Executive Summary

This study aimed to develop a model that can accurately predict sepsis up to 6 hours before a doctor's analysis. Through various approaches such as deep learning and machine learning, the team focused on data explanatory, feature selection and transformations to achieve significant results and insights.

- The data contains many missing values. Addressed the missing values and handled them using various imputation techniques.
- Feature transformation for timeseries data, using a single aggregation contributed more than using each value in time separately.
- Machine learning was more effective than deep learning approaches.
- We recommend that future researchers focus on machine learning to better address this critical issue. Overall, the study offers promising potential for improving sepsis diagnosis and treatment.
- We have tried various model such as LSTM, XGBoost, Random Forest, AdaBoost, GradientBoosting, LightGBM - All achieved F1 score surpassing the required F1 score.
- XGBoost appeared to be the most prominent candidate to address this problem.
- Thoroughly examined feature importance of trained models and implemented insights to the study.
- Used the common best practice of creating a benchmark in each step, to assess our improvement. In the feature selection, we used the top non missing value columns as benchmark, in the model development we used a logistic regression as the benchmark.

As part of trying to make this study creative and innovative, we also incorporated gifs into our data visualization.

Literature Review

We decided to start off by researching the previous literature in the domain. Giving us a better understanding of the assignment and its real-world implications.

According to the World Health Organization (WHO)[1] , in 2017 there were 48.9 million cases and 11 million sepsis related deaths worldwide, which accounted for almost 20% of all global deaths. In America, at least 1.7 million adults develop sepsis each year and approximately 270,000 adults die during hospitalization.
Sepsis is the body's extreme response to an infection. Common signs of sepsis include fever, shortness of breath, shivering or chills, severe pain, or discomfort, clammy or sweaty skin[2].
Using machine-learning models to detect bloodstream infections (such as sepsis) have been used

[1] https://www.who.int/news-room/fact-sheets/detail/sepsis
[2] https://www.verywellhealth.com/early-signs-of-sepsis-5498608

throughout the years and have shown worthy results. SERA[3], TREAT[4] and the third article[5] presented by the assignment paper are such examples of the use of such models in the field.

Early detection of sepsis can prevent deaths. Therefore, finding the best model to solve the following assignment has a potential impact of saving many lives.

Explanatory Data Analysis

Our raw training data contains 20,000 patients and a total of 766,884 records. On average each patient logged 38.34 records, meaning between 38 and 39 hours. Therefore, our raw training data contains a total of 87.54 years of hours in the ICU.

a. Describing available features in the dataset.

There are 41 features to the data, of which we decided to describe the following features:

Feature	Description	Reason Chosen
Lactate	Lactic acid (mg/dL)	High lactate levels may indicate poor tissue perfusion, which is a common sign of sepsis
WBC	Leukocyte count (count*10^3/µL)	Elevated or low WBC count can suggest infection or an immunocompromised
HR	Heart rate (beats per minute)	Tachycardia can be a sign of the body's response to infection and inflammation.
Temp	Temperature (Degrees C)	Fever or hypothermia can be indicative of infection or sepsis.
Resp	Respiration rate (breaths per minute)	Rapid breathing can be a sign of respiratory distress or compensation for metabolic acidosis, both of which can be seen in sepsis.
MAP	Mean arterial pressure (mm Hg)	Hypotension can be a sign of septic shock, which is a severe form of sepsis.
O2Sat	Pulse oximetry (%)	Low oxygen saturation can indicate respiratory failure, a potential complication of sepsis.
PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)	Altered PaCO2 levels can indicate respiratory distress or compensation for metabolic acidosis, which can be seen in sepsis.
Age	Years (100 for patients 90 or above)	Descriptive feature
Gender	Female (0) or Male (1)	Descriptive feature
SepsisLabel	Patients diagnosed with sepsis: $\begin{cases} 1 \; if \; t \geq t_{sepsis} - 6 \\ 0 \; if \; t < t_{sepsis} - 6 \end{cases}$ For non-sepsis patients, Sepsis Label is 0	Descriptive feature

[3] Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JJL, Tan GYH. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun. 2021 Jan 29;12(1):711. doi: 10.1038/s41467-021-20910-4. PMID: 33514699; PMCID: PMC7846756.

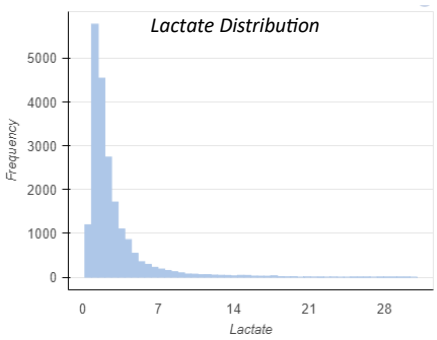
[4] Paul M, Andreassen S, Nielsen AD et al (2006) Prediction of bacteremia using TREAT, a computerized decision-support system. Clin Infect Dis 42:1274–1282.

[5] Roimi, M., Neuberger, A., Shrot, A., Paul, M., Geffen, Y., & Bar-Lavie, Y. (2020). Early diagnosis of bloodstream infections in the intensive care unit using machine-learning algorithms. Intensive Care Medicine, 46, 454-462.

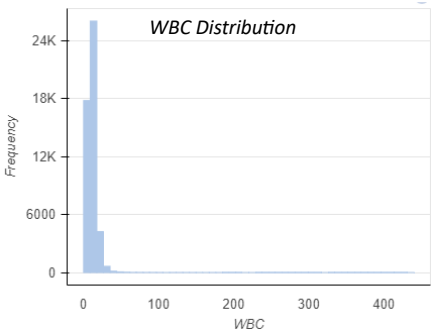
The relevance of all the features not with the “descriptive feature” in the reasoning column, were chosen due to their probability of being a highly insightful feature for predicting symptoms of sepsis, based on physiological changes related to sepsis.

b. Inspecting features distribution.

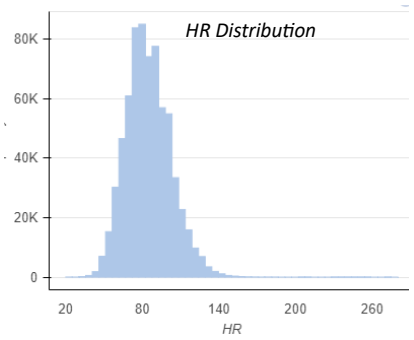
Lactate	
Min	0.2
Max	31
Median	1.89
Mean	2.68
Skewness	3.28
Missing Values	97.3% (!!!)

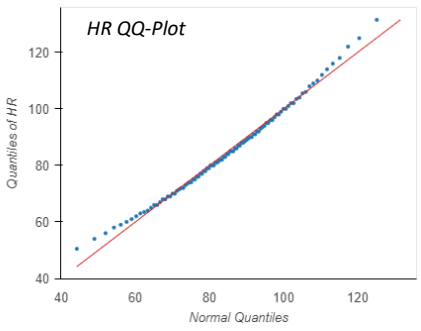


WBC	
Min	0.1
Max	440
Median	10.3
Mean	11.45
Skewness	11.99
Missing Values	93.6% (!!!)

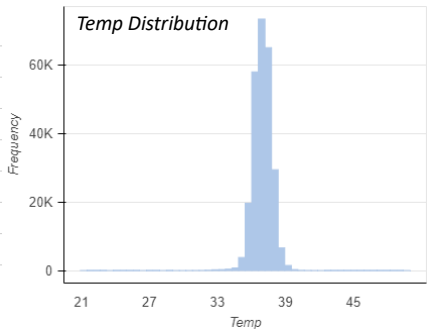


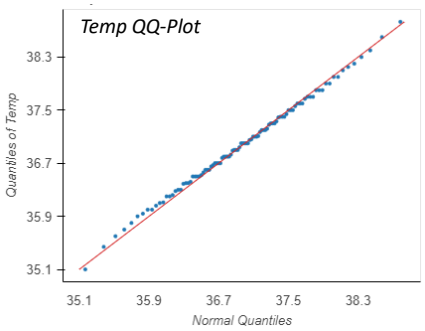
HR	
Min	20
Max	280
Median	84
Mean	84.6
Skewness	0.44
Missing Values	9.9%



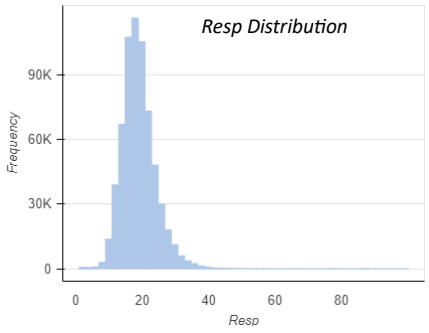


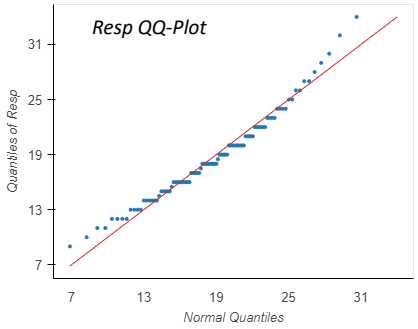
Temp	
Min	20.9
Max	50
Median	37
Mean	36.97
Skewness	-0.42
Missing Values	66.1%



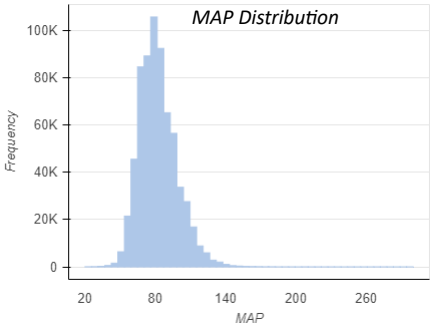


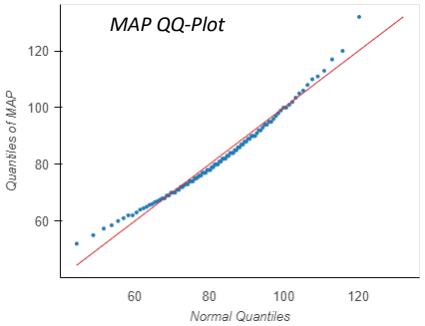
Resp	
Min	1
Max	100
Median	18
Mean	18.75
Skewness	1.03
Missing Values	15.4%



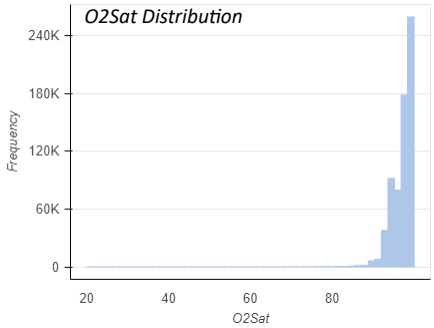


MAP	
Min	20
Max	300
Median	80
Mean	82.29
Skewness	1.04
Missing Values	12.5%

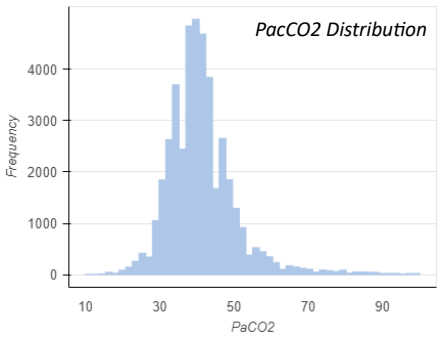


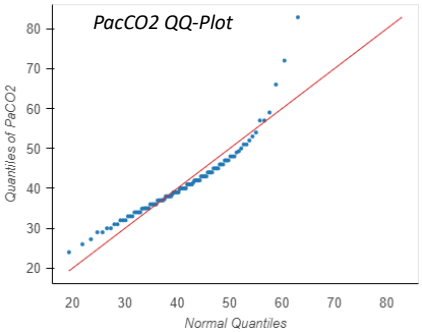


O2Sat	
Min	20
Max	100
Median	98
Mean	97.19
Skewness	-4.19
Missing Values	13.1%



PaCO2	
Min	10
Max	100
Median	91
Mean	41.16
Skewness	1.61
Missing Values	94.4%





Age & Gender

To get the correct distribution of age and gender, we needed to look at each patient as 1 record in the data alone. The minimum age is 15 and maximum 100, mean 61.6 and median 63.3. 44.45% of the training data patients are females.

Sepsis Label ("SepsisLabel")

The training data available is unbalanced, 7.07% of the patients in the data are diagnosed with sepsis, while 92.93% are not diagnosed with sepsis. We addressed this issue when choosing a model.

c. Comparative analysis between features.

I. Highly correlated raw features.

See Appendix 1 for the full heat map of the correlation between the raw data features. We decided to present here the pairs of features with correlation above 0.75 and below -0.75 (calculated using the accepted Pearson correlation).

Features	Correlation
Hct – Hgb	0.95
Bilirubin_direct - Bilirubin_total	0.95
MAP - DPB	0.85
BaseExcess – HCO3	0.85
SBP - MAP	0.78
Unit1 – Unit2	-1

Highly correlated features can be useful when we are researching feature selection. To reduce high dimensionality of the data and redundant columns, we can use a single feature from each highly correlated pair.

II. Hypothesis testing.

We aimed to investigate the typical symptoms associated with sepsis. The null hypothesis assumes that there is no significant difference in the recorded values of these symptoms between patients diagnosed with sepsis and those without.
In all the hypotheses below, we calculated the mean of a certain feature, and as our sample is large enough (20,000), under the assumption of normality (by the central limit theorem) we used the t-test to test the hypothesis. Using $\alpha = 5\%\ (0.05)$ as statistically significant.

Hypothesis 1: Sepsis patients will have a higher heart rate (HR) than non-sepsis patients.

We calculated at first the average HR for each patient and tested the hypothesis (see figure 1). The t-test yielded a statistically significant result with a p-value < 0.001, to reject the null hypothesis. We discovered a bias in the nature of the real-world ICU data we are working with. Each patient may be diagnosed with Sepsis at a different time throughout the time spam they are in the ICU. For example,

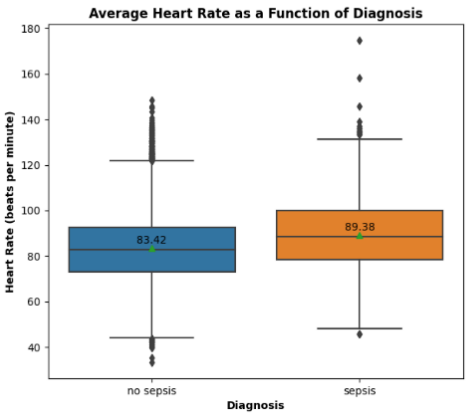


Figure 1. *Average HR per patient.*

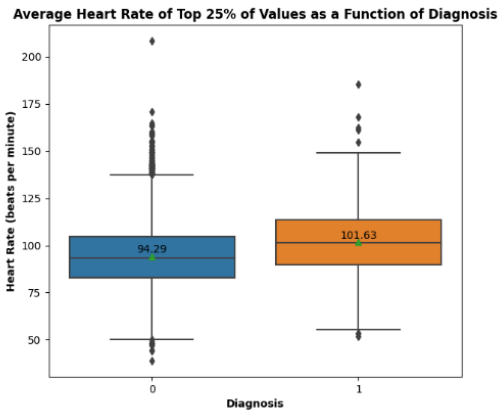


Figure 2. *Average of top 25% HR values per patient.*

a patient diagnosed with sepsis after 20 hours in the ICU, and then decides to leave the hospital, has 20 hours of HR data recorded prior to being diagnosed as opposed to a patient being diagnosed immediately when being admitted to the ICU. To address this issue, we decided to use the average of the top 25% of values of HR per patient (See figure 2). The t-test calculated for figure 2 was also statistically significant with a p-value below 0.001, rejecting the null hypothesis.

Hypothesis 2: Sepsis patients will have a higher temperature (Temp) than non-sepsis patients.

Testing both the average temperature per patient (see figure 3) and the average temperature of the top 25% values per patient (see figure 4), both resulted in statistically significant results indicating we can reject the null hypothesis based on our training sample, with a p-value below 0.001.

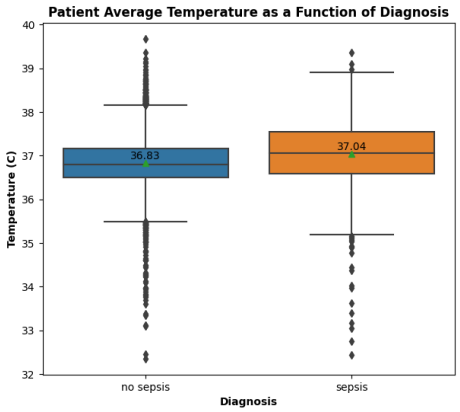


Figure 3. *Average Temp per patient.*

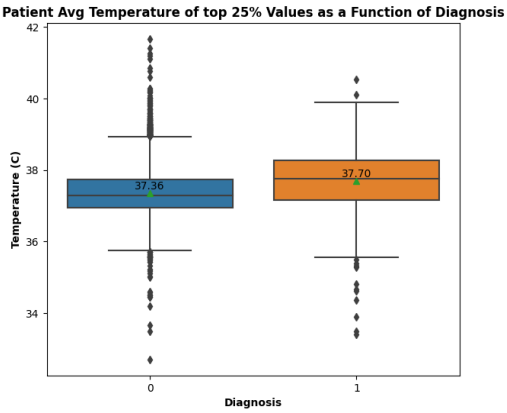


Figure 4. *Average of top 25% Temp values per patient.*

Hypothesis 3: Sepsis patients will have a lower O2Sat values than non-sepsis patients.

Testing both the average O2Sat per patient (see figure 5) and the average O2Sat of the bottom 25% values per patient (see figure 6), both resulted in non-statistically significant results indicating we cannot reject the null hypothesis based on the training data, with a p-value of 0.36 and 0.17 respectively.

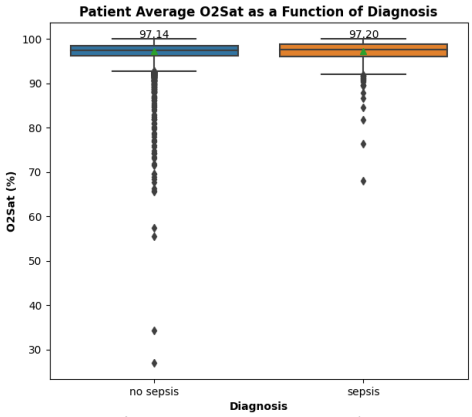


Figure 5. *Average O2Sat per patient.*

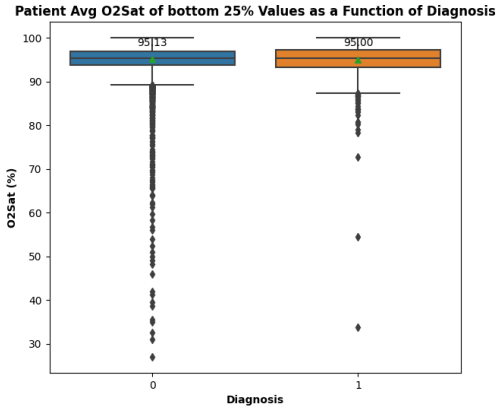


Figure 6. *Average of bottom 25% O2Sat values per patient.*

Hypothesis 4: Sepsis patients will be older than non-sepsis patients.

Testing the above hypothesis (see figure 7) the t-test yielded non-statistically significant results indicating we cannot reject the null hypothesis based on our training sample, with a p-value of 0.16.

These insights found from the statistically significant hypothesis results will contribute to our feature selection and transformations processing.

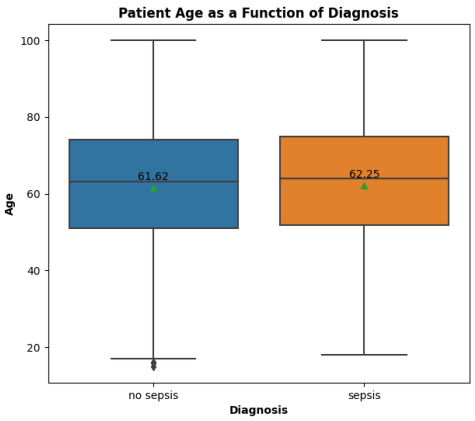


Figure 7. *Patients age hypothesis.*

d. Handling missing data

Missing value analysis on the entire rows in all training data, prior to any pre-processing (figure 8).

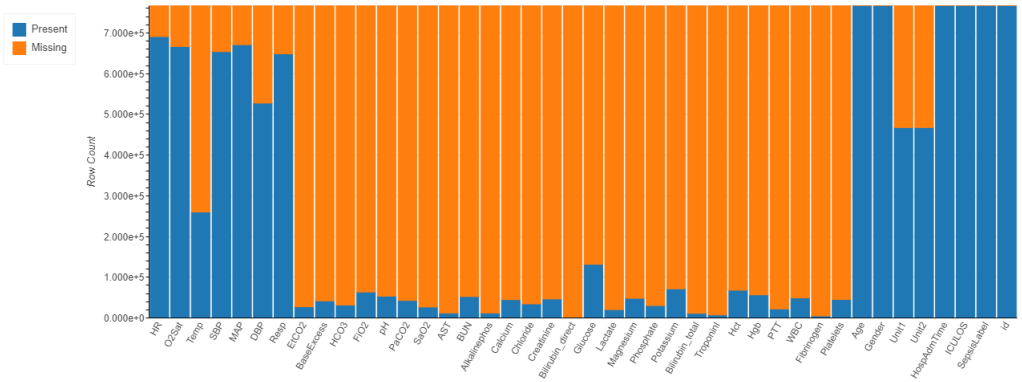


Figure 8. *Missing values, raw data.*

After processing the data as required to input into the model, dropping all redundant rows labelled as ones, we checked if there was at least one value in each column for each patient – which will allow us to use the value from a transformation on the full column. Only if the whole column is missing, we will get a nan value when using such transformations, therefore we felt this is the correct missing value analysis to look at (see figure 9 below).

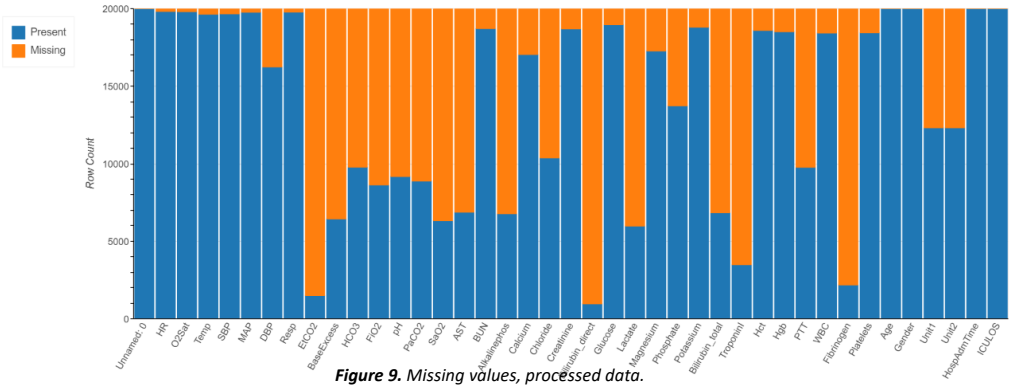


Figure 9. *Missing values, processed data.*

There are many ways to handle the missing data. While working on handling the missing data, feature selection and feature transformation simultaneously, we decided to put the different missing value imputation methods to test when running the different models.
We used mean imputation, nearest-neighbour imputation, and a zero-imputation method which we will justify when talking about the feature transformations.

Feature Engineering

a. Feature Selection

- **All features** – used as a sanity check / benchmark. Why select features if all features are resulting in the best result and we can run the models in a reasonable time?
- **Top non missing value columns** – our intuition was that these columns contain the most information.
- **Drop highly correlated columns** – During the EDA stage, we found pairs of columns which were in extreme correlation with each other, dropping these columns is generally accepted to cause higher performance in the models.
- **Forward Sequential Feature Selection** – commonly used greedy procedure that iteratively finds the best new feature to add to the set of selected features. This method was too time consuming for our computers, we believe it has the most potential to be the best of all the other methods, but as we were satisfied with the other results and were limited in time, we stopped implementing this method during evaluation.
- **Chat-GPT recommendation** – Asked ChatGPT to select the columns based on "his" opinion, we were curious to see if the chosen features were any better than the other methods used.

There are many more feature selection techniques, we chose to present only the methods we used and implemented throughout the project.

b. Feature Transformation

"The models' prediction was based mainly on patterns of the time-series variables…Values of parameters at a specifc time were infrequently used by the model. Patterns of parameter changes over time contain more information than their absolute values at a single, specific time-point."[6]

For the machine learning classifiers, we used this approach, which was easy to implement and resulted in great results with strong reasoning and backup from the literature. We decided to use the mean and standard deviation of each column as the informative pattern variable.
When using this transformation approach, filling the missing values in with mean and std equal to 0, is an informative imputation, representing non movement in the column values. Which will be shown to have resulted in the best performance classifiers.

[6] Roimi, M., Neuberger, A., Shrot, A., Paul, M., Geffen, Y., & Bar-Lavie, Y. (2020). Early diagnosis of bloodstream infections in the intensive care unit using machine-learning algorithms. Intensive Care Medicine, 46, 454-462.

Prediction

a. Report

We decided to present only the best configuration per model. We tested for each model the different imputations and the different feature selection techniques.

Algorithm	Hyperparameters	Missing Values Method	Feature Selection Method	Feature Transfo-mation	Training F1 Score	Validation Metrics		
						F1	Prec-sion	Re-call
Random Forest	The main hyperparameter we increased until we didn't achieve any improvement in the model performance was the "n_estimators". Best result at 1,000. Evaluated the different functions evaluating the quality of the split, chose "gini".	Zero-Imputation	Keeping all	std & mean	1.0	0.676	0.899	0.541
Gradient Boosting	As mentioned above, here we found "n_estimators" = 1,000 also the best option. Learning rate kept as default (=0.1), after using larger learning rates which caused degraded results.	Zero-Imputation	Keeping all	std & mean	0.96	0.71	0.843	0.614
AdaBoost	As mentioned above, here we found "n_estimators" = 1,000 to achieve the best performance with a learning rate of 1.	Zero-Imputation	Keeping all	std & mean	0.79	0.694	0.821	0.601
LightGBM	As mentioned above, here we found "n_estimators" = 2,000 to achieve the best performance.	Zero-Imputation	Keeping all	std & mean	1.0	0.705	0.907	0.576
XGBoost	Tested the "max_depth" param which defines the maximum tree depth of each estimator, best result was for 5. Used the "tree_method"="hist", which creates histogram bins for creating the trees. "n_estimators" = 2,000.	Zero-Imputation	Keeping all	std & mean	1.0	0.714	0.883	0.599
LSTM	Inputsize (number of columns)-40. Hidden_size-250x1. Outputsize-1 (prediction). batch_first-True. num_layers-3.	Forward fill when value present at start, backwards fill when missing at start, zero - imputation if missing all column.	select columns with less than 25% missing values. Top non missing column.	Data Type Transform: all values into Tensors.	0.0005 binary cross entropy loss	0.675 (**see appendix 3**)		

b. Post Analysis of Chosen Models

i. Performance of subsets

We tried to separate the data set into reasonable groups that will help us check how the model deals with an equal number of patients in the data but with different categories. After a lot of checking, we found equal and reasonable separations.
The first separation is based on Age and Gender, we use this separation since both features are the most basic features to collect per patient, and finding models which perform substantially superior then other model on certain subsets – can be a valuable insight to be used.

Then we separated the patients intro groups by their Mean Arterial Pressure (MAP) and Temperature (Temp) values. This separation is interesting because it allows us to analyze model performance based on two important clinical features, while we have found in the hypothesis testing above (hypothesis 2) that temperature is a statistically significant feature as a function of sepsis diagnosis.

We then predict our models on those groups and got these results:

Age Group 30-59				
Model	**Gender**	**F1 Score**	**Precision Score**	**Recall**
XGBoost	Male	0.727	0.935	0.594
	Female	0.689	0.879	0.567
AdaBoost	Male	0.694	0.847	0.588
	Female	0.675	0.836	0.567
RandomForest	Male	0.664	0.908	0.524
	Female	0.638	0.882	0.5
LSTM	Male	0.667	0.821	0.561
	Female	0.647	0.741	0.573

Low Temperature (below median of all values) Group				
Model	**Condition**	**F1 Score**	**Precision Score**	**Recall**
XGBoost	High MAP	0.68	0.847	0.568
	Low MAP	0.784	0.934	0.675
AdaBoost	High MAP	0.669	0.787	0.582
	Low MAP	0.76	0.876	0.672
RandomForest	High MAP	0.65	0.814	0.541
	Low MAP	0.757	0.961	0.624
LSTM	High MAP	0.693	0.787	0.618
	Low MAP	0.676	0.791	0.59

we can learn from it that:

- In the age group 30-59, XGBoost performs the best in terms of F1 score for both male and female patients.
- For the low temperature group, XGBoost also performs the best in terms of F1 score for patients with low MAP, while LSTM performs the best for patients with high MAP.
- The LSTM performs better than all other models on the Low Temp High MAP group (F1 score).

- In both age and temperature groups, the precision score is generally higher than the recall score. This indicates that the models have a higher tendency to correctly identify non-sepsis cases but may miss some actual sepsis cases. Depending on the context and the potential consequences of misclassification, it might be necessary to optimize the models to improve recall at the expense of precision.

To conclude this part, we believe that the prediction part shouldn't be based solely on 1 model. If we can separate the data into equal groups which one of them got significantly better results with one of the models, we should use this model on the next patient who will belong to this specific group.

ii. Interpretability

Interpretability is essential for understanding how a model makes its predictions and helps build trust in the model's decision-making process. There are many techniques used to provide interpretability for any model, we decided to present these methods for XGBoost (representing the tree based, machine learning models) and LSTM (representing the deep learning family of models).

XGBoost - Interpretability techniques for XGBoost primarily revolve around understanding the importance of each feature in the model. Feature Importance built in method: metric based on the number of times a feature is used to split the data across all trees. This can help in understanding which features contribute the most to the model's predictions.

LSTM (Long Short-Term Memory) - Attention Mechanism: This technique helps understand which parts of the input sequence have the most significant impact on the model's output. Attention mechanisms can be integrated into the LSTM architecture, allowing the model to weigh the importance of each input token or time step when making predictions.

Generally, deep learning models are more challenging to interpret, due to their large complexity.

We used the feature importance for 3 XGBoost, AdaBoost and Random Forest (see appendix 3) and got interesting results. These results may lead to the use of only the top importance features in the future feature selection or in a lack of time to prediction.

Summary and Discussion

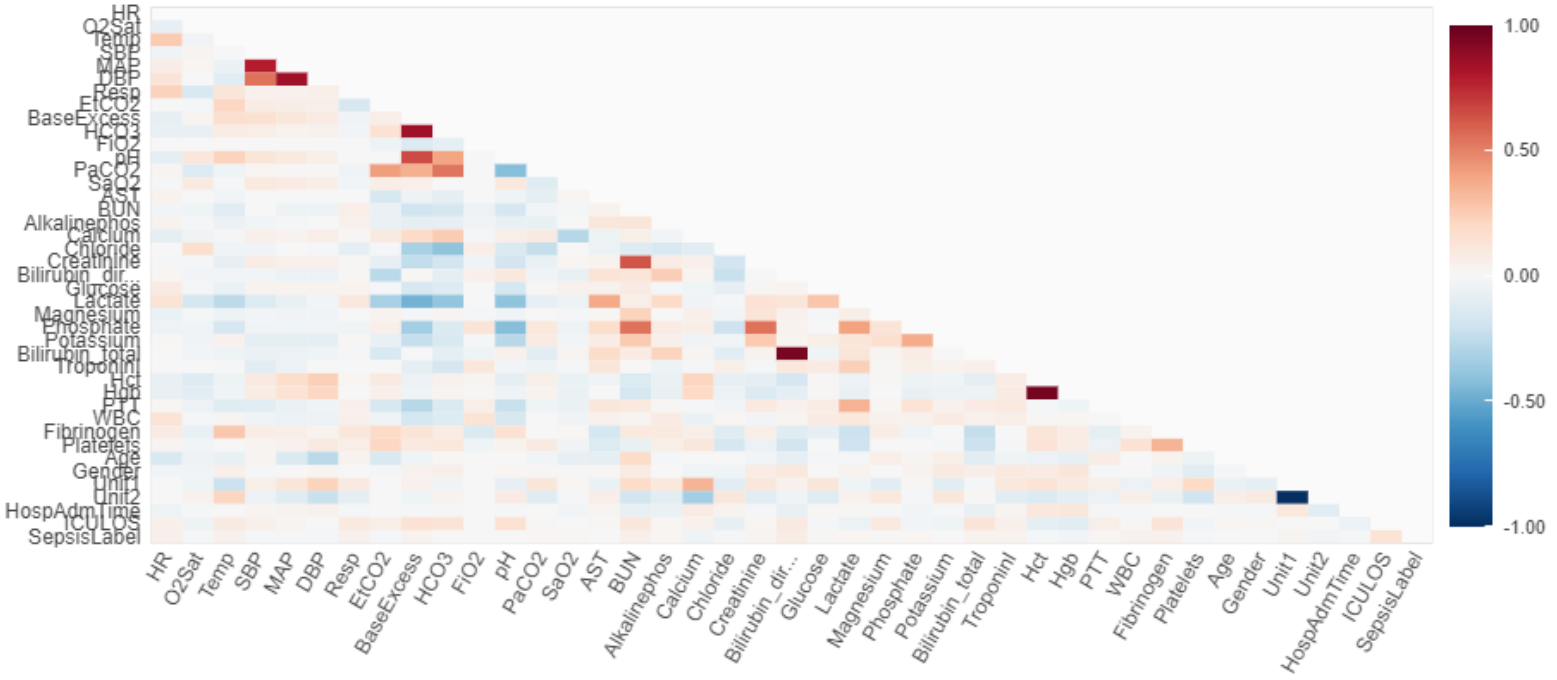
Although we found a strong model to handle sepsis label prediction, there is still ample opportunity for enhancement. In critical classification tasks, we should aim for high recall to avoid making false negative classifications. Our future research will focus on addressing this issue. Despite the inability of deep learning to offer a solution in our study, we plan to gather more data and explore cutting-edge techniques such as CNN and Transformers.

There is still room for improvement in handling missing values, and we opted for simple but effective methods due to time constraints. However, we are eager to explore more advanced imputation techniques such as MICE and Doubly Robust.

Throughout the project, we worked in an asynchronized workflow, team working simultaneously on different topics and combining the knowledge accumulated. We chose to emphasize common best practice teamwork tools, learning to work correctly with Git and GitHub. Thank you for this opportunity. We learnt a lot.

Appendix 1

Correlation Heat Map of Raw Data

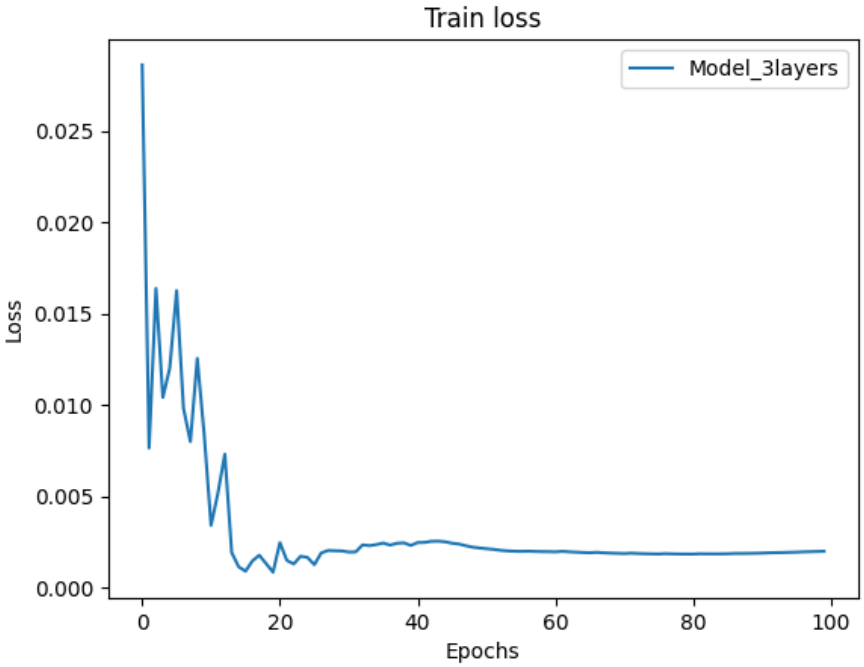


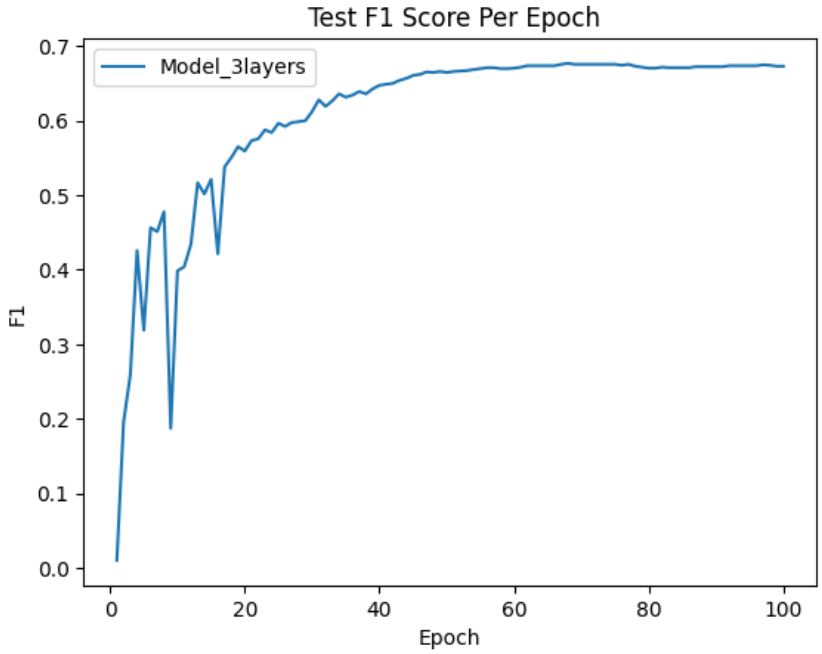
Appendix 2

LSTM Performance Plots

Train loss and F1 score on the test data as a function of Epoch (over 100 epochs)

*** See git of F1 score as a function of Epoch in the git repo “LSTM.gif” ***





Appendix 3

Feature interpretability tables – Top 25 features (out of 80, 40 mean, 40 std)

AdaBoost

Feature	Importance
HR_std	0.041
O2Sat_std	0.034
Creatinine_mean	0.031
MAP_std	0.031
MAP_mean	0.028
Resp_std	0.027
HR_mean	0.027
ICULOS_mean	0.026
SBP_std	0.025
Glucose_mean	0.023
Resp_mean	0.022
O2Sat_mean	0.022
SBP_mean	0.021
Hgb_mean	0.021
WBC_std	0.02
Potassium_std	0.02
DBP_std	0.02
Hct_mean	0.02
ICULOS_std	0.02
BUN_mean	0.02
Calcium_mean	0.019
Temp_mean	0.019
HospAdmTime_mean	0.019
Potassium_mean	0.019
Platelets_mean	0.018

RandomForest

Feature	Importance
ICULOS_std	0.18236
ICULOS_mean	0.15739
SBP_std	0.03542
Temp_std	0.02668
HR_std	0.02547
Temp_mean	0.02527
SBP_mean	0.025
HR_mean	0.02059
Resp_mean	0.02047
O2Sat_mean	0.02027
O2Sat_std	0.01952
Resp_std	0.01848
MAP_std	0.01699
MAP_mean	0.01456
HospAdmTime_mean	0.01275
WBC_mean	0.01248
Creatinine_mean	0.01201
Platelets_mean	0.01196
Glucose_mean	0.01128
Potassium_mean	0.01114
Hct_mean	0.01106
BUN_mean	0.01096
Hgb_mean	0.01074
DBP_std	0.01072
Age_mean	0.01005

XGBoost

Feature	Importance
ICULOS_std	0.16093
Unit2_mean	0.04593
Bilirubin_direct_mean	0.03875
AST_std	0.03818
Fibrinogen_std	0.03695
ICULOS_mean	0.03117
HCO3_mean	0.02225
Lactate_std	0.02163
EtCO2_std	0.01974
Lactate_mean	0.01839
TroponinI_std	0.01809
Fibrinogen_mean	0.01699
FiO2_mean	0.01644
PaCO2_std	0.0164
BaseExcess_mean	0.01603
Bilirubin_total_std	0.01466
TroponinI_mean	0.01372
PTT_std	0.01293
Temp_mean	0.0126
pH_mean	0.01204
pH_std	0.0117
Alkalinephos_std	0.01151
EtCO2_mean	0.01116
Bilirubin_total_mean	0.01101
Chloride_std	0.011