337977045
316250877

Github: https://github.com/TamirShazman/Data-analysis-and-presentation.git

In folder HW1

Executive Summary:

- We achieve an F1 score of 0.68 on the provided test set using a bidirectional RNN model together with fully connected layers.
- Out of the ~40 features, we limit ourselves to 14 features based on our online research on Sepsis and perform various transformations that capture trends such as mean, variance, MAR and time-series. This approach seems to be validated in our Post Analysis.
- Our model performs best on men and on patients whose unit is not marked.

EDA and feature engineering:

In this section, we analyze 14 features that we found from our exploration and research to be related to Sepsis or indicative of serious illness. While we also examined other features, we do not expand on them here for the sake of brevity. We chose to use this limited amount of features in order to avoid our model discovering spurious correlations which could negatively impact our model's OOD inference on the test set.

We begin by examining the rate of non-null values, mean and variance of our features. We employ confidence intervals (unadjusted) as a tool for quickly identifying features that are different between the sepsis and non-sepsis population. We understand that these CI do not have statistical significance. However, they provide an indication of the likely amount of difference between the two groups given the variance inherent in that feature. Because the cost of adding a feature to our model is low, we are not concerned with the problem of multiple testing.

Based on our clinical research, the features that are clinically relevant to pursue in depth are:

1. Temperature - because Sepsis can cause fever or low body temperatures (https://www.sepsis.org/sepsis-basics/symptoms/)

2. Heart Rate - because Sepsis can cause elevated heart rate (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5538252/#:~:text=One%20of%20the%20most%20common,typically%2C%20a%20primary%20respiratory%20alkalosis.)

3. Respiratory Rate - because Sepsis can caue elevated respiratory rates (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5538252/#:~:text=One%20of%20the%20most%20common,typically%2C%20a%20primary%20respiratory%20alkalosis.)

4. WBC - Because Sepsis can cause an increase or decrease in the number of white blood cells

5. Lactate - because Sepsis can cause an increase in lactate levels (https://www.sepsis.org/sepsis-basics/symptoms/)

6. Base Excess - Sepsis can trigger metabolic acidosis which can be identified by excess base levels (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6402550/#:~:text=Sepsis%20causes%20hemodynamic%20instability%20through,prognosis%20in%20critically%20ill%20patients.)

7. O2 Saturation - Due to organ dysfunction, sepsis patients often suffer from low oxygen levels. (https://www.sepsis.org/sepsis-basics/symptoms/)

8. SBP - The heart pressure of the patient (systolic) is a general indicator of patient health.

9. MAP - Similar to SBP, MAP is also a general indicator of patient health. Specifically, due to the critical nature of sepsis. Organ failure can cause a severe drop in arterial pressure.

10. Age - Age is often correlated with severity of illness and the likelihood to get infections

11. Gender - men have been found to be more susceptible to women in the literature (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4641283/#:~:text=Out%20of%20100%20patients%20with,females%20(p%3D0.30).)

12. ICU LOS - Logically speaking, a sepsis patient will spend much of his LOS in a state of sepsis often resulting in the final LOS viewed in the series being lower than for non-sepsis patients.

13. Unit - Despite not knowing the actual difference between the two units, we check (and indeed likely confirm) that different units are used for different severities of illness. Therefore, if our models identify that a patient is in a more critical state based on the unit, our model will assign that patient a higher likelihood of sepsis.

14. Hospital Admittance Time - how long between entry to the ICU until hospitalization in the general wing of the hospital. We view this feature as an indicator for patients who are already admitted to the hospital for long standing health problems who are likely to develop more serious illnesses, such as Sepsis.

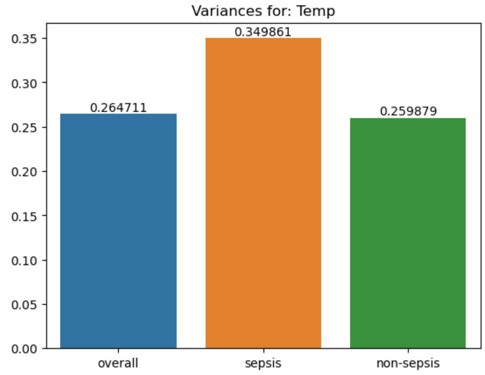
We examine these field from three main perspectives:

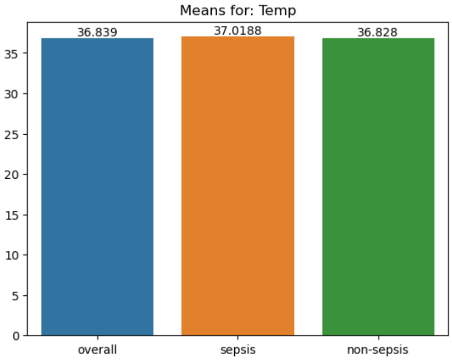
1. The rate of non-null values. This is informative to us for two reasons. One, if a field is very sparse, we cannot use its time series without imputing the missing values and introducing serious bias. Secondly, by comparing the rate of non-null values between Sepsis and non-Sepsis patients, we hope to discover is the field is MAR or MCAR. If the field is MAR, then the rate of non-null values can be indicative of Sepsis. For example, if a field is only filled for Sepsis patients than the rate of non-null values is MAR and a good signal for our model to use.
2. The mean of each patient's time series for each feature
3. The variance of each patient's time series for each feature

We now present our findings and explain how we translated these findings into feature engineering.

Temperature-

We found temperature to be a relatively sparse field. The average rate of non-null entries is roughly 33%. However, we did find indications that both the difference in mean temperature and the variance are different between Sepsis and non-Sepsis patients. We use both of these as features for our model.



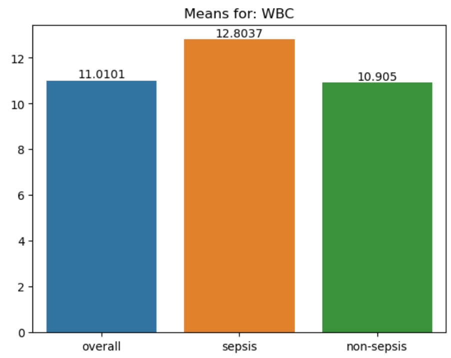


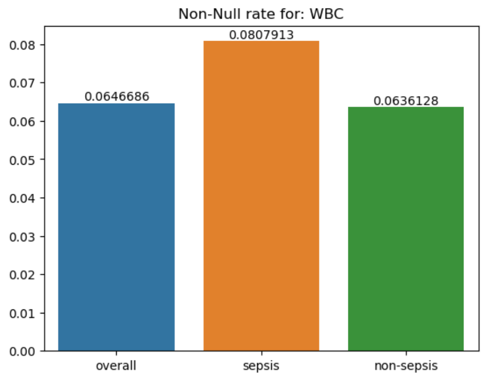
Heart Rate & Respiratory Rate:

We found that the non-null rates are high and the difference in means seems to be different. Therefore, we believe these to be good features to pass to our model as a time series.

WBC:

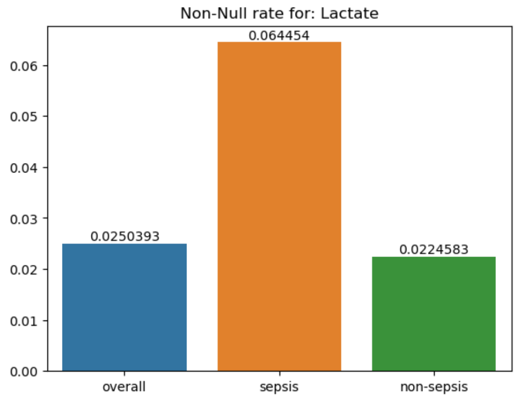
We find that the non-null rate is very low, therefore the series of WBC cannot be used. However, we see a difference between the non-null rate between the sepsis and non-sepsis population. We also see a difference between the means of the two populations as expected based on the literature. Therefore, we chose to use both the not-null rate and the means of WBC.





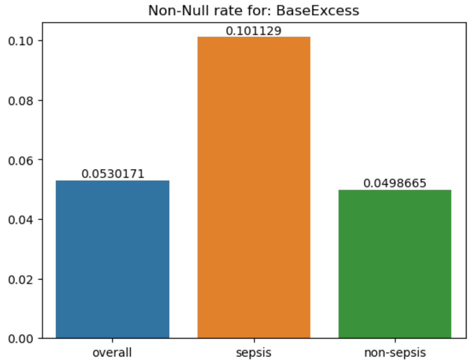
Lactate:

We cannot use the series because it is so empty, but the non-null rates show a difference as could be expected based on the literature.



Base Excess:

Like lactate, we cannot use the series because it is so empty, but the non-null rates show a difference as could be expected based on the literature.



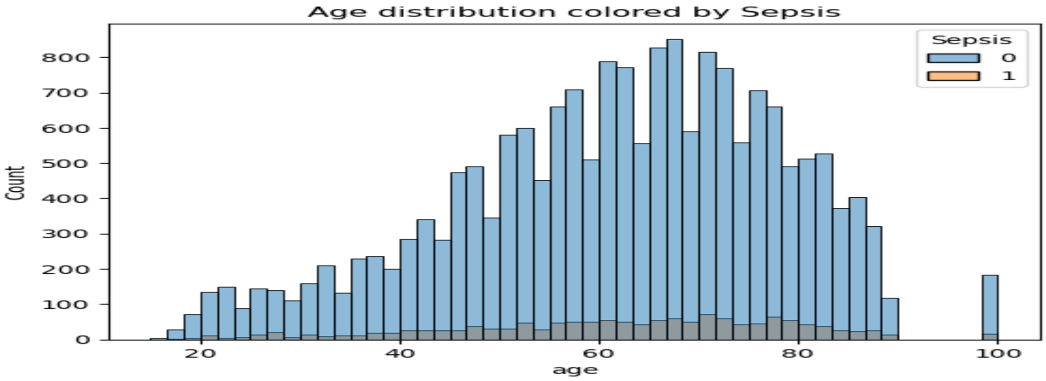
O2 Saturation, SBP, MAP:

These features are highly populated, O2 Saturation (85%), SBP (84%), MAP (86%). Although not finding differences between the mean and variation between the two groups, we use this feature as a time series based on the basis in the literature. We suspect that the significance of these features will be quick changes between time sections that do not affect the overall statistics.

Age:

We compared between sepsis and non-sepsis, with and without separation by gender. We checked by gender because the literature indicates a difference between men and women. Men are prone to sepsis at younger ages. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4641283/#:~:text=Out%20of%20100%20patients%20with,females%20(p%3D0.30).)

All the CI include zero. However, advanced age is highly correlated with serious illness in general. Indeed, all our data is taken from hospitals which already biases the data towards adverse health conditions relative to the general population. Therefore, we are not surprised to see that the overall distributions mode is in the 60s.



We will also note that we examined the group of ~200 patients aged 100. We suspect that perhaps given that the hospitals that the data is collected from have the resources to create and clean a dataset such as the one that we are using, they are top line hospitals. Perhaps, the centenarians were specifically brought to receive treatment there, thus explaining the aberration from the natural age distribution.

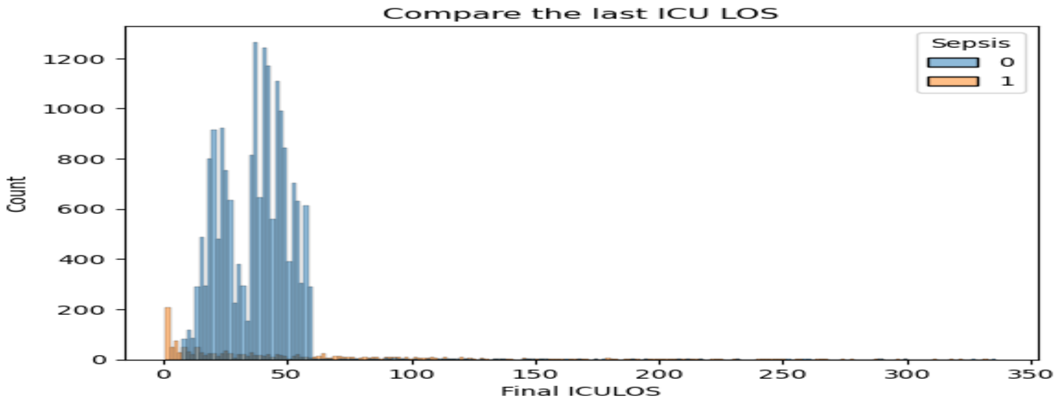
We pursued this hypothesis by checking if the centenarians were especially sick, which would warrant special treatment. However, using the proxy of ICULOS we did not find significant differences between the centenarians and the rest of the population.

Hospital Admittance Time:

This field seems to be somewhat significant. The bootstrap CI does not include zero, but it is not far $(-51.1, -2.88)$. If we would adjust for multiple tests, it is likely (and certainly using the conservative Bernoulli adjustment) that the difference would not be significant. However, logically we believe this to be a good indication of serious illness and therefore we use this feature.

Final ICU LOS:

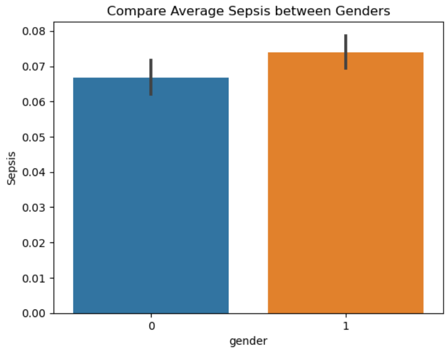
This field seems to be very significant based on both the distribution and the confidence intervals. We believe this will be a significant feature. As we can see, even a simple decision tree based only on this feature could capture a large mass of the Sepsis patients.



Gender:

We find that gender is correlated to Sepsis at a p-value of around 0.05 using the Pearson's chi squared test.

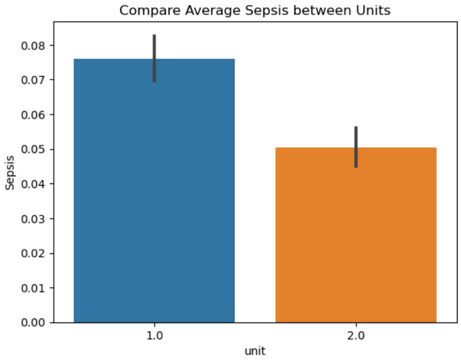
This makes sense because based on the literature men are more susceptible to sepsis and we find that to be the case here. However, we will note that would we adjust for multiple comparisons this would not be considered statistically significant (even by less conservative adjustments then Bernoulli). Below we see that there is a difference, but the confidence intervals overlap one another.



Nevertheless, based on the literature we still use this feature.

Unit:

We look at which unit the patients were treated and find that at a very significant level unit is correlated to sepsis, $7.21e^{-9}$. Even if we would control for multiple testing on all the tests, Pearson's chi squared test indicates deep correlation between sepsis and the unit.



It is likely that unit 1 treats more ill patients. However, we will note that this field is often null and therefore the statistic could be biased. We dealt with this by creating a three one-hot encodings: for unit 1, unit 2 and null unit.

Lastly, we will discuss how we dealt with null values.

Our primary treatment for dealing with null values are to use summary statistics which compress information from the non-null entries of a specific feature. We also use the presence of null values as its own feature.

For the time series that have null values we use two methods. First, linear interpolation. This assumes that entries in the time series are equidistant in time. In our case, this is true because each entry is a certain hour. We consider this to be a reasonable solution because if a truly unlikely event (which would require significant extrapolation as opposed to interpolation) occurred, the nurses would likely measure the change and the field would not be empty. We also noticed that the first entry of many time series is empty while the second entry is populated. Perhaps, this is due to the time it takes the nurses to take the measurements as they treat other patients. We filled the first value with the measurement afterwards.

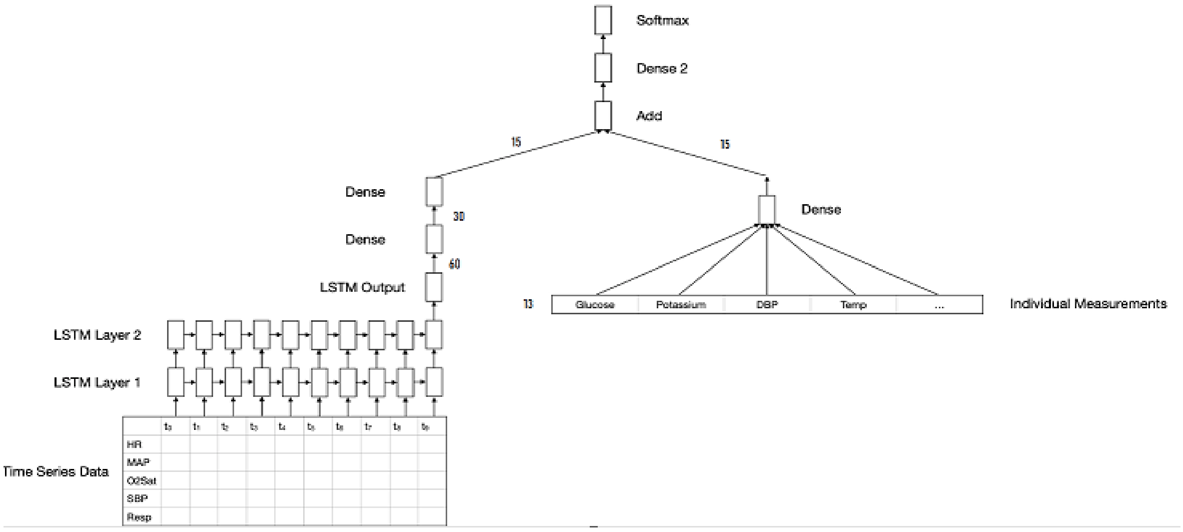
For the non-time series features, we used a simple average. We added most of these based on the difference in average between the sepsis and non-sepsis groups. The effect of using the average of everyone is to smooth the difference between the two groups, giving a higher weight to the non-sepsis group due to its size. This essentially inclines the model to predict non-sepsis if it has less information which is reasonable due to the data imbalance, only 7% of the train population has sepsis.

Algorithms:

Drawing inspiration from a previous study on sepsis prediction (https://github.com/nerajbobra/sepsis-prediction), our approach aims to leverage the strengths of LSTM and Fully connected network to improve prediction accuracy over time series data.

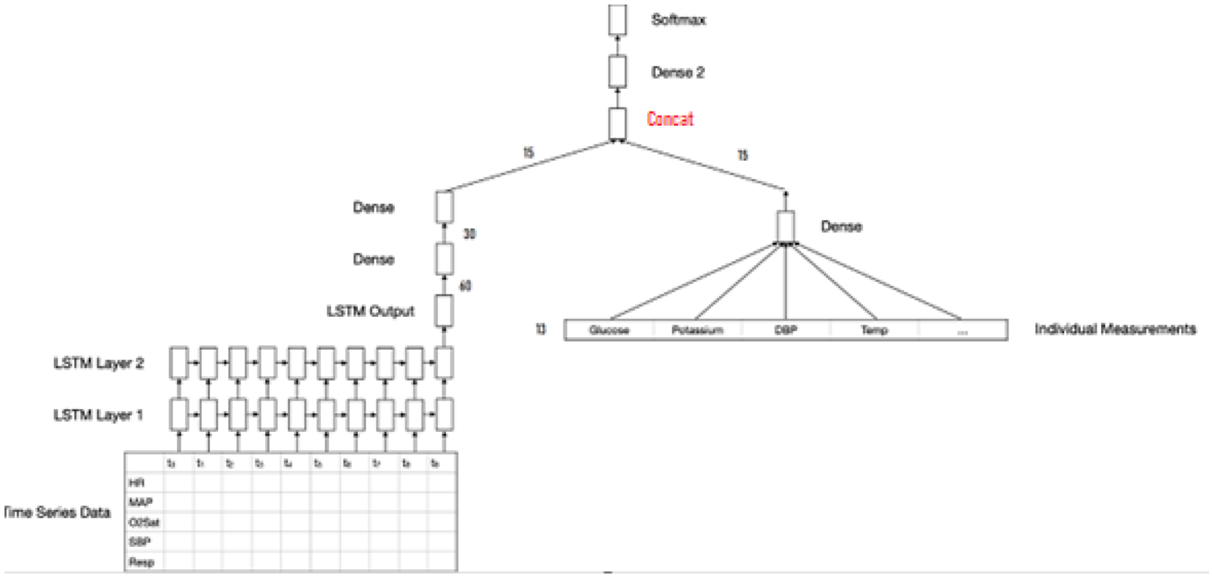
We elaborate on the 3 models that we use here:

1. "Add" model – In order to show our improvement over the previous study, we implemented almost the same model as the study suggests. The original model uses the following architecture:



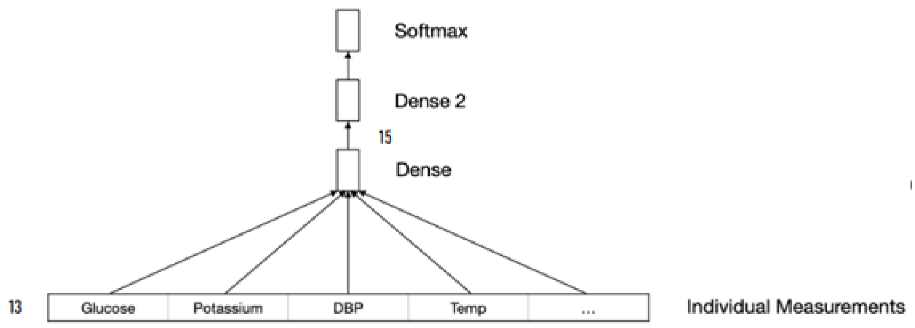
The LSTM model processes the time-series data, which is then fed into a 2-layer dense network. The individual (non-time series) measurements, on the other hand, are passed through a separate dense network. The outputs from both networks, each having the same shape, are added together, and propagated through another dense layer before reaching a SoftMax layer. The resulting output is a 2-dimensional vector that serves as an indicator for the presence of sepsis in the patient.

2. “Concat” model



Upon examining the "add model", one may question whether concatenating the right output with the left output would yield improved outcomes. This would allow the propagation of clearer signals from each part of the architecture. This notion prompted us to create the "Concat model", which shares the same architecture as the "add model", with the sole difference being that the left and right outputs are concatenated instead of added. The only necessary adjustment involved modifying the dimensions of the final dense layer.

3. “Ignore Time-Series” –



Given that time-series data is more intricate than single measurements data, one might wonder if the model could perform better without it. To investigate this possibility, we devised the "Ignore Time-Series model", which consists of a dense layer that follows the same architecture as that used for the single measurements data and is subsequently followed by another dense layer.

Hyperparameter selection, regularization

The hyperparameters for the Architect:

1. bidirectional-LSTM (Input size-5, hidden size-30, number of layer-2)
2. The sizes of all the dense layers can be found in the visual description above. The activation function following each of those layers is ReLU.

The hyperparameters for the Training:

1. Batch size-64
2. Optimizer-Adam with learning rate 0.0005
3. Loss function: For the unbalance data we use focal loss (https://arxiv.org/pdf/1708.02002.pdf) with $\alpha = 0.5, \gamma = 2$

The hyperparameters that are mentioned above are carefully chosen after training different values and checking their performance.

Regularization:

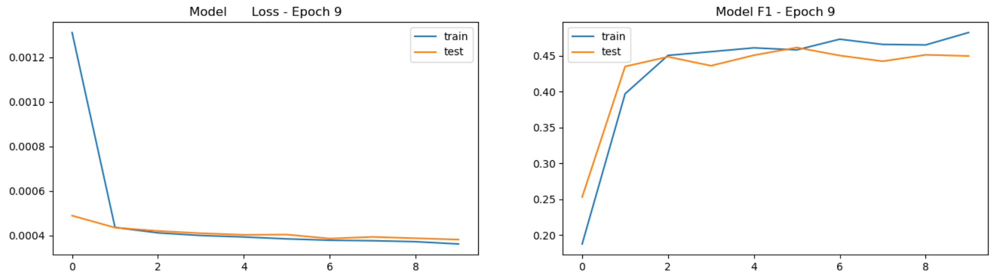
Dropout is a regularization technique commonly used in neural networks to prevent overfitting. However, if the validation loss is not significantly diverging from the training loss, it means that the model is not overfitting, and the training data is generalizing well to the validation data. In this case, using dropout may not be necessary, as it may not improve the model's performance.

Instead, we used early stopping to prevent overfitting. Early stopping involves monitoring the validation loss during training and stopping the training process when the validation loss starts to increase, indicating that the model is starting to overfit. This helps to find the point where the model achieves the best generalization performance on the validation set and prevents it from continuing to train and overfit the training data.

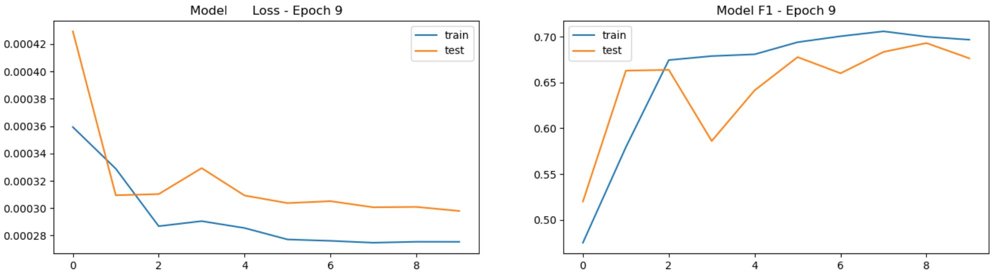
Training and validation results:

For training data, we use all the files that were in the train folder and for the evaluation we use all the files that were in the test folder. The following training graphs for each model:

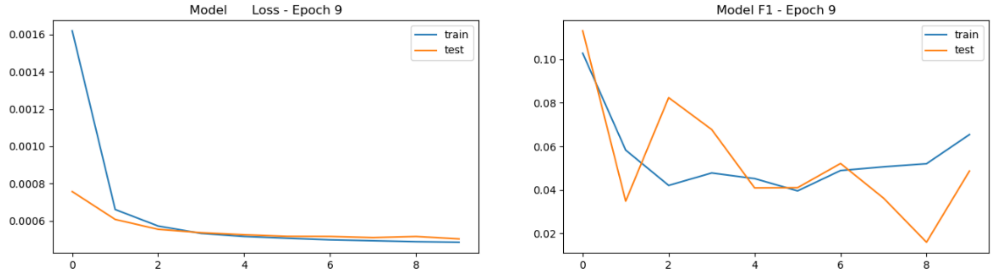
Add model:



Concat model:



Ignore TS model:



The Concat model proved to be the most effective among all models, achieving the highest F1 score of 0.68 on the test set. Notably, the results obtained from the Ignore model clearly indicate that the time-series data is indispensable to achieving optimal performance.

<u>Post analysis:</u>

One of the many problems Deep learning have is the problem to interpreted it complex networks. To interpreter our Concat model we implemented the following experiment:

- Features importance – To examine how to model interprets the features we'll sample with random a subgroup from the test data and follow the following procedure:
    a. Calculate the F1 score of the (we'll denote this value as Standard F1)
    b. Pick some feature we think are important and change its values.
    c. Calculate the new F1 and check the predictions.

<u>Experiment results</u>

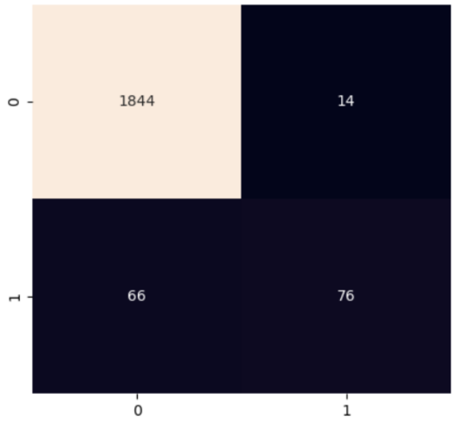
Feature Importance – We decided to change the following two features: 'Final ICULOS' and Temp_mean as follow:

'Final ICULOS:' $new\ value = \max(value - 30, 0)$

Temp_mean: $new\ value = value + 3$

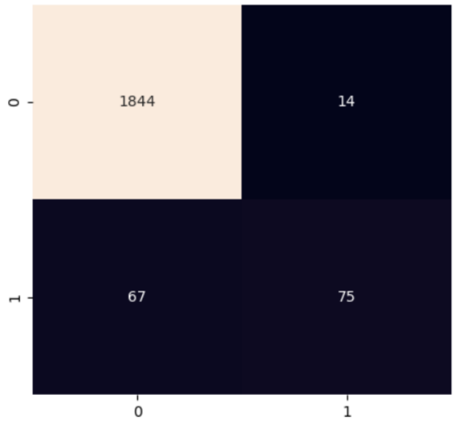
We’ll elaborate on the result here:

‘Final ICULOS’ – Standard F1- 65.5 with the following Confusion matrix:



	0	1
0	1844	14
1	66	76

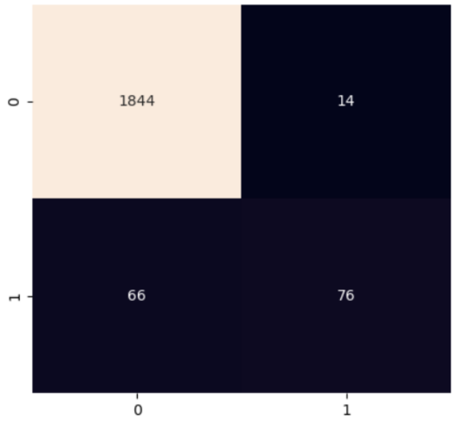
after change in value F1 – 64.9 with the following Confusion matrix:



	0	1
0	1844	14
1	67	75

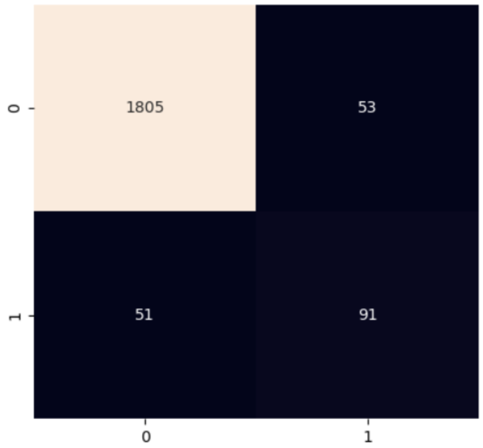
From those results we can conclude that ‘Final ICULOS’ is not that important to the model predictions as we drastically change the values and the model prediction almost reminds the same. This is surprising in light of differences we saw between the distribution of Final ICULOS in the exploration.

Temp_mean - Standard F1- 65.5 with the following Confusion matrix:



	0	1
0	1844	14
1	66	76

after change in value F1 – 63.6 with the following Confusion matrix:



We can see that the number of FP is much higher, thus we can conclude that the feature Temp_mean is highly relevant to the task as we change it by only adding 3 to it values and the F1 score dramatically degraded. Furthermore, it seems that the model interprets high values of Temp_mean with a plausible chance for Sepsis as expected from the literature.

For the sake of brevity, we include these two examples which show interesting behavior by our model because it prioritizes factors that we expected to be less important.

We also check our model on several sub-populations. We found that our model performs better on men than women (0.7 vs. 0.66). However, this could be random because when we switch the gender label the model's performance barely changes (decreases from 0.68 to 0.67).

It works best on unknown unit, second best on unit 2 and lastly on unit 1 (0.74 > 0.73 > 0.63). This could be due to the relative number of true positives in each of the unit groups. The least are found in unknown unit, then unit 2, then unit 1. Perhaps, our model is conservative about labeling 1 for sepsis and therefore performs better on populations where 0 is correct.

Summary and Discussion

In summary, we have a model that performs 0.68 on the provided test set. We used a bidirectional LSTM model on the timeseries data in conjunction with fully connected layers for the individual measurements. Despite using a relatively limited number of features, the model succeeds in capturing relevant information. We hope that by limiting the number of features to clinically relevant features, we avoid the pitfalls of spurious correlations and succeed at "deployment" on the real test set. Perhaps we see the success of this overall approach in the features that we expanded upon above, Final ICULOS and mean temperature. Our EDA showed that Final ICULOS should be significant but it does not seem to have been learned by our model. As opposed to temperature which was indicated as being significant in the literature but not in our EDA and our model does learn.

Our model performs better on men than women, a common bias in medicine. Together with the observation that it works better on patients whose unit is not marked, we notice that it performs best on the larger categories. This could indicate that our model suffers from underfitting due to lack of data. This could be caused by our use of a deep learning model with only 20,000 data points. Nevertheless, we achieve reasonable results.

We believe that further work could include using classic models or transformers in place of LSTM. Often transformers outperform RNNs in high resource environments with millions of data points, which led us to use RNNs. We also see further work including more features and checking the lift relative to our “conservative” feature selection.