Final Report
Predicting Sepsis in ICU Patients

By Aisling Casey – 06/10/2021

Table of Contents

Overview

Sepsis is a leading cause of death in US hospital patients. Sepsis occurs when.. "when the body's response to infection causes tissue damage, organ failure, or death"[1][3]. Prompt intervention in sepsis patients can improve the likelihood of their condition improving significantly, while unnecessary treatment in non-sepsis patients drains hospital resources and can lead to worse patient outcomes.

The purpose of this project was to create a model to predict if an ICU patient would develop sepsis using hourly vital sign, laboratory and demographic data. This model could then be used in an ICU as a warning system for clinicians to consider further intervention during patient treatment.

Results

A gradient boost model that classified each hour of a patient's data as being in the pre-sepsis period or not was selected. It had an ROC-AUC of 0.84 and a precision-recall curve AUC of 0.496 on testing data.

Using the hourly classification predictions from the model, per-patient results were also determined, to assess model performance across all patients more accurately. This meant that if a patient was ever predicted to develop sepsis during their hospital stay, they were classified as pre-sepsis; if not, they were classified as non-sepsis patients. If a sepsis patient was only caught during sepsis or after, it was considered a false negative case. This classification scheme had lower performance metrics that the hourly classification, with an ROC-AUC of 0.716 and a precision-recall curve AUC of 0.38 on testing data.

Use cases and recommendations for the classification scheme thresholds, which affect the false positive and true positive rates, as well as future directions are explored in the Discussion section of this report.

Methods

Data Wrangling

The data came from a data science competition[1], which provided the hourly data for over 40,000 different ICU patients that included vital sign (e.g. heart rate), lab and demographic data, along with a sepsis label indicating if the patient had sepsis or not at that time.

Variable Type Column #	Vital Signs 1-8	Laboratory Values 9-34	Demographics 35-40	Sepsis Label 41
t_0	…	…	…	…
t_1	…	…	…	…
….	…	…	…	…
t_n	…	…	…	…

Figure 1: Data structure of one patient's dataset; there are as many rows as hours the patient spent in the ICU. There are 40,366 patient datasets in total.

Exploratory Data Analysis

Of 40336 patients available in the data set, 7.27% develop sepsis at some point during their hospital stay. Of the 1552210 data points in the data set (each one representing an hour) 9.2% occur during the pre-sepsis period of a sepsis patient. So, this is a very imbalanced data set.

Vital Signs

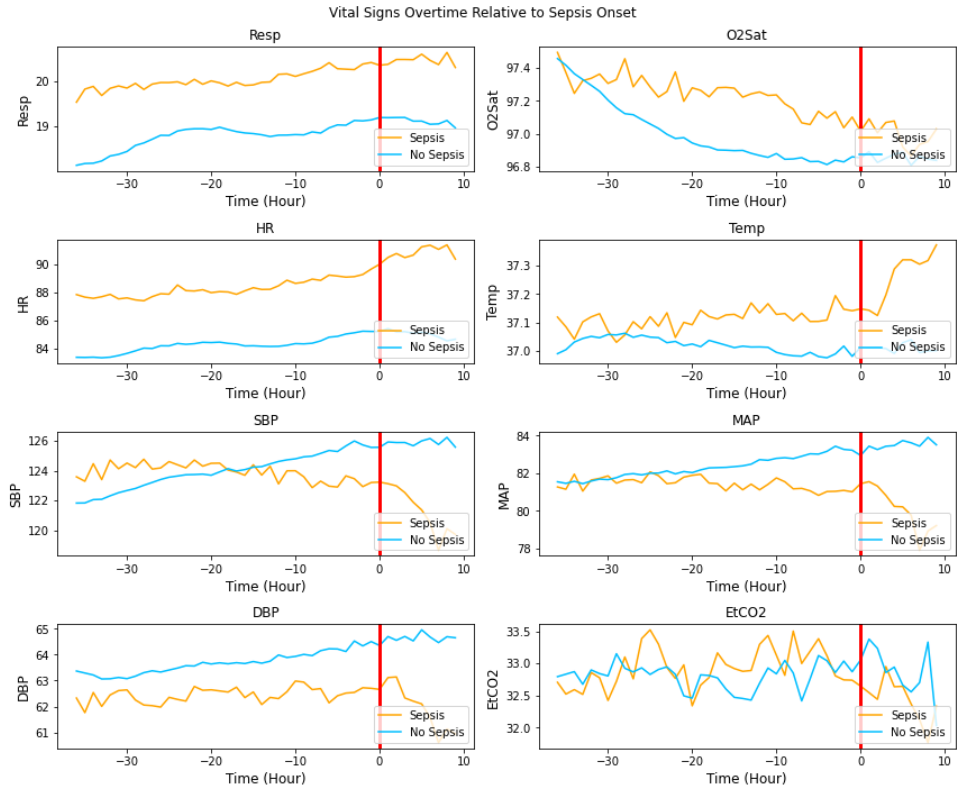


Figure 2: Average vital sign values relative to sepsis onset (red line) for sepsis patients compared to a random selection of values for non sepsis patients.

As seen in figure 2, there are some clear differences in the average time course of vital signs between sepsis and non-sepsis patients. In particular, sepsis patients have consistently higher average respiration & heart rate at any point compared to non-sepsis patients. Decrease in blood pressure is clearly seen in all sepsis patients post sepsis onset; pre-sepsis onset, SBP and MAP shows the strongest downward trend a few hours out, with DBP not as clear. O2Sat is a bit puzzling, as you'd expect it to clearly go down for sepsis patients, more so than non-sepsis patients, but the non-sepsis patients have a starker downward trend.

Lab Values

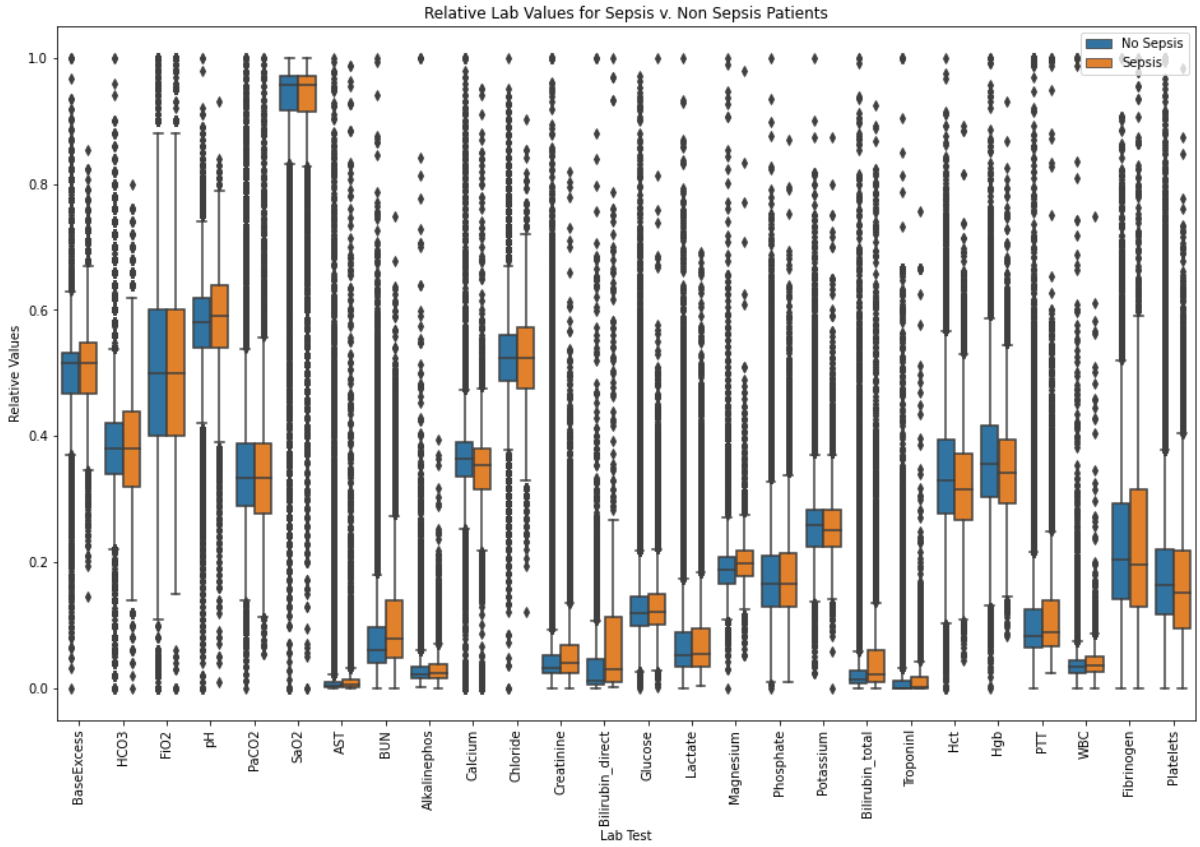


Figure 3: Distribution of relative lab values for sepsis vs. non sepsis patients.

As seen in figure 3, almost no clear pattern exists between the groups of any of lab values. For example, while there is a visual difference in distributions between groups for Bilirubin direct, that could easily be attributed to the small number of lab values available (i.e. noise). Bilirubin_total is the only exception, having lab values for over 30% of patients, and a seemingly different distribution. Still, there may be latent interactions between variables that are impossible to tell from this graph.

Demographic Data

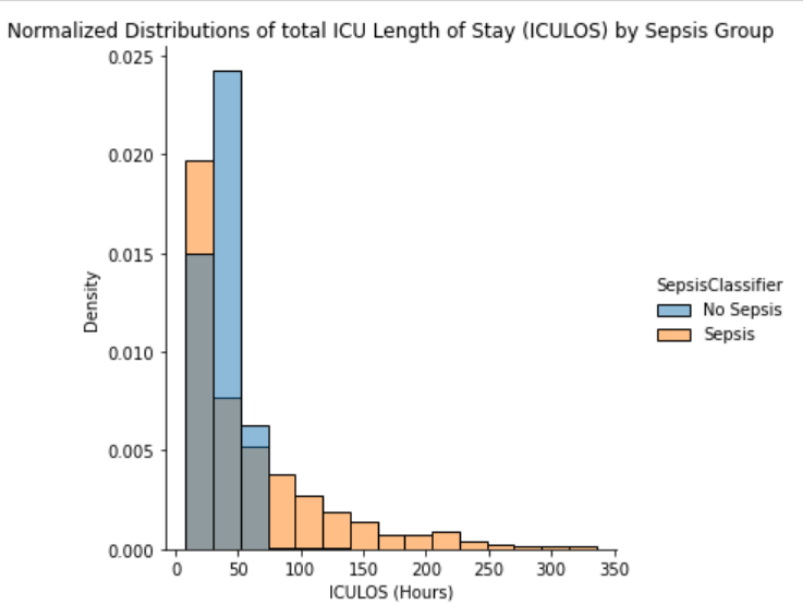


Figure 4: Normalized Distribution of ICU length of stay, sepsis vs. non-sepsis patients

Marginal differences existed between most demographic data, such as 7.71% of males having sepsis compared to 6.71% of females. By far the largest difference was ICU length of stay; a greater share of Sepsis patients stayed longer than 50 hours in the sepsis vs. non-sepsis patients, as seen in figure 4.

Preprocessing

The following changes were made to the data to prepare it for modeling, in this order:

1. Removal of outliers in laboratory & vital sign data
2. Interpolation of vital sign data
3. Forward filling of lab value data
4. Filling in remaining vital sign & lab data with median data
5. Addition of indicator variable column for lab values (if that lab value was present for that patient currently or anytime moving forward)
6. Addition of change in vital sign columns, for previous one, two and three hours.
7. Addition of pre-sepsis classifier column; 1 if period before sepsis onset, 0 otherwise.
8. Log transform of skewed distributions
9. Prevention of data leakage in test & train sets – all data of each patient was placed either in the training set or the testing set.
10. Min/Max scaling based on test dataset

Modeling & Performance Metrics

Logistic regression, random forest, and gradient boost models were created, using the pre-sepsis classifier column as the target variable. Randomized and grid search cross validation were used to determine optimal hyperparameters; sequential folding was used to prevent data leakage within the folds.

The ROC curves for all models are shown in figure 5. They had AUCs between 0.79 and 0.85 for the testing data. The random forest model overfits to its training data, but still has the highest AUC for testing data at 0.85. The gradient boost model has a similar testing data metric at an AUC of 0.84, without overfitting of the training data. The Logistic Regression model proved the least skilled with a ROC AUC of 0.79.

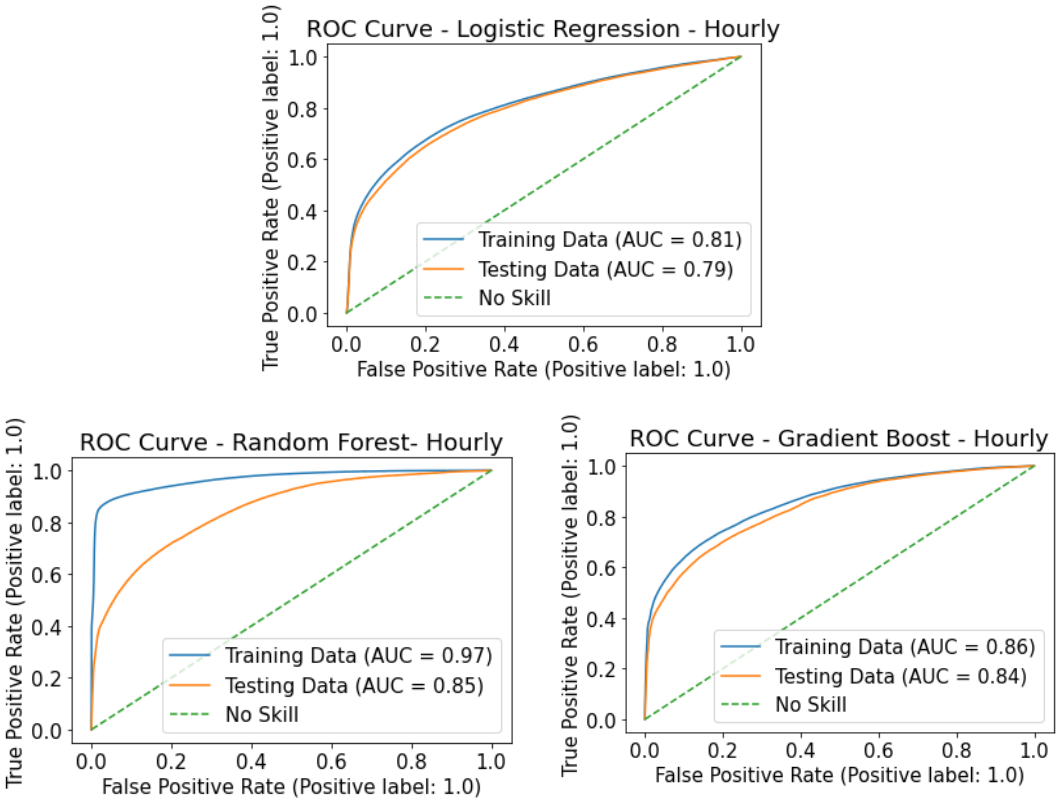
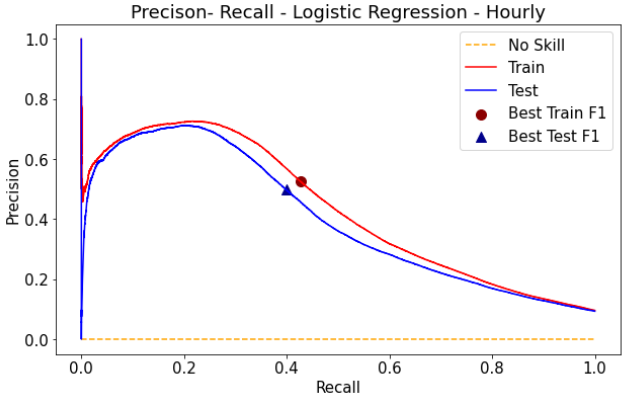


Figure 5: ROC curves for all three models with the hourly classification scheme.

The precision recall curves tell a different story, because this is an imbalanced data set. While the models do still show skill in distinguishing hourly patient data in the pre-sepsis period from those that are not, their AUCs are not as relatively strong as they are for the ROC curves with values between 0.39 and 0.52. The AUC patterns across models are similar to that of the ROC curves.



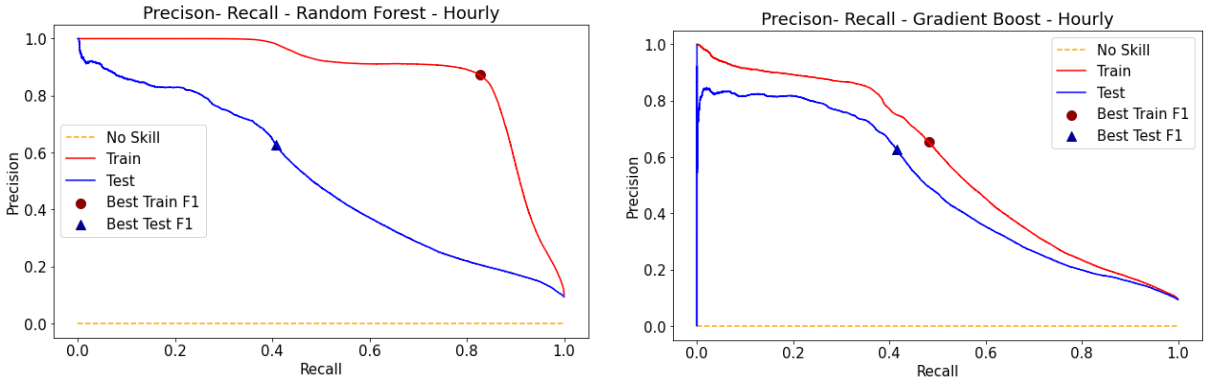


Figure 6: Precision-recall curves for all three models with the hourly classification scheme. The testing data has a Precision-Recall AUC of 0.398 with the logistic regression model, and 0.514 with the random forest model, and 0.496 with the gradient boost model

While the random forest model produced the highest AUC for the ROC and precision and recall curves the gradient boost model had very similar metrics. Ultimately the gradient boost model was chosen because the training data overfit less than for the random forest so it should generalize better to other unseen data, and the gradient boost models took less time to build.

Once the gradient boost model was selected, various probability thresholds were explored. This shown in the confusion matrixes below, illustrating in detail how the model performs when favoring precision or recall. When the threshold is set to 30%, very few hours are false positive classifications (1.6%) at the expense of a low true positive rate (37%). At 21.6%, the true positive rate grows to 41.5%, while the false positive rate only grows to 2.5%. Finally at a threshold of 10%, there are more false positive cases than true positive ones, though the percentage of true positive cases (67%) is still greater than false positives (17.9%).

Hourly Confusion Matrix: Counts

Actual	Predicted					
	0	1	0	1	0	1
0	344789	75264	409363	10690	413244	6809
1	13861	29275	25214	17922	27140	15996
Threshold	10%		21.6%		30%	

Hourly Confusion Matrix: Rates

Actual	Predicted					
	0	1	0	1	0	1
0	82%	17.9%	97.5%	2.5%	98.4%	1.6%
1	32.1%	67%	58.5%	41.5%	62.9%	37%
Threshold	10%		21.6%		30%	

Figure 7: Confusion Matrices for hourly data. The first table contains hourly counts while the second contains rates according to actual patient class.

A classification report is show below for a threshold of 21.6%. The overall accuracy of the model is quite high, as expected with imbalanced classes; importantly, the f1-score for the positive class is reasonably high at 0.5.

	precision	recall	f1-score	support
0.0	0.94	0.97	0.96	420053
1.0	0.63	0.42	0.50	43136
accuracy			0.92	463189
macro avg	0.78	0.70	0.73	463189
weighted avg	0.91	0.92	0.92	463189

Figure 8: Classification report for hourly data, set to a threshold of 21.6%.

While the model has proven to have skill in classifying whether data from a specific timepoint indicates pre-sepsis, ultimately what makes the model useful being able to predict sepsis in all patients, not just the ones that stay in the ICU for a long time. In its current form, the classification scheme may be obscuring sepsis patients with short pre-sepsis ICU stays and showing good performance because it accurately classifies the data of sepsis patients in the ICU for a long time before sepsis develops, but not those with shorter stays.

Another classification scheme was developed, where a sepsis patient with a positive prediction at any point during the pre-sepsis period was considered a pre-sepsis patient; a sepsis patient that was only caught during sepsis or after was considered a false negative case. Non-sepsis patients with positive predictions at any point during their stay were considered false positive cases. The ROC and precision-recall curves for this by-patient classification scheme can be seen in figure 9 and 10 below.

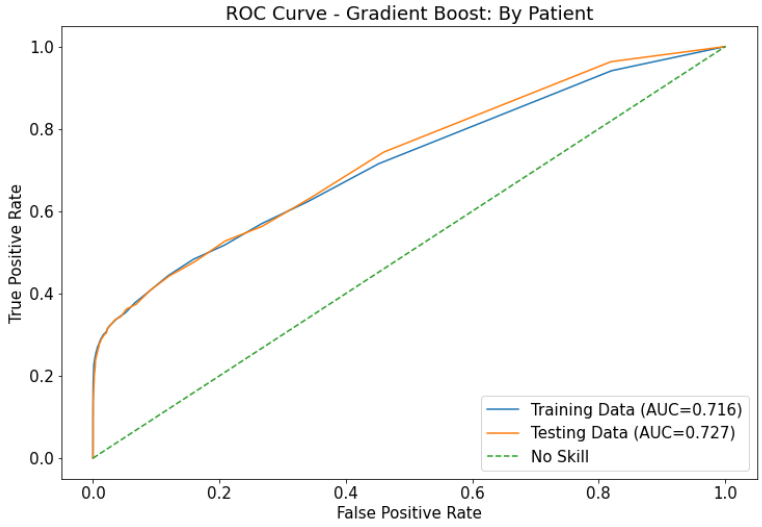


Figure 9: ROC curve for gradient boost model, by patient classification scheme. Testing data has an AUC of 0.727

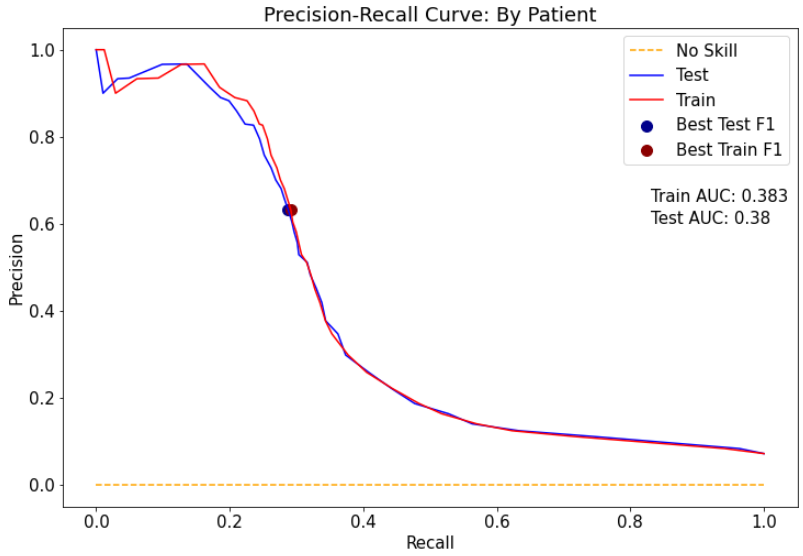


Figure 10: Precision-recall curve for gradient boost model, by patient classification scheme. Testing data has an AUC of 0.38

As expected, the results of this classification scheme were not as a strong as they were for the hourly sepsis classification. The AUC of the ROC curve of the testing data was reduced from 0.84 to 0.727; for precision-recall, it went from 0.496 to 0.38. Nonetheless, the model still shows some skill in distinguishing patients in the pre-sepsis period from non-sepsis patients.

Figure 10 shows a plot of true positive cases vs false positive cases. It illustrates the trade-off between precision and recall in terms of patient numbers – for example, at a threshold of 6%, about 650 of 870 sepsis cases are caught; but about 5,000 of 11,231 negative cases are marked as positive.

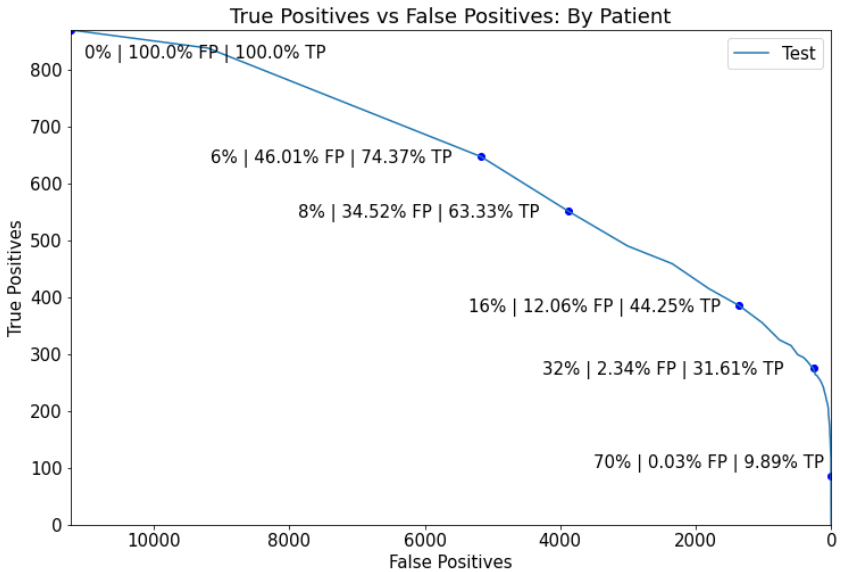


Figure 10: True positive cases by false positive cases, by patient

This tradeoff is further explored in confusion matrices below. They show the patient count and classification rates at three different thresholds: 4%, 11% and 34%.

By Patient Confusion Matrix: Counts

Actual	Predicted					
	0	1	0	1	0	1
0	2030	9201	8534	2697	10996	235
1	32	838	399	471	606	264
Threshold	4%		11%		34%	

By Patient Matrix: Rates

Actual	Predicted					
	0	1	0	1	0	1
0	18.1%	81.9%	76%	24%	97.9%	2.1%
1	3.7%	96.3%	45.9%	54.1%	69.7%	30.3%
Threshold	4%		11%		34%	

Figure 11: Confusion Matrices for by patient. The first table contains patient counts while the second contains rates according to actual patient class.

A classification report for the by patient scheme is show below for a threshold of 11%. The overall accuracy of the model is reduced significantly from the first model, down to 74% from 92%, as well as the f1 score for the positive class which has gone from 0.5 to 0.23. This threshold favors recall over precision – that is, identifying positive patients over not identifying negative patients.

	precision	recall	f1-score	support
0	0.96	0.76	0.85	11231
1	0.15	0.54	0.23	870
accuracy			0.74	12101
macro avg	0.55	0.65	0.54	12101
weighted avg	0.90	0.74	0.80	12101

Discussion

Sepsis is notorious for manifesting itself differently in different patients, thus making it difficult to predict. Using hourly data from over 40,000 patients in the ICU, the selected gradient boost model has skill in distinguishing between sepsis and non-sepsis patients in the pre-sepsis period. Though it performs better on classifying if a patient is in the pre-sepsis period every hour, to assess model performance across all patients it was necessary to use a wholistic, by-patient classification scheme.

The classification performance of the model then depends on the probability threshold set for identifying if a patient's hourly data is in the pre-sepsis period or not. This can be set according to the use case of the model. Three use cases, with extremely precise, extremely sensitive or in-between model performance are discussed below.

One use case is for the model to be used as a "Critical Sepsis Intervention Warning". This model results in extremely confident positive predictions. So, if a patient is classified as being in the pre-sepsis period, then the clinician should assume that they will develop sepsis and intervene accordingly. A threshold of 34% yields a model acceptable for this use case, with a false positive rate of 2% and a true positive rate of 30%. Looking at patient numbers, this still translates to about half of the positively identified patients receiving unnecessary intervention with the current model. The risks of intervening unnecessarily would need to be weighed against failing to intervene before using the model in this manner. Furthermore, many true sepsis cases (70%) are missed at this threshold.

Another use case is for the model to be used as a "Sepsis Concern Eliminator". In this scenario, if the model classifies a patient in the negative class, then they are extremely unlikely to develop sepsis – far less likely than the average incidence of sepsis in ICU patients. A threshold of 4% yields a model reasonable for this use case, with a false negative rate of 3.7% and a true negative rate of 18.1%. Of course, the ideal clinical tool would accurately identify all patients soon to develop sepsis, not some of the patients that will not develop sepsis. Still, using the model in this way would help to ensure that hospital resources such as antibiotics and clinician attention are not spent where they are not needed, and the model proves considerable skill in being able to do so.

A final use case if for the model to be used as an "Elevated Sepsis Intervention Warning". For this use case, the model warns if a patient is at higher risk of sepsis. It would be intended for the clinician to reference as they go about their standard treatment, not to rely upon. With a threshold of 11% the model best supports this use case, having a true positive rate of 54% and a false positive rate of 24%. Unfortunately this model threshold picks up many false positive patients – about 2,700 for every 470 true sepsis patients. Nonetheless, it could help the clinician identify pre-sepsis patients that they would otherwise not consider to be at risk of developing sepsis and monitor them accordingly.

In the end, the "Elevated Sepsis Intervention Warning" is recommended, because it could augment clinical intervention decisions while not being relied upon too greatly. But it is recognized that the model would still have limited utility in this regard.

To improve the model, one suggestion would be to train and use it only on patients that have been in the ICU for 24 hours or more. This would limit the scope of the model's utility, but could improve overall classification performance. This is based on the fact that the length of ICU stay was such an important feature in the model, as well as the higher performance of the model with the hourly classification scheme. Additionally, it would be interesting to look at the distributions of lab and vital sign values between sepsis patients that the model identified versus ones they did not – perhaps there are certain

patient profiles it is better at distinguishing from non-sepsis patients than others, which would be useful insight for clinicians.

It is also worth mention that the competition that provided this data called for an algorithm to identify sepsis six hours before sepsis onset[1]. With the supervised models used here, which build models based on individual data points, it was not possible to do this with any accuracy. Another group in the competition found success in doing so using a recurrent neural net[4]. Recurrent neural nets can accept sequences of data as input, such as a patient's hourly data, to identify patterns overtime in a single entity. Other than the change in vital sign columns added to the data during preprocessing, this information was essentially lost on the models used in this project, because each patient's data point is fed into the model separately. Making use of a patient's data trends would be critical in making a sepsis prediction with any reliable timing accuracy.

The final model chosen had skill in predicting whether a patient was going to develop sepsis but would ultimately not be recommended for use in a clinical setting, given its performance on the training data at various probability thresholds. Future work using deep learning models such as recurrent neural networks is likely to be more useful at this task.

Citations

[1] Reyna, M., Josef, C., Jeter, R., Shashikumar, S., Moody, B., Westover, M. B., Sharma, A., Nemati, S., & Clifford, G. (2019). Early Prediction of Sepsis from Clinical Data -- the PhysioNet Computing in Cardiology Challenge 2019 (version 1.0.0). *PhysioNet*. https://doi.org/10.13026/v64v-d857.

[2] Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, Clifford GD, Sharma A. Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge. Critical Care Medicine 48 2: 210-217 (2019). https://doi.org/10.1097/CCM.0000000000004145

[3] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, Hotchkiss RS, Levy MM, Marshall JC, Martin GS, Opal SM, Rubenfeld GD, van der Poll T, Vincent JL, Angus DC. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016 Feb 23;315(8):801-10. doi: 10.1001/jama.2016.0287. PMID: 26903338; PMCID: PMC4968574.

[4] Liu, L., Wu, H., Wang, Z., Lieu, Z, Zhang, M. Early Prediction of Sepsis From Clinical Data via Heterogeneous Event Aggregation. arXiv.org. 2019 Oct 12; 1910.06792v1.