evaluation

The models were tested on the same balanced set of randomly selected encounters. Our primary metric for prediction is an area under the receiver operator characteristic curve (AUROC), one of the most commonly reported metric in deterioration prediction research.

Analysis of Data

a) Statistical Classifiers

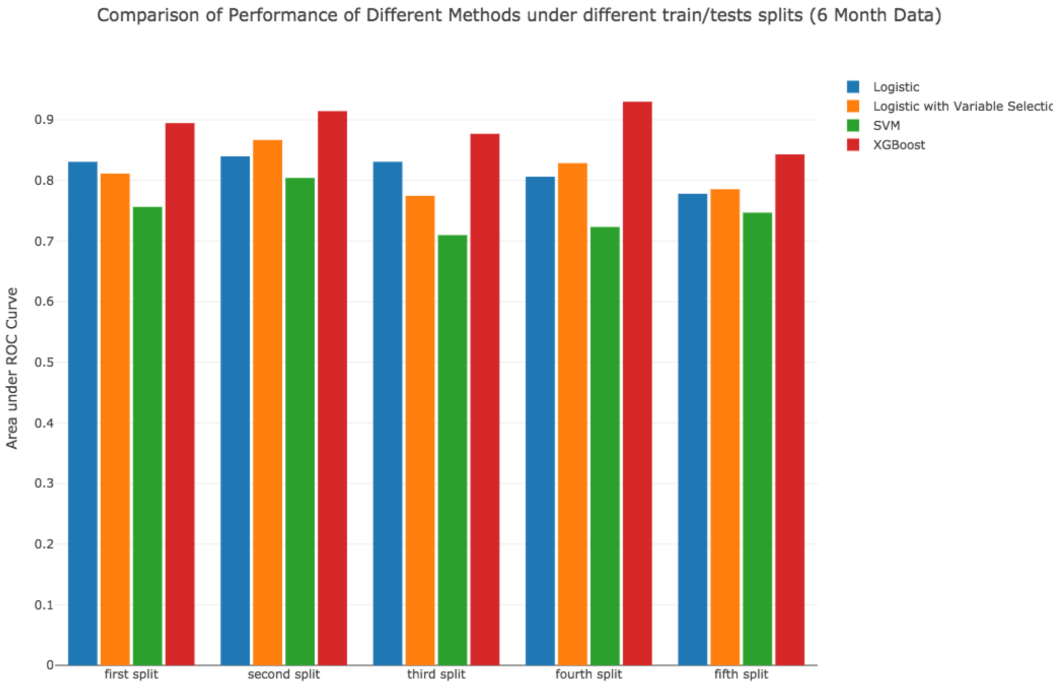

Fig 2. Performance of Logistic regressions, Support Vector Machine, and XGBoost

Fig. 2 shows the performance of statistical classifiers on five random 70/30 cross validation splits. XGBoost performs consistently best among them, registering AUROC scores in the range of 0.85-0.91, achieving similar precisions as, if not outperforming, the state-of-the-art NEWS score with 0.87 AUROC [8].

It is also interesting to note that feature selection did not contributed much to the performance of logistic regression models (blue and orange bars in fig. 2), suggesting that most of the extracted features contributes to the prediction. Further research need to be conducted to examine the features that are primary signals of risk for patients in our models.

b) Survival Analysis

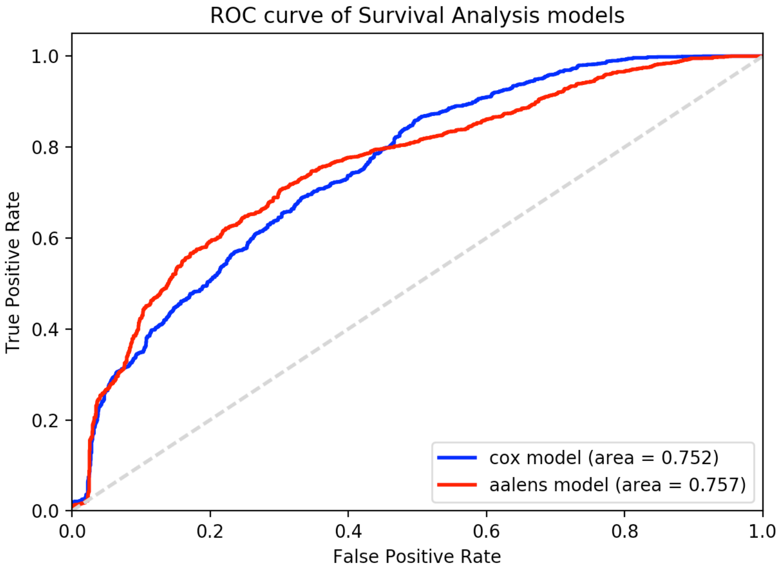

Fig 3. ROC curves of survival analysis models

Survival analysis models fit hazard functions that estimates the instantaneous probability of a hazard (in our case cardiac deterioration given a patient's conditions). The two models examined, the Cox Proportional Hazard model and Aalen's Additive model differs in the way they weight an individual's feature values at time *t* by learned regression coefficients *b* to the baseline hazard function (see equation 1 and 2 below). The advantage of survival analysis models is that, in an early warning setting, they are able to predict not only a binary outcome, but also time to event. Fig. 3 shows that while yielding similar performance (AUROC = 0.75), Aalen's model has higher sensitivity while Cox more specific.

$$\lambda(t) = b_0(t)\exp(b_1x_1 + ... + b_Nx_n)$$

Eq 1. Hazard function for Cox PH

$$\lambda(t) = b_0(t) + b_1(t)x_1 + ... + b_N(t)x_T$$

Eq 2. Hazard function for Aalen's Additive

c) Recurrent Neural Networks

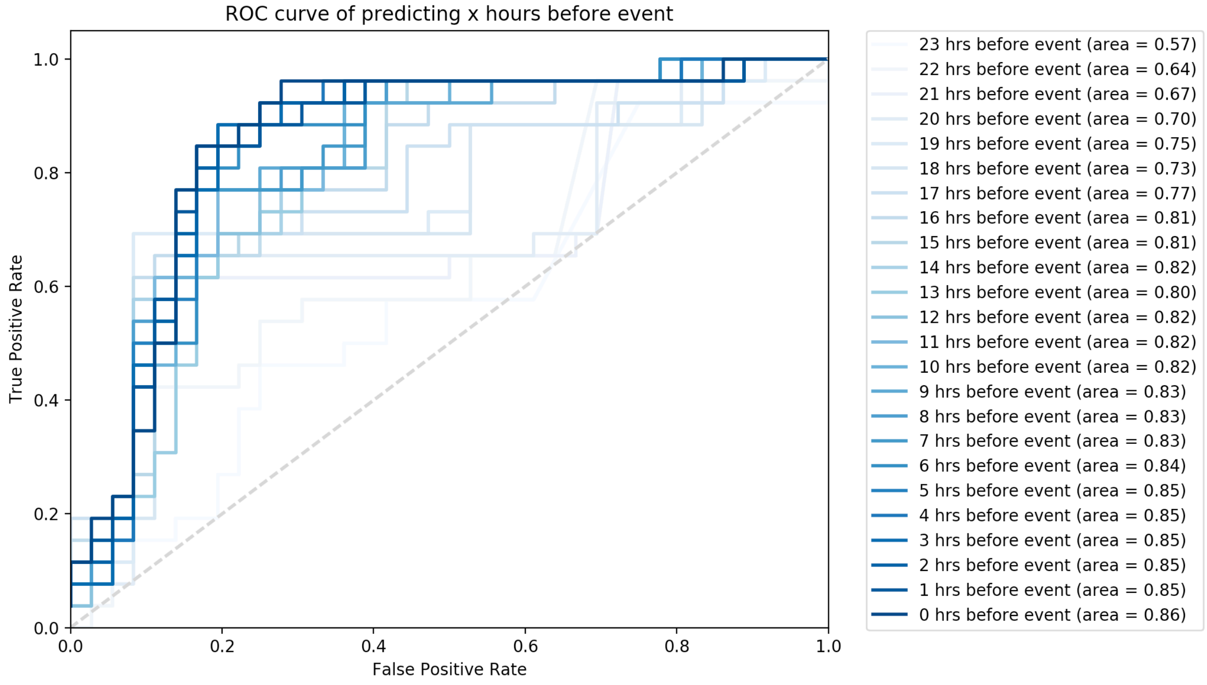

Fig 4. ROC curves of RNN predictions x hours before event

The most promising result came from a sequence-to-one LSTM network, which predicts at every hour from a sliding window of 24 hours. The 23-hours-before-event prediction is based on data 48 to 24 hours before event, the 22-hours-before prediction is based on data 27 to 23 hours before event, and so on. The marginal increase in sensitivity and specificity decreases as we approach the 12th hour. This experiment shows that outcomes can be relatively reliably predicted (AUROC > 0.82) up to 12 hours in advance.

Conclusions

The performance of these baseline models suggest that with the selected group of features, cardiac deteriorations can be feasibly modeled using machine learning algorithms, especially XGboost and LSTMs. Early warning and prediction from real-time data are achievable. This work serve as a proof-of-concept for further development of the cardiac deterioration prediction system.

References

1. CDC, NCHS. Underlying Cause of Death 1999-2013 on the CDC WONDER Online Database.
2. Puymirat, Etienne, et al. "Cardiogenic shock in intensive care units: evolution of prevalence, patient profile, management and outcomes, 1997–2012." *European journal of heart failure* 19.2 (2017): 192-200.
3. T. J. T. Hodgetts, G. G. Kenward, I. G. I. Vlachoniko- lis, S. S. Payne, and N. N. Castle. The identification of risk factors for cardiac arrest and formulation of activation criteria to alert a medical emergency team. *Resuscitation*, 54(2):125–131, Aug. 2002.
4. C. Sandroni, G. Ferro, S. Santangelo, F. Tortora, L. Mistura, F. Cavallaro, A. Caricato, and M. Antonelli. In-hospital cardiac arrest: survival depends mainly on the effectiveness of the emergency response. *Resuscitation*, 62(3):291–297, Sept. 2004.
5. A. C. Andréasson, J. Herlitz, A. Bång, L. Ekström, J. Lindqvist, G. Lundström, and S. Holmberg. Characteristics and outcome among patients with a suspected in-hospital cardiac arrest. *Resuscitation*, 39 (1-2):23–31, Oct. 1998.
6. M. M. Churpek, T. C. Yuen, M. T. Huber, S. Y. Park, J. B. Hall, and D. P. Edelson. Predicting Cardiac Arrest on the Wards: A Nested Case-Control Study. *Chest*, 141(5):1170–1176, May 2012.
7. J. McBride, D. Knight, J. Piper, and G. B. Smith. Long-term effect of introducing an early warning score on respiratory rate charting on general wards. *Resuscitation*, 65(1):41–44, Apr. 2005
8. Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., & Featherstone, P. I. (2013). The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of

early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, *84*(4), 465–470. https://doi.org/10.1016/j.resuscitation.2012.12.016

9. K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (trewscore) for septic shock," *Science Translational Medicine*, vol. 7, no. 299, pp. 299ra122–299ra122, 2015
10. Mortazavi, B., Desai, N., Zhang, J., Coppi, A., Warner, F., Krumholz, H., & Negahban, S. (2017). Prediction of Adverse Events in Patients Undergoing Major Cardiovascular Procedures. *IEEE Journal of Biomedical and Health Informatics*, *PP*(99), 1.
11. Alvarez, C. A., Clark, C. A., Zhang, S., Halm, E. A., Shannon, J. J., Girod, C. E., … Amarasingham, R. (2013). Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. *BMC Medical Informatics & Decision Making*, *13*, 28.
12. Churpek, M. M., Adhikari, R., & Edelson, D. P. (2016). The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation*, *102*, 1–5. https://doi.org/10.1016/j.resuscitation.2016.02.005
13. Cao, H., Eshelman, L., Chbat, N., Nielsen, L., Gross, B., & Saeed, M. (2008). Predicting ICU hemodynamic instability using continuous multiparameter trends. *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2008*, 3803–6. https://doi.org/10.1109/IEMBS.2008.4650037
14. Rochefort, C. M., Verma, A. D., Eguale, T., Lee, T. C., & Buckeridge, D. L. (2014). A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data. *Journal of the American Medical Informatics Association*, 155–165.
15. Zachary C Lipton, David C Kale, Charles Elkan, and RandallWetzel. (2016). Learning to Diagnose with LSTM Recurrent Neural Networks. *International Conference on Learning Representations, 2016*
16. Hrayr Harutyunyan, Hrant Khachatrian and David C. Kale, Aram Galstyan. (2017). Multitask Learning and Benchmarking with Clinical Time Series Data. In Proceedings of *ACM Conference,Washington, DC, USA, July 2017 (Conference'17)*, 11 pages.
17. Rajpurkar P., Hannun A. Y., Haghpanahi M., Bourn C., Ng A.Y. (2017). Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. abs/1707.01836, 2017. URL: https://arxiv.org/abs/1707.01836

Appendix A: Program Questions

1. *Research Ethics - What was done in the lab to support this topic?*

   The research protocol was approved by The Johns Hopkins University Institutional Review Board (IRB) which concluded that the research presented no more than minimal risk of harm to subjects. Therefore, the IRB waived the need for informed consent. All patient data were de-identified and stored in an encrypted database. We were only allowed access to patient data after we completed all ethics training and IRB approval of our research statement. The data were not used in any context other than this particular research topic. As the design of this research project changed several times during the span of the program, all changes were submitted to the IRB and were approved.

2. *Value of the Program - How will you apply the knowledge and skills you have gained in your research experience to subsequent college enrollment (undergrad, grad, PhD) and/or future employment?*

   This program was an extremely valuable experience for me. Not only did I get to work on a machine learning project and engage in the whole process from design to data ETL to modeling, I also learned about doing computer science research in the medical context. I learned about statistical learning techniques, and read about current research in the area as a part of the machine learning reading group of the lab. These were things I am always interested in and am grateful to have the opportunity to explore along with graduate students. I also received guidance on code organization, modularization, and documentation from my mentors in the lab. I am now more capable of developing a machine learning prediction system from the REU experience.

   The program provided me with new perspectives about working as an engineer / computer scientist to develop real world applications. I communicated with physicians and hospital administrators extensively in the process of the 10 weeks, and learned little by little to see our project from their perspective and translate their inputs back into the realm of computer science. These communication skills will definitely be useful for future collaborations with not only physicians but also other non-CS professionals.

3. *Overview of the Program - If you were to recommend the program to a friend, what were the highlights, and what could be changed or strengthened?*

   A list of highlights:
   - The research. Working in Saria Lab has been my best research experience so far. Both our PI and graduate student mentor were insightful and very helpful during the program. We learned about research of other lab members and collaborated with many of them.
   - The community. It's my first time surround by so many mechanical and electrical engineers interested primarily in research and the experience was inspiring. It's inspiring to learn about their perspective and their projects.
   - The housing is superb. I appreciate the concierge, front desk, and all amenities and services available to us at Charles Commons.

   I think the program is well-organized and has been a valuable experience for all of us. I only hope the program was longer so we can work in the lab more :). One aspect I hope can be change in the future was the organization of the presentation classes. One-on-one appointments or small group meetings might be more efficient and effective in the later stage of the class.

Appendix B: ICD-9 codes relevant to definition of outcomes

ICD-9 codes relevant to acute heart failure: 402.01, 402.11, 402.91, 425.1, 425.4, 425.5, 425.7, 425.8, 425.9, 428.0, 428.1, 428.2, 428.21, 428.22, 428.23, 428.3, 428.31, 428.32, 428.33, 428.4, 428.41, 428.42, 428.43, 428.9

ICD-9 codes relevant to placement of mechanical support devices:

- implant of pulsation balloon (37.61)
- insertion of implantable heart assist system (37.66)
- insertion of percutaneous external heart assist device (37.68)
- insertion or implantation of two external VADs for simultaneous right and left heart support (37.60)
- Extracorporeal membrane oxygentation (39.65)

Appendix C: A complete list of the extracted features

alt liver enzymes, eeg procedure, weight, lorazepam dose, paco2, platelets, arterial ph, inr, height, buprenorphine-naloxone dose, phenobarbital dose, metoprolol dose, sodium bicarbonate IV dose, spo2, creatinine, metoprolol IV dose, cortisol, hematocrit, hydralazine dose, lisinopril dose, temperature, tsh, bilirubin, ast liver enzymes, bnp, lactulose dose, bicarbonate, enalapril dose, phytonadione dose, anion gap, labetalol, abp sys, albumin, systemic steroids dose, respiratory rate, glucose, protamine dose, rifaximin dose, co2, flumazenil dose, lactate, aspirin dose potassium, bmi, carvedilol dose, losartan dose, ekg procedure, naloxone injection dose atenolol dose, bun, mchc, isosorbide dinitrate dose, wbc, sodium bicarbonate injec dose, telemetry procedure, arterial blood gas procedure, pao2, abp dias, troponin, furosemide IV num admin, valsartan dose sodium, nitroglycerin IV dose, heart rate, lymph abs, urine output, phenytoin dose, diazepam dose, age, gender, immunodeficiency history, heart arrhythmias history, heart failure history, chronic airway obstruction diagnosis, existence of previous event