Prediction of the Impending Onset of Septic Shock in Patients with Sepsis

Marion Pang and Emily Chang

***Abstract* — Septic shock is a major public health concern associated with increased patient mortality and significant costs to the healthcare system. Timely identification and treatment of patients progressing from sepsis to septic shock is critical, as early interventions can significantly improve patient outcomes. We present the use of a generalized linear model with lasso regularization to analyze 28 features commonly recorded in electronic health records for 2,147 patients, identifying key factors that may lead to septic shock, as defined using Sepsis-3 criteria. Our model achieved a 0.94 maximum area under the receiver operating characteristic curve, 86.4% sensitivity, and 87.2% specificity, resulting in a median early warning time of 7.47 hours. For septic patients at risk for progression to septic shock, implementation of novel prediction algorithms can anticipate these developments and allow for more timely and effective treatment.**

I. INTRODUCTION

Sepsis, defined as a "life-threatening organ dysfunction caused by a dysregulated host response to infection" [1], affects more than 30 million people worldwide each year [2] and is the most costly condition in U.S. hospitals, responsible for over $20 billion of aggregate inpatient hospital costs in 2011 [3]. The consequent progression of untreated sepsis into septic shock is of even further concern: delays in treatment lead to an increase in mortality of about 8% per hour [4], and full onset of septic shock is associated with a mortality of 45% [5]. Initiation of interventions in a timely manner is therefore critical to improving clinical outcomes, ensuring immediate treatment and management upon identification of impending septic shock.

Utilization of machine learning and other computational techniques to analyze clinical data and predict the onset of severe sepsis or septic shock has been a central focus of many groups in recent years [6-9]. These studies have applied a variety of approaches, implementing signal processing techniques, Bayesian networks, and classification algorithms to predict instances of sepsis based on a variety of patient factors. Most recently, Liu *et al.* [6] demonstrated the efficacy of three different machine learning methods in predicting the transition from sepsis into a novel "pre-shock" state that puts them at heightened risk for eventual progression to septic shock. Analysis of this "pre-shock" state allows for improved performance compared to previous models.

Here, we demonstrate a replication of the generalized linear model (GLM) mentioned in Liu *et al.*, applied to the same dataset but with different preprocessing techniques. With a maximum area under the receiver operating characteristic curve (AUC) of 0.94, corresponding to 66.9% accuracy, 86.4% sensitivity, and 87.2% specificity, our model demonstrates a median early warning time (EWT) of 7.47 hours. The features most indicative of passage into the "pre-shock" state were identified as being decreased Glasgow Coma Score (GCS), systolic blood pressure, nervous system sequential organ failure assessment (SOFA) score, cardiovascular SOFA score, lactate levels, urine output, heart rate, respiratory SOFA score, respiration rate, liver SOFA score, partial pressure of carbon dioxide, partial pressure of oxygen, and fraction of inspired oxygen.

II. METHODS

A. Data Preprocessing

Data was extracted from both training and test datasets and processed to remove clear outliers and incomplete data. Notably, the following changes were made:

- All temperatures above 50 degrees were assumed to be in Fahrenheit and converted to Celsius.
- All extreme heart rates less than 20 beats per minute (BPM) or greater than 500 BPM were removed from the data set.
- All year, month, and day values were removed from timestamps, as the dates did not match up between the septic onset timings and measurement timestamp timings (e.g. Septic Onset in Dec 2196, whereas measurements were in April 2131).

B. Model Fitting

Next, a logistic regression model was trained with lasso (L1) regularization on the provided training data. Using the scikit-learn machine learning library, a regularized linear model with stochastic gradient descent (SGD) learning and a logistic regression loss function was trained on all the training data with stratified 10-fold cross-validation.

1) Lasso Regularization. Lasso (Least Absolute Shrinkage and Selection Operator) is a penalty applied to the linear model that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the model. In our dataset, we are presented with a list of 28 different features, which, without regularization, results in models with high variance. Introducing lasso regularization thus reduces variance at the cost of introducing some bias. Specifically, we seek to maximize the following objective function, where the regularization parameter, λ, represents the bias:

$$-\log\left\{\prod_{i=1}^{n} f\left(\sum_{j} x_{ij}\beta_j\right)^{y_i} + \left\{1 - f\left(\sum_{j} x_{ij}\beta_j\right)^{1-y_i}\right\}\right\} - \lambda \sum_{j=1}^{p} |\beta_j \quad (1)$$

where $f(\alpha) = \frac{e^{\alpha}}{1+e^{\alpha}}$.

Lasso regularization produces sparse models with few coefficients, making some coefficients zero and eliminating those from the model. This thus helps us refine our model and prevent overfitting by only considering the important features that affect the outcome. In particular, λ controls the strength of the L1 penalty such that larger λ eliminates more variables from the model.

2) Stochastic Optimization. The logistic regression model SGDClassifier was used in this project. SGDClassifier implements regularized linear models with SGD learning. That is, the gradient of loss is independent of the number of terms in the sum, instead optimizing the sum of a finite number of smooth convex functions.

3) Stratified K-fold Cross Validation. Cross validation is a method used to ensure that trained models are not overfitted or biased due to variance within the dataset or unknown dependencies between measurements. In particular, K-fold cross validation is performed by splitting the dataset K-folds, and using K-1 portions of the data to train the model, while testing the model on the remaining 1 portion of data. This is repeated K times, and the mean metrics of the model is obtained. Stratification seeks to ensure that each fold is representative of all strata of the data. Generally, this is done in a supervised way for classification and aims to ensure each class is (approximately) equally represented across each test fold (which are of course combined in a complementary way to form training folds). In our case, since prediction outcomes are either 0 (not in septic shock), or 1 (in septic shock), it is important to ensure equal representation of these 2 classes of outcomes and avoid heavy bias.

III. RESULTS AND DISCUSSION

A. With Data Balancing

10-fold cross validation was performed to ensure that the splitting of training data into training and testing sets for the model was not heavily biased. The performance of 10 folds of cross-validated models is shown here.

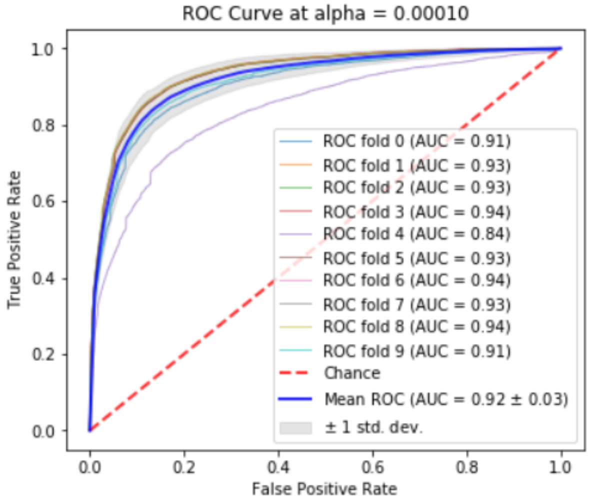


Fig. 1. ROC curves of 10 K-Fold Cross Validation at α = 0.0010

$$\alpha_{SGD} = \frac{\lambda}{n_{samples}} \tag{2}$$

Using 10-fold cross validation, where 90% of the data is used for training, we thus obtain:

$$\lambda = \alpha_{SGD} * 0.9N_{training} \tag{3}$$

As λ increases, bias increases and thus helps to reduce variance in the model, resulting in greater AUC and greater accuracy. However, at a large λ, the penalty effect dominates and essentially eliminates all variables from the model, returning only “0” outputs. Thus, the accuracy decreases to 0.835, the proportion of 0 (no septic shock) measurements in the training data set.

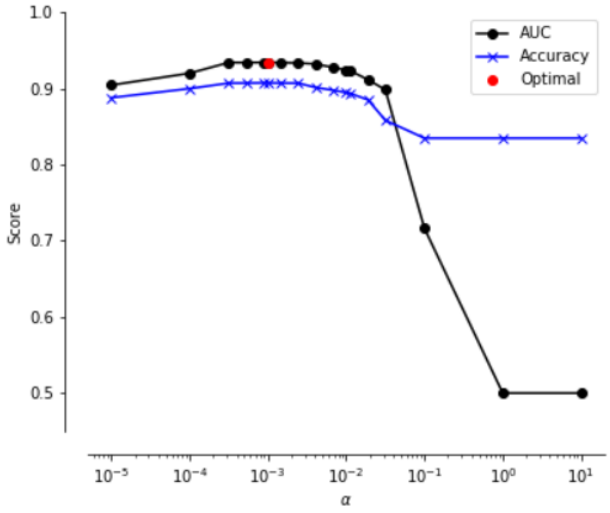
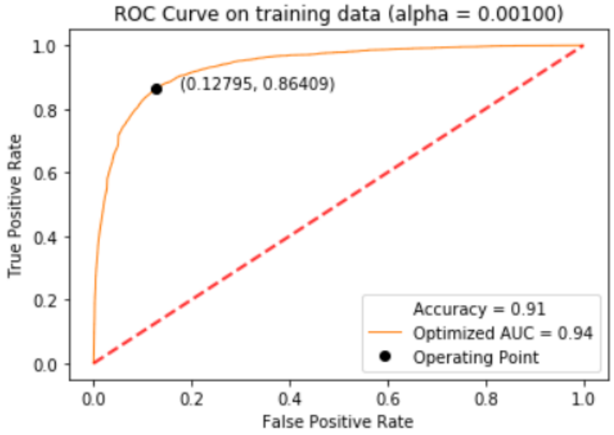


Fig. 2. AUC and Accuracy of mean 10-fold cross validation across range of α on balanced classes. Optimal α was found to be 0.00100.

An optimal value of $\alpha = 0.00100$ corresponding to the maximum AUC obtained was used to train the entire set of training data. This corresponds to $\lambda = 1418.46$. The ROC Curve was plotted on training data, and an AUC of 0.935 and accuracy of 0.908 was obtained. The operating point was defined to be the min of the norm from ideal (TPR, FPR) = (1,1) and was found to be (0.128, 0.864).



\Fig. 3. ROC curve on training data at optimal α = 0.00100. Metrics: AUC = 0.94, Accuracy = 0.91.

Next, the model was used to predict outcomes based on the testing data, and its performance was evaluated.

TABLE 1
GLM METRICS ON **TEST DATA** FOR BALANCED CLASSES

Metric	**Score**
AUC	0.731
Accuracy	0.669
Sensitivity	0.864
Specificity	0.872

Finally, the EWT between predicted and actual septic shock onset was calculated and is presented below.

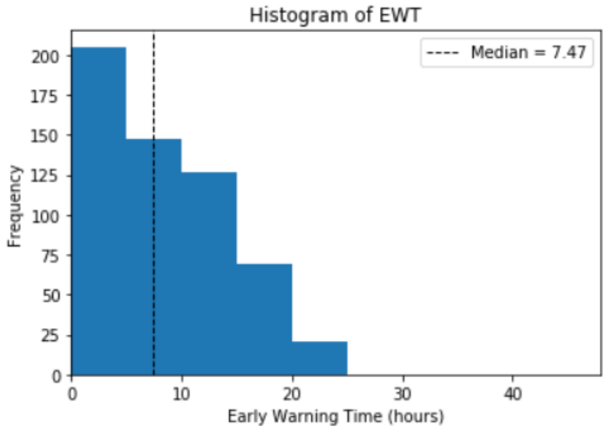


Fig. 4. EWT Histogram for balanced class model. (Median EWT = 7.47h)

B. Without Data Balancing

The training and estimation procedure was then repeated without using data balancing, by switching from a stratified k-fold to a normal k-fold with shuffle, and changing the class weights of the logistic regressor to "balanced". This results in a data set with greater variance, which is expected to decrease the performance of the model.

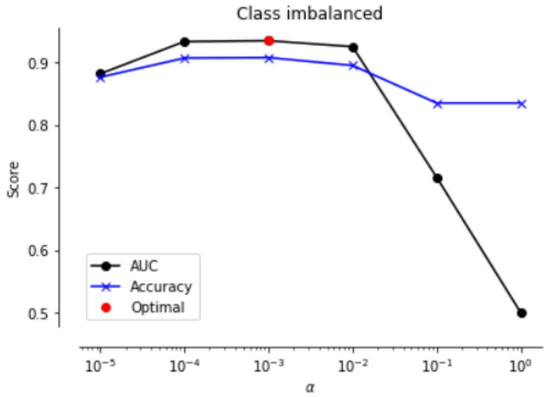


Fig. 5. AUC and Accuracy of mean 10-fold cross validation across range of α on imbalanced classes. Optimal α was found to be 0.00100.

$\alpha = 0.00100$ and $\lambda = 1418.46$ were used to train the entire set of training data, obtaining an AUC of 0.940 and accuracy of 0.875. The operating point was defined as before and was found to be (0.122, 0.871).

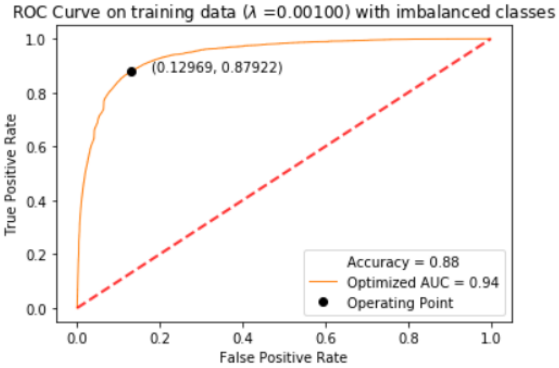


Fig. 6. ROC curve on training data at optimal α = 0.00100. Metrics: AUC = 0.94, Accuracy = 0.88

A ROC curve similar to that from the balanced data was obtained, with a comparable AUC but with lower accuracy (0.88 compared to 0.91 with the balanced classes). Interestingly, the same features were identified as before, but with stronger coefficients.

TABLE 2
GLM METRICS ON **TEST DATA** FOR IMBALANCED CLASSES

Metric	**Score**
AUC	0.680
Accuracy	0.669
Sensitivity	0.878
Specificity	0.871

Lastly, the median EWT obtained from the imbalanced class model was 7.25h, which is 0.22h later than the median EWT from the balanced class model. **As expected, the imbalanced class model does perform worse due to the higher variance in dataset.**

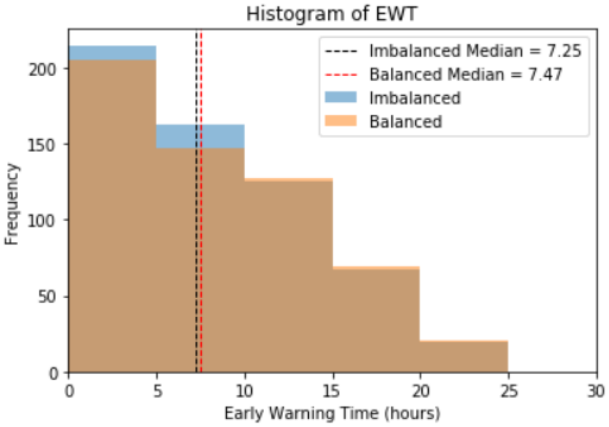


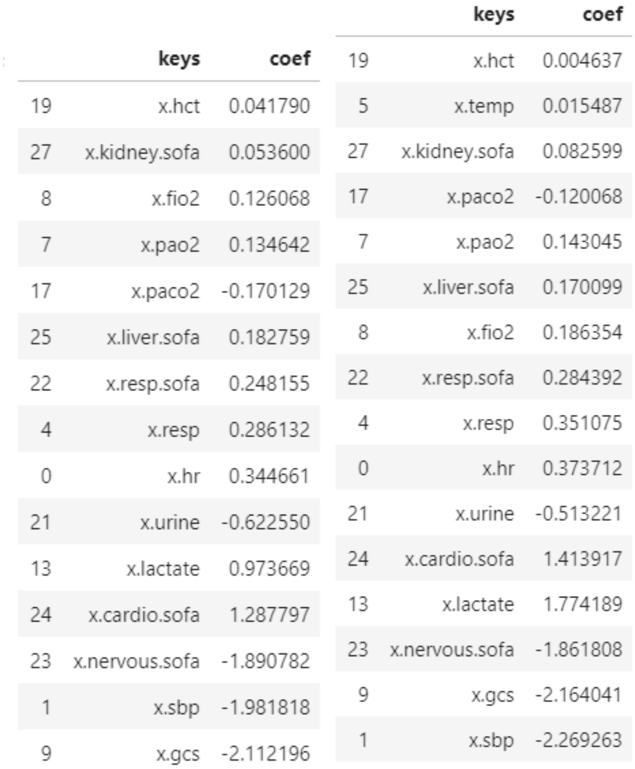
Fig. 7. EWT Histogram for balanced class model vs imbalanced class model. Balanced Median = 7.47h, Imbalanced Median = 7.25h.

IV. CONCLUSIONS

Logistic regression using a stochastic optimization strategy was successfully applied to a dataset to identify a "pre-shock" state and classify septic patients who may be at risk for impending septic shock. The model achieved a median EWT of 7.47 hours using data balancing, with an AUC of 0.741. This would provide ample information to clinicians well in advance of the development of septic shock, allowing for more timely treatment and intervention to prevent further complications.

However, though simple logistic regression does assist in the early prediction of septic shock, as shown in Liu, *et al.*, neural networks are more likely to produce a more robust model, as it allows for the consideration of dependent variables. For example, heart rate may influence cardiovascular SOFA, which a logistic regression model cannot account for, as it assumes independence between all features. Future implementations of these methodologies would benefit from developing the neural network model further, rather than solely relying on logistic regression.

APPENDIX



	keys	coef
19	x.hct	0.041790
27	x.kidney.sofa	0.053600
8	x.fio2	0.126068
7	x.pao2	0.134642
17	x.paco2	-0.170129
25	x.liver.sofa	0.182759
22	x.resp.sofa	0.248155
4	x.resp	0.286132
0	x.hr	0.344661
21	x.urine	-0.622550
13	x.lactate	0.973669
24	x.cardio.sofa	1.287797
23	x.nervous.sofa	-1.890782
1	x.sbp	-1.981818
9	x.gcs	-2.112196

	keys	coef
19	x.hct	0.004637
5	x.temp	0.015487
27	x.kidney.sofa	0.082599
17	x.paco2	-0.120068
7	x.pao2	0.143045
25	x.liver.sofa	0.170099
8	x.fio2	0.186354
22	x.resp.sofa	0.284392
4	x.resp	0.351075
0	x.hr	0.373712
21	x.urine	-0.513221
24	x.cardio.sofa	1.413917
13	x.lactate	1.774189
23	x.nervous.sofa	-1.861808
9	x.gcs	-2.164041
1	x.sbp	-2.269263

Fig. S1. Coefficients for logistic model. Left: balanced. Right: imbalanced.

REFERENCES

[1] M. Singer et al., “The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3),” JAMA, vol. 315, no. 8, pp. 801–810, Feb. 2016.

[2] “Sepsis.” [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/sepsis. [Accessed: 18-Oct-2019].

[3] C. M. Torio and R. M. Andrews, “National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2011: Statistical Brief #160,” in Healthcare Cost and Utilization Project (HCUP) Statistical Briefs, Rockville (MD): Agency for Healthcare Research and Quality (US), 2006.

[4] A. Kumar et al., “Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock,” Crit. Care Med., vol. 34, no. 6, pp. 1589–1596, Jun. 2006.

[5] F. Daviaud et al., “Timing and causes of death in septic shock,” Annals of Intensive Care, vol. 5, no. 1, p. 16, Jun. 2015.

[6] R. Liu et al., “Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU,” Sci Rep, vol. 9, no. 1, pp. 1–9, Apr. 2019.

[7] A. Bravi, G. Green, A. Longtin, and A. J. E. Seely, “Monitoring and Identification of Sepsis Development through a Composite Measure of Heart Rate Variability,” PLoS One, vol. 7, no. 9, Sep. 2012.

[8] J. S. Calvert et al., “A computational approach to early sepsis detection,” Computers in Biology and Medicine, vol. 74, pp. 69–73, Jul. 2016.

[9] S. K. Nachimuthu and P. J. Haug, “Early Detection of Sepsis in the Emergency Department using Dynamic Bayesian Networks,” AMIA Annu Symp Proc, vol. 2012, pp. 653–662, Nov. 2012.