Heinz 95-845: Final Project Report

Nettie Lian JIAOYINL/JIAOYINL@ANDREW.CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

Stephen Lin DONGLIEL/DONGLIEL@ADDRESS.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

1. Introduction

Sepsis is defined as "life-threatening organ dysfunction caused by a dysregulated host response to infection"[1], and it is a leading cause of hospital mortality in the United States. While physicians administer fluid as a first-line strategy to combat sepsis onset, it is unclear if the patient will respond to the treatment positively. As aggressive fluid administration on unresponsive patients can lead to serious adverse events[2][3][4] , a reliable early prediction of volume responsiveness on sepsis patients must be developed.

We have utilized the Medical Information Mart for Intensive Care III (MIMIC-III) dataset[5] and The MIMIC-III Waveform Database Matched Subset[6] to build machine learning models. As an extension to *Lian et al. (2019)*, we successfully developed a pipeline to clean up both the EMR and waveform data, performed feature extraction from high-resolution waveform, enhanced EMR with waveform features, and applied machine learning algorithms on the focused dataset.

The goal of the study is to provide data-driven insights that would better inform clinical practitioners in developing fluid strategies on sepsis 3 patients, backed by a high performance prediction model. The major contributions of the study are listed as follows:

- This is the first ever study that incorporates high resolution waveform data into the prediction landscape in identifying volume responsiveness on patients in intensive care units.
- Our proposed model, UAENA, provides clinical interpretability by performing feature selection using l1 sparsity, identifying key indicators.
- Our proposed model, UAENA, proves that inclusion of waveform derived features is meaningful, as lesion studies indicates a performance boost in terms of AUC.

Keyword: Sepsis-3, Volume-Responsiveness, MIMIC-III, MIMIC-III Waveform, Feature Extraction, Logistic Regression, Regularization, Interpretability, Lesion Studies, AUC, Machine Learning

2. Related Work

As far as our literature exploration goes, there lacks a systemic view of volume responsiveness predictors despite full understanding of the importance to early predict responsiveness.

Monnet, Marik, and Teboul (2016) has reviewed several methods for prediction, and compared them with limitations and applicable use cases. However, the study did not identify clinical features and characteristics for early prediction. *Girkar et al. (2018)* has built attention-based recurrent neural network to predict successful response to fluid treatment. The best performing model of *Girkar el al.* is a stacked LSTM with attention mechanism, achieving accuracy of 0.852 and AUC of 0.925. Despite the promising results, *Girkar et al.* has relied on EMR data, which can be error-prone and retrospectively filled, as well as time-consuming laboratory examination results. Moreover, the fluid bolus volume threshold of 248 ml/hr does not comply with conventional 500ml/hr fluid challenge.

3. Background

In this paper we present a Logistic Regression using L1 sparsity feature selector, or UAENA (L1 featUre spArsity rEgresióN logísticA), to predict patient volume responsiveness. The following interpretation and explanation of L1 sparsity feature selector is partially borrowed from lecture notes of Prof. Jeremy Weiss, a distinguished assistant professor of health informatics at Carnegie Mellon University, Heinz College.

Considering learning task L as a process that finds the parameters that minimize some objective function J(θ). Regularization is a common technique to shrink parameter space by applying an additional penalty term on top of the original target function. The new objective function under L1 regularization is defined as:

$$JL_1(\theta) = J(\theta) + ||\theta||_1$$

The L1 regularization does "automatic feature selection" where the addition of penalty term yields sparse parameter vectors. The UAENA model construction relies on this quality to make informed decision of feature selection.

4. Data

4.1 MIMIC-III

We have collected data from the MIMIC-III database, which contains 61532 ICU stays, or 58976 unique ICU patients admitted to Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. The relational database contains 26 tables and more than 400 M rows, documenting patients demographics, vitals signs, and laboratory examination results. *Lian et al. (2019)* has developed a pipeline enabling MIMIC-III clean-up, query, and pivoting values suited for our study, and we have utilized the MIMIC-III sub-schema they created for all further analysis in this study.

4.2 MIMIC-III Waveform Matched Subset

We have collected data from the MIMIC-III Waveform Matched Subset, an intersection of MIMIC-III Waveform Database and MIMIC-III clinical database. The waveform data contains recordings collected from patient bedside monitors in ICUs. The Matched Subset contains 5960 unique patients that can be matched with the MIMIC-III database. *Lian et al* has developed a feature extraction pipeline that transforms high resolution time series data

into characterized statistics such as autocorrelation and binned entropy. We have utilized the *Lian et al.* pipeline to extract waveform features for all further analysis in this study.

5. Data Preprocessing

5.1 *Lian et al.* Waveform Extraction Pipeline

We have utilized the waveform extraction pipeline *Lian et al.* introduced. To integrate the extracted results with the MIMIC-III EMR data, we merged the results on per fluid event level. Since patients may receive multiple fluid treatments in a given ICU stay, and patients may respond to crystalloid differently based on their unique situation at the time fluid bolus is administered, we treated fluid events that are 6 hours away from the previous fluid event as separate treatment episode, thus different observations in our model.

5.2 Imputation Strategy: (For waveform extraction results only)

5.2.1 MCAR

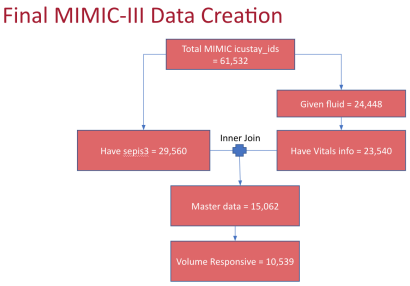
In our dataset, 94% of the waveform records contain EKG lead II information. There are 313 out of 524 records have EKG lead V waveform information and 351 out of 524 records have PLETH waveform information. We determined that the propensity for these records to be missing is completely unsystemic. The reason of missingness in lead V and lead PLETH is probably related to the fact that lead II signals have already been collected in our dataset. These missingness can be due to nurses' discrete, but random decision that monitoring lead II is standard practice according to the manual, but other leads are not. We've concluded that the missing data in our dataset is MCAR but not other mechanisms. Based on the early data exploration we've performed on our original dataset, there is no evidence that indicates patients with certain characteristics are more likely to not include lead V or PLETH signal information than others.

5.2.2 Multiple Imputation by Chained Equations (MICE)

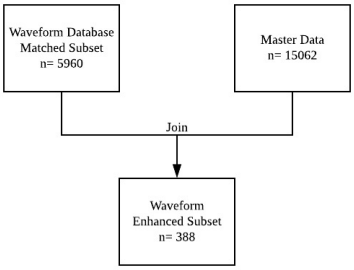
We've performed MICE for the purpose of our project. Multiple imputation is a process where we fill the missing values multiple times in order to create multiple "complete" datasets. It has a lot of advantages over single imputation methods, for instance, replacing empty values by mean or median. The MICE algorithm works by running multiple regression models and each missing value is modeled conditionally depending on the other variables in the dataset.[7] Here we used IterativeImputer function from sklearn library.

5.3 Integration with MIMIC-III

From *Lian et al.* sub-schema of MIMIC-III, we have collected data from patients who developed sepsis-3, received fluid treatment at infusion rate greater than 500 ml/hr, and had vitals recorded. In this focused dataset, there are 15062 unique fluid events and in 10539 occasions patients are responsive to fluid volume. After matching waveform extraction results with MIMIC-III database, we obtained 388 unique fluid event observations with vital signs and waveform records present. The integrated dataset has 274 features, including 219 waveform extracted characteristics, 48 vital signs, and 7 patient demographics.



(a) Final MIMIC-III Data Creation



(b) Joining waveform data with MIMIC-III

6. Method

With the dataset at hand, we have built several machine learning models with the aim to predict volume responsiveness on sepsis-3 patients and identify key indicators. Due to the concern of sample size and interpretability, we have excluded recurrent neural networks and other deep learning models. In our study, we have explored RandomForest, XGBoost,, unregularized Logistic Regression without feature selection, Logistic Regression with feature selection using L1 sparsity, and Support Vector Machines. Due to the model choices, we have normalized all features using standard scaler available in scikit learn. For all prediction model families aforementioned, we have implemented 70-30 train-held out split and conducted 10-fold cross validation on an extensive hyperparameter space. The best performing model of our study is Logistic Regression using L1 sparsity feature selector, UAENA, reaching accuracy of 88.89% and AUC of 0.863.

6.1 UAENA

Building our best performing model, UAENA, requires a two-step process, first a feature selection using L1 sparsity, then construct logistic regression model accordingly. The logistic regression constructed in the second phase uses the liblinear solver that uses l2 norm penalization. The inverse strength of regularization, controlled by hyperparameter C, is decided by 10-fold cross validation using the vector space of [0.01, 0.1, 1, 10].

6.2 Evaluation criteria

In this study, we use a combination of accuracy, AUC, and F1 score as evaluation criteria. Those metrics were chosen because they are suited for evaluating classification tasks, and F1 score tells a compelling story by weighing both precision and recall. As a tiebreaker, performance robustness and clinical interpretability should be taken into account since comprehension, which leads to commitment, of target users is imperative.

7. Results

7.1 Baseline characteristics

We use an unconstrained, non feature-selected logistic regression as baseline model. The baseline model has an accuracy of 81.2% and AUC of 0.812 in the held-out test set of

Model	Accuracy	AUC	F1
RandomForest	82.05%	0.912	0.901
XGBoost	86.32%	0.850	0.925
Logistic Regression	81.20%	0.812	0.820
UAENA	88.89%	0.863	0.890
LinearSVC	81.20%	0.570	0.800
SVC	80.34%	0.566	0.790

Figure 2: Performance Comparison

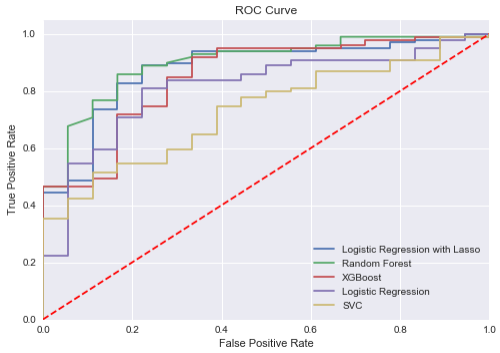
117 in size, and 84.61% unresponsiveness prevalence. Out of the top 30 features, 15 are extracted from waveform, 12 are relevant to vital signs and 3 demographic are included. The baseline model on its own indicates the importance of waveform features in predicting volume responsiveness even on a relatively small sample size. However, the unconstrained construction includes all 274 features and is likely to overfit the train set.

7.2 Comparison between UAENA and other Models

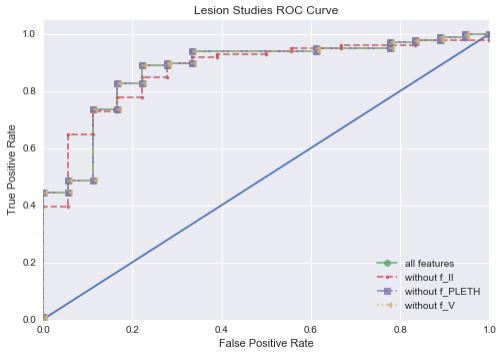
Compared to other prediction models, the support vector machine family clearly performs suboptimally as it presents the lowest accuracy, AUC, as well as F1 score. After eliminating the support vector machine models, the baseline unconstrained Logistic Regression is also eliminated from the competition of best prediction model for similar reasons. We are left with RandomForest, XGBoost, and UAENA, the constrained logistic regression. From the given metrics, accuracy, AUC, and F1, there is no strictly dominant model that has a clear advantage in all three dimensions against any of the two other models. For example, accuracy of RandomForest is less than that of XGBoost, its AUC outperforms. UAENA has a clear lead in accuracy against both contestants, but its F1-score is the lowest among the three, although there is only marginal difference. As a matter of fact, we performed a simple robustness check by rerunning those three models with different random states and training testing splits. RandomForest and XGBoost pass the robustness check as we expected for those ensembles. To our surprise, UAENA also presents steady performance in the test, showing strong performance metrics across all randomness we tested. (The result files can be accessed in the repository as well) . The determining factor comes down to interpretability, which UAENA is the clear winner as it provides clean, understandable interpretations for the clinicians. UAENA weakly dominates both ensemble models for being a simpler, more interpretable model that is equally robust in terms of all performance metrics. By applying Occum's Razor, we have concluded that UAENA is the best performer in our study.

7.3 UAENA features

UAENA contains 25 features after using L1 sparsity feature selector. Out of the 25 features, 6 are vital related, and 19 of which are extracted from waveform.



(a) Model Comparison: ROC Curves



(b) ROC curve, UAENA Lesion Studies

7.4 Lesion studies

For this particular lesion study, we'd like to gain insight into what contributes to our model performance by lesioning components of it, aka, evaluating models with lead II records, lead PLETH records, or lead V records.

The ROC curves above show how performance is affected when different feature types are removed from UAENA. It's clear to see without lead II information, UAENA performance is not as good as before. As we can see from Appendix 3, among all three categories of waveform lead records, only lead II records have been included in our UAENA. Hence, when we removed lead PLETH or lead V records, UAENA performance is not affected because the model does not consider lead PLETH or lead V information if lead II features are present when making predictions.

7.5 Learning Curve

From the learning curve of UAENA, we observe a reasonable improvement as training sample size increases. UAENA construction is able to quickly improve performance from the additional variation it observes from a marginally increased sample size. With only or the original training set size, UAENA is able to generate reasonably good scores; and with half of the training set size, it achieves similar scores compared to consuming the entire training data.

8. Conclusion

In this study, we have integrated high resolution waveform data from MIMIC-III Waveform Matched Subset with EMR data from MIMIC-III to predict volume responsiveness on sepsis 3 patients. Prior studies on similar prediction tasks have focused on the EMR data as well as laboratory examination results from MIMIC-III database, which can be error-prone and retrospectively filled. Our work successfully extends works of Girkar et al. and Lian et al. to improve volume responsiveness prediction with waveform feature extraction. With UAENA, a constrained logistic regression that uses L1 sparsity feature selector, we demonstrated that it is achievable to develop a robust, interpretable model with strong performances

.5

Figure 4: UAENA Learning Curve

in accuracy and AUC. UAENA has an accuracy of 88.89%, and AUC of 0.863. In terms of clinical interpretability, UAENA also provides clear indicator that waveform extracted features are beneficial in improving the prediction model, as there are 19 waveform features among the 25 features selected. Our work should be able to demonstrate the benefits of keeping waveform records in assisting clinicians in determination of fluid strategies for sepsis patients in the MIMIC-III dataset. Systematic review of external validity of our work should be implemented to better understand volume responsiveness, and better inform the clinical practitioners. Reproducible scripts can be found at *https://github.com/netlian/AAMLP-Project*

References

[1] Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801–810. *doi:https://doi.org/10.1001/jama.2016.0287*

[2] Monnet X, Marik PE, Teboul JL. Prediction of fluid responsiveness: an update. Ann Intensive Care. 2016;6(1):111. *doi:10.1186/s13613-016-0216-7* .

[3] Micek ST, McEvoy C, McKenzie M, Hampton N, Doherty JA, Kollef MH. Fluid balance and cardiac function in septic shock as predictors of hospital mortality. Crit Care. 2013;17(5):R246. Published 2013 Oct 20. *doi:10.1186/cc13072* .

[4] Boyd JH, Forbes J, Nakada TA, Walley KR, Russell JA. Fluid resuscitation in septic shock: a positive fluid balance and elevated central venous pressure are associ-

ated with increased mortality. Crit Care Med. 2011; *39(2):259–65. Epub 2010/10/27. 10.1097/CCM.0b013e3181feeb15* .

[5] MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). *DOI: 10.1038/sdata.2016.35. Available at: http://www.nature.com/articles/sdata201635* .

[6] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; *http://circ.ahajournals.org/cgi/content/full/101/23/e215]; 2000 (June 13).*

[7] Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. International journal of methods in psychiatric research, 20(1), 40–49. *doi:10.1002/mpr.329*

[8] Uma M. Girkar and Ryo Uchimido and Li-wei H. Lehman and Peter Szolovits and Leo Celi and Wei-Hung Weng (2018) Predicting Blood Pressure Response to Fluid Bolus Therapy Using Attention-Based Neural Networks for Clinical Interpretability *https://arxiv.org/pdf/1812.00699.pdf*

[9] Jiaoying (Nettie) Lian, Dong-Lien (Stephen) Lin, Himagar Molakapuri, Sri Mannikanth Nunna, Parth Shah (2019) Predicting Volume Responsiveness for sepsis-prone patients . Carnegie Mellon University Heinz College of Information System and Public Policy Management Capstone Project *Publication in progress*

Appendix 1: Top 30 features in baseline model

f_V__first_location_of_minimum	0.322542051
f_PLETH__longest_strike_above_mean	-0.262178203
f_V__longest_strike_below_mean	-0.201085014
f_V__binned_entropy__max_bins_10	0.181604874
f_RACE_BLACK	0.175414042
f_V__ar_coefficient__k_10__coeff_3	0.16760352
f_V__large_standard_deviation__r_0.15000000000000	0.16110354
f_SYSBP.MIN	-0.159399788
f_IS_MALE	-0.155161536
f_MEANBP.MIN	-0.143933561
f_PLETH__linear_trend__attr_rvalue	-0.137704139
f_V__linear_trend__attr_pvalue	-0.133066271
f_V__agg_autocorrelation__f_agg_mean__maxlag_40	0.132931805
f_V__ar_coefficient__k_10__coeff_2	-0.12840549
f_MEANBP.COUNT	0.126497724
f_HEART_RATE.COUNT	0.126497724
f_DIASBP.COUNT	0.126497724
f_RESPRATE.COUNT	0.126497724
f_SYSBP.COUNT	0.126497724
f_SPO2.COUNT	0.126497724
f_PLETH__fft_aggregated__aggtype_variance	0.125759842
f_SPO2.MAX	0.122011198
f_DIASBP.MIN	-0.120263661
f_V__large_standard_deviation__r_0.2	0.118980839
f_RACE_HISPANIC	0.118280599
f_MEANBP.25.	-0.117450208
f_V__ar_coefficient__k_10__coeff_4	-0.114836364
f_PLETH__large_standard_deviation__r_0.05	-0.114588163
f_SYSBP.25.	-0.11096553
f_V__autocorrelation__lag_1	0.109245464

Appendix 2: Model Performance Comparison

Model	Accuracy	AUC	F1
RandomForest	82.05%	0.912	0.901
XGBoost	86.32%	0.850	0.925
Logistic Regression	81.20%	0.812	0.820
UAENA	88.89%	0.863	0.890
LinearSVC	81.20%	0.570	0.800
SVC	80.34%	0.566	0.790

Appendix 3: UAENA Feature Importance

UAENA Features	
f_DIASBP.COUNT	1.31713
f_SYSBP.COUNT	1.31713
f_HEART_RATE.COUNT	1.31713
f_MEANBP.COUNT	1.31713
f_MEANBP.MIN	-0.54418
f_SYSBP.MIN	-0.52651
f_II__ar_coefficient__k_10__coeff_1	0.00084
f_II__autocorrelation__lag_5	-0.00084
f_II__ar_coefficient__k_10__coeff_4	-0.00084
f_II__autocorrelation__lag_3	0.00084
f_II__ar_coefficient__k_10__coeff_0	0.00084
f_II__autocorrelation__lag_6	-0.00084
f_II__autocorrelation__lag_4	-0.00084
f_II__autocorrelation__lag_2	-0.00084
f_II__agg_autocorrelation__f_agg_var__maxlag_	-0.00084
f_II__autocorrelation__lag_7	-0.00084
f_II__agg_autocorrelation__f_agg_median__max	-0.00084
f_II__agg_autocorrelation__f_agg_mean__maxla	-0.00084
f_II__autocorrelation__lag_8	-0.00084
f_II__ar_coefficient__k_10__coeff_3	-0.00083
f_II__autocorrelation__lag_1	-0.00083
f_II__ar_coefficient__k_10__coeff_2	0.00083
f_II__abs_energy	-0.00074
f_II__absolute_sum_of_changes	0.00074
f_II__autocorrelation__lag_0	0.00000