Augmenting SOFA Scores using Machine Learning, Feature Augmentation, and Explainable AI

Dylan Sain, Kathleen Fort, Maria Jorge

July 2024

1 Introduction

This paper will utilize the BOLD dataset in order to predict the 24 hour deterioration or improvement of a patient. It includes investigation of important features through feature selection and XAI. Especially, it utilizes three different machine learning models and various different feature selection tools to bring about a new understanding of SOFA scores and the prediction of patients status after 24 hours. Additionally, it will investigate the potential applications of AI in sepsis prediction, the challenges associated with implementing ML in a clinical setting, and areas where further research is required. Furthermore, this review will explore the gap between ML outputs and user comprehension, emphasizing considerations for creating an effective AI user interface (UI) for clinicians. Specifically, it will address Explainable AI (XAI), strategies to avoid screen fatigue, and the overall usability of the UI. Pulse oximeters are small clip-like machines that typically go on a patient's finger and measure a patient's arterial oxygen saturation (SpO2) levels and heart rate. Pulse oximeters clip onto a patient's finger and shines a light which helps obtain SpO2 and heart rate readings. The more HgB that is saturated with oxygen, the higher the SpO_2 readings should be, which normally read above 95%. Using pulse oximeters is the fastest less invasive way to measure a patient's SpO2 levels. However, decreased accuracy of pulse oximetry in patients with dark skin tones has been demonstrated since as early as 1985. It is seen how, pulse oximeters may overestimate the true oxygen saturation in individuals with dark skin tones, leading to higher rates of occult hypoxemia (unrecognized low blood oxygen saturation).

1.1 Sepsis

Sepsis is one of the leading causes of death world wide, however the first problem many prediction algorithms need to topple before getting to the data or the models themselves is its definition. The third international consensus for sepsis [42], sepsis-3, attempted to re-define the phenomenon in 2016. The definition

they came to accept at this consensus was simply the immune system causing damage to its own body[24]. While easy to define in words, it is much more difficult to translate this definition into some way a computer can understand and learn. For example, many of the early symptoms are extremely common in an ICU setting and can get overlooked before it gets worse. While worse systems make diagnosing sepsis easier, a patient's mortality rate also increases the longer it does untreated. Being an arbitrary condition, the exact moment in time a patient transitions from non-sepsis to sepsis is extremely difficult to point out, even for an experienced doctor[14]. Together these reasons make it difficult for a computer to prediction the condition and thus these are the first problems a machine learning algorithm must face before moving forward.

Many, if not all, boil sepsis down into a binary classification[26]. Based on a series of time series data where a patient has been pre-determined to have either developed sepsis or not are given a binary flag based on this pre-determined data. While simplistic this approach allows machine learning researchers the freedom to reduce complexity in their models. Unfortunately, due to the complex nature of sepsis, most of the finer details of this condition are lost. Clinicians need more than a simple yes or no in order to effectively treat a patient. Furthermore, the definition of sepsis has changed over the years and as it continues to get redefined the previously created machine learning algorithms will also need to get retrained. A fact that very few are taking into account at this time.

1.2 SOFA

SOFA, sequential organ failure assessment, scores are used to quantify the severity of illness daily in an ICU setting. SOFA was designed to be a simple and generalized system in order for clinicians to qunatify the status of a patient. It is based on the assessment of the degree of dysfunction of six vital organ systems: respiration, cardiovascular, central nervous system (CNS), coagulation, liver, and renal. With the number of component scores limited to six, it is less complicated than many earlier severity of illness scores. While a simple system SOFA scores are able to help clinicians with patient status and prevent sepsis.

1.3 Data Pre-processing

The first step, when creating an early prediction of sepsis tool, is creating or procuring a sufficiently large enough, complete, and bias free dataset to work off of. However, especially in the medical field where data privacy is especially important datasets are riddled with missing values, large positive and negative disparities, biases and more. A projects success and failure can be decided within the dataset. In this case all of the papers reviewed used time-series data and also had to create an algorithm that could evaluated this data.

First problem many of these papers hit was how to deal with the large amounts of missing data. Most of the datasets come from different hospitals or region and each one will have a slightly different way of treating patients. This leads to many labs present in one patient, but absent in another. Especially

when it comes to time-series data where labs can only be order every 4 hours yet others are ordered every 24 [22]. Some of the papers used simple yet effective approaches like dropping columns all together with missing values more than 20% [25][29]. More sophisticated approaches include involving filling in values with the most recent value [26][35][22]. This is called linear interpolation, and while this theoretically makes sense for values close in time, it fails for values farther apart in time. Others suggest utilizing missing values as there could be a pattern data scientist cannot obtain within the values [26][14][34]. Some even utilized methods like Gaussian Processes [14] or models like XGBoost [29] to make missing values have less impact on the model.

While missing values are a large problem, there was only one paper that touched upon another important problem in machine learning, the disproportion between the 'positive' and 'negative' values [26]. Due to the nature of sepsis, many data sets are hit with a very particular problem: there are more cases of non-sepsis patients to sepsis patients. Inadvertently this causes the machine learning model to be more biased towards non-sepsis patients, a kind of disparity that causes problems for many models.

Another issue that is brought up in all papers is feature selection. Some claim that good feature selection [22] with a less complicated model can beat out neural networks any day. Many of the papers utilized various tests such as ANOVA, KS(Kolmogorov-Smirnov), Mann Whitney U, Chi-squared, Fisher's exact, and more to determine the exact amount of features to be used within the model [26][46][25][22][29][5][24]. The lack of these tests in other papers [14][35][34] shows that the authors trusted their models to determine the features with the most information gain without having to do any prior work. This approach is popular within neural networks field as it allows a researcher to fully utilize the learning power of the neural network.

1.4 Methods

Various methods were utilized in the creation of these methods. Specifically the ones that seem to be dominating the machine learning world are Random Forests, XGBoost classifiers and LSTM RNNs.

Random Forest classifiers seem to have good results for a handful of reasons. They rely heavily on statistical analysis of the features for good feature selection[26] [46] [25]. This allows a machine learning researcher to carefully pick out the features that extract the least amount of bias, troubling data with missing values and even specify various features that are known by professionals to be important for predicting sepsis. Due to their natural bagging capabilities much of the variance in the data is reduced. Other approaches like XGBoosting algorithms shine because of their innate ability to handle missing values with less problems than the random forests [22] [29]. This allows XGBoost algorithms to rely less on imputation strategies talked above and more on the data itself. This would greatly reduce bias introduced by the imputation strategies. Finally RNNs and other neural network approaches tend to be the best when it comes to accuracy [14] [35] [34]. The ability to handle large amounts of data with

ease and pick up patterns doctors cannot might allow one to find patterns and features that other models cannot. They are also equipped with many different techniques and ability to adapt through transfer learning for various hospitals across the globe.

1.5 Evaluation

While the methods to get to a prediction can be quite different, almost all of these papers rely on the same metrics for evaluation. The favorite for these models is the AUC (area under the ROC curve)[14]. Sense these kinds of predictions cannot be boiled down to a simple percentage based accuracy, a more sophisticated way of evaluating the models must be found. Utilizing the ROC curve a researcher can visualize the accuracy, false negatives and false positives on the same curve. This also allows the model to be adjusted based on the criteria being fit.

One of the biggest problems in the medical world is a phenomenon called alarm fatigue[14][35][24]. According to one paper 63.4% of alerts were cancelled by the nurses who received them[14]. A good model cannot be the boy who called wolf. Ensuring that a model does not over estimate the prediction leading to a high level of false positives is paramount when developing a model. Researchers can use various techniques such as maximizing F1 score instead of accuracy, or analyzing the confusion matrix and adjusting accordingly to ensure their models are not attributing towards alarm fatigue.

Another thing that is not talked about enough was the affect in fairness [5] in these models. Due to the nature of these datasets the models were trained on majority white people. People of color were found to be under represented. One study found that there is a positive correlation between fairness and better predictability overall. Naturally ensuring these models can accurately predict for all people, age, race, gender, social status, ect. must be touched upon, but is not nearly talked about enough in the mainstream papers.

1.6 Medical World and Black Blox Models

Larger machine learning models have been dubbed black box models by the majority of the statistics community. Data goes in, magic happens, and a value pops out, with the majority of the time the magic is either extremely hard or outright impossible to explain what happened. Interpretability is extremely important in the medical field where doctors need to be making decisions on a patient's life. Not nearly enough of the machine learning community are approaching this problem with the intention of ensuring a clinician understands the reasoning behind a prediction[35][24].

Models like random forests are perfect for interpretability. While graphs like importance plots can explain which of the features are more important to the decision making than others, the natural decision process for random forests allows a clinician insight into directly the weight a model places upon a certain lab or data point for its decision. On the other side, XAI methods

for neural networks are much more complicated and difficult to work with. XAI methods for neural networks are both difficult to implement correctly and difficult to verify leading to an inherit distrust of the models due to their Black Box nature. Overall medical professionals need to be able to trust the outputs of these models. One paper suggests instead of a decision system, their model is an early warning system with physicians still making the final call[35]. The majority of these models and papers are taking data out of an academic setting without much thought about how a physician might use it in a realistic hospital setting.

2 Pulse Oximeters and Potential Bias

Pulse oximeters are small clip-like machines that typically go on a patient's finger and measure a patient's arterial oxygen saturation (SpO2) levels and heart rate. A typical pulse oximeter has two light-emitting diodes (LEDs), one emitting red light and the other emitting infrared light. Oxygen saturation is defined as the measurement of the amount of oxygen dissolved in blood, based on the detection of Hb and HbO2. Using pulse oximeters is the fastest less invasive way to measure a patient's SpO2 levels. However, decreased accuracy of pulse oximetry in patients with dark skin tones has been demonstrated since as early as 1985. It is seen how, pulse oximeters may overestimate the true oxygen saturation in individuals with dark skin tones, leading to higher rates of occult hypoxemia (unrecognized low blood oxygen saturation). This is due to the high levels of melanin in their skin. The shielding effect of melanin has against UV light, the natural substance produced by specialized cells that determine the color of hair, eyes and skin, proves difficult for devices such as pulse oximeters to get an accurate reading.

Recent studies have linked pulse oximeter inaccuracy to worse clinical outcomes, suggesting that pulse oximeter inaccuracy contributes to known racial health disparities. Failure to control increased absorption of red light by melanin and inadequate regulatory standards for device approval play major roles in decreased accuracy.

2.1 Pulse Oximetry Analysis

When analyzing pulse oximeters, it was seen that the true faults lay in the device design. Many engineers overlooked the fact that the melanin found in patients that have darker skin tones would absorb most of the applied UV light and not give an accurate reading.

Due to this known bias, the BOLD data needed to be analyzed and corrected for any of the foreseen bias. Utelizing a kernel density estimation plot, it was seen that most of the patients (Caucasian and minority patients) were within in healthy SpO2 ranges (see figures 1 and 2).

To be able to compare both of the graphs easier, both graphs were overlapped one on top of each other. The Caucasian patients' data are represented in the

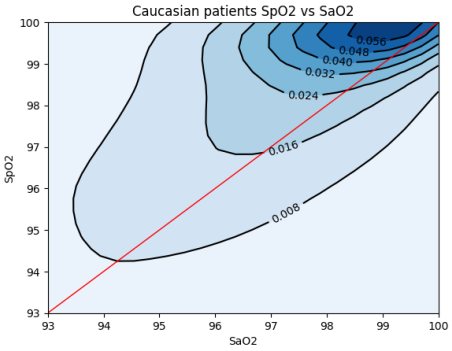

Figure 1: Density plot - Caucasian patients SpO2 vs SaO2

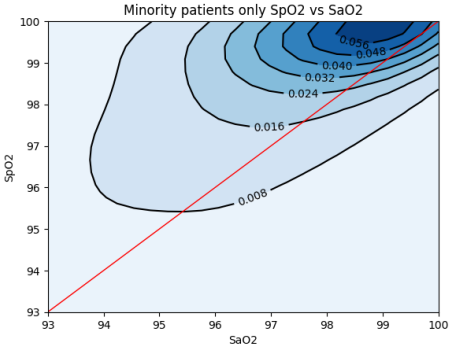

Figure 2: Density plot - Minority patients SpO2 vs SaO2

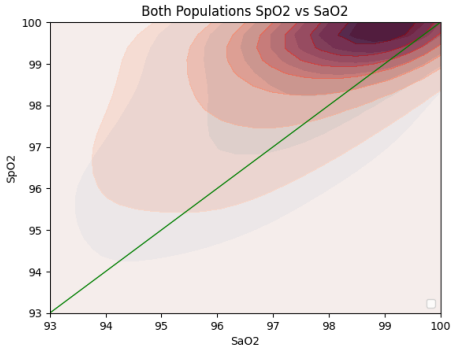

Figure 3: Density plot - Minority patients SpO2 vs SaO2

color blue and the minority patients are represented in the color red (see figure 3).

2.2 Pulse Oximetry Dissucssion

In order for a sepsis patient to be deemed as healthy, their SpO2 ranges need to be in the 90%-100% range. The data within the BOLD dataset had a majority within this acceptable range. This is also where the bias in readings existed. The unhealthy range, useful for the 24 hour predictions, did not exhibit the same bias. Dude to this distinction a corrected algorithm was not implemented for this project.

3 Data Preparation

Due to the large amount of research in sepsis prediction requiring extensive time series data to make a sepsis prediction this paper decided to tackle another approach. Instead of using a large section of time series data to predict whether or not a patient will develop sepsis, this paper attempted to see if a patient will deteriorate or improve 24 hours after admission to the ICU. The BOLD dataset was perfect for this, a combination of three datasets: MIMIC-III, MIMIC-IV, and eICU-CRD, it allowed for a larger variety in the data. Due to the tabular nature of this dataset, with values being taken immediately as a patient got into the ICU, it provided a unique opportunity to approach sepsis prediction from a different angle.

However, as with many medical datasets, and due to the combining of multiple different datasets, there was an extensive amount of missing data within BOLD (see figure 4). Specific patients were removed during initial ispecition of the data due to missing critical prediction values. This drooped the dataset from $49,000$ to approximately $30,000$. Furthermore, various patients were dropped due to large amounts of missing values within that row. Any row with over 25%

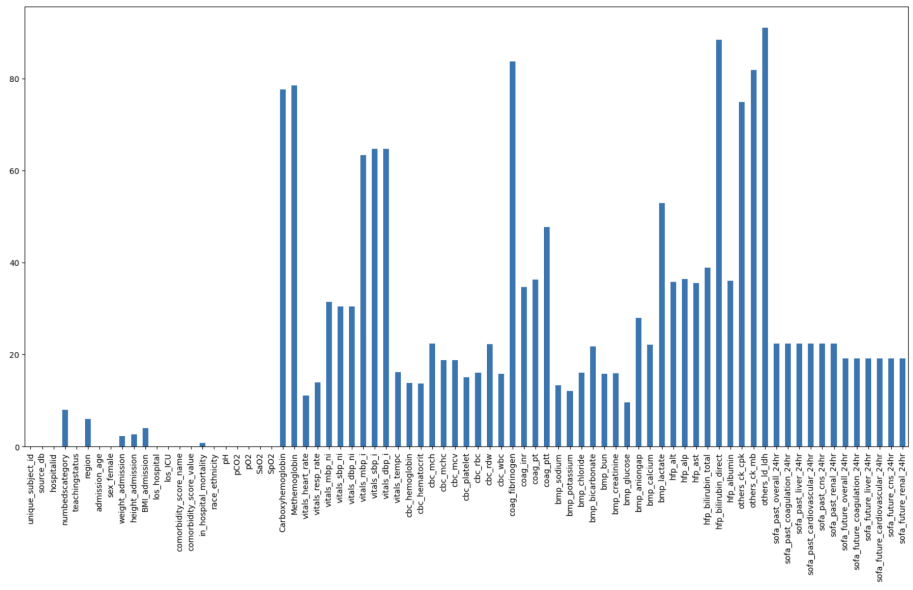

Figure 4: A figure containing all features in BOLD on the X axis and the percentage (%) of missing values on the Y axis

of its values missing were dropped. However, there was still a significant amount of missing values. Due to the significant missing values various imputation techniques, from linear interpolation to moving averages or a full KNN imputation, were tested for their effectiveness. The most efficient technique that was tested utilized a combination of computer science imputation algorithms and clinical insight into the data. Using knowledge of the various labs and how they affect each other, and the KNN imputation technique the missing data was effectively filled.

- Hepatic function panel (HFP) $\rightarrow$ Coagulation labs and patient's age
- Vitals $\rightarrow$ patient's age and SOFA past cardiovascular
- Basic Metabolic Panel (BMP) $\rightarrow$ HFP and patient's age and SOFA past renal
- Complete Blood Count (CBC) $\rightarrow$ COAG and patient's age
- Coagulation labs $\rightarrow$ HFP and CBC and patient's age

Its important to note that using the smaller KNN subsets were important to reduce bias within the data. Using smaller subsets instead of the full dataset allowed the previous knowledge of the human body and how one field might affect the other. For example, there is a significant negative correlation between the

measured hemoglobin and the measured fibrinogen, however they exist in two different lab subsets: CBC and COAG respectively. Armed with this knowledge the KNN imputation strategy is able to accurately predict the missing values based on the various labs that had been done.

Additionally there were various columns that were deemed unnecessary for the model or would provide the model with an unfair advantage. Columns such as Methemoglobin and Carboxyhemoglobin were deemed to have high levels of missing values and the values themselves were only ordered by specific hospitals, thereby introducing a potential bias between hospitals that order these test and ones that do not. Thus some columns were removed for these reasons. Others such as hospital ID created a model that was learning to separate patients by hospital first, before learning the 24 hour prediction. Adjusting the comorbidity score by using a min/max scaling was necessary as the dataset included two different scales, Charlson and Elixhauser. One-hot encoding was used on gender and other text based columns.

4 Methods

4.1 Developing a Machine Learning Model

Once the data was pre-processed and ready, various experiments were put into place. Based on the tabular data, the binary classification class and previous work on the subject three different models were used to predict the status of the patient after 24 hours, an XGBoost model, a Random Forest classifier, and a simple multi-layer Perceptron. Through investigation these models have the innate ability to analyze tabular data as well as have good results in classification tasks. Ideally the XGBoost or the Random Forest model to be the selected model due to their higher levels of explainability.

Optuna trials were run for hyperparameter tuning on a (60, 20, 20) train/test/validation split. The bayesian technique that the Optuna trials had allowed for a smaller number of models to be trained, while increasing accuracy over time by narrowing down the ideal criteria. Regularization and early stopping were used in the XGBoost and MLP models respectively, to prevent the overfitting of the data. The XGBoost model was tuned for number of estimators, max depth of the trees, learning rate, subsample, column sample by tree, gamma, regularization alpha, and regularization lambda. The Random Forest model was tuned for number of estimators, max depth, minimum sample spilt, and minimum samples per leaf. The MLP was tuned for first layer neurons, second layer neurons, learning rate, activation functions, solver, and an alpha value. Each of the trials were set to optimize the log-loss of the validation set. A graph of the training can be seen in figure 5.

Extensive work was also put into feature engineering. While the optuna trials utilized all 62 features, reducing that number to a much more feasible number was important. Using a technique called recursive feature elimination (RFE) and feature importance plots, a ranking was created of the top features based on

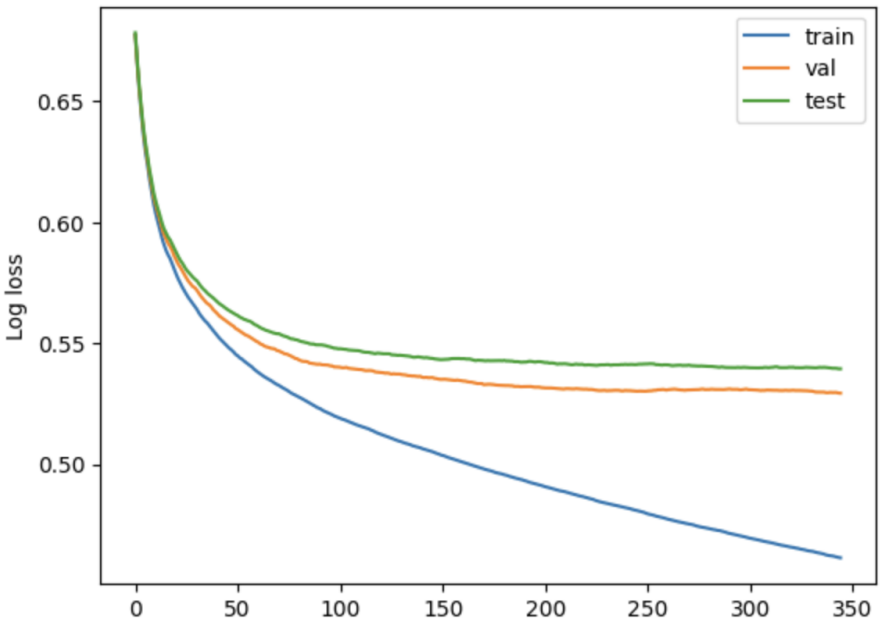

Figure 5: A figure representing the log loss of a XGBoost model over the training process.

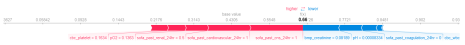

Figure 6: SHAP Local Explainability for a patient x

how often they would be chosen for predictions, and their contributions towards a model's F1 score. This ranking was used to both choose the top features to include in the final model and also as a benchmark to potentially augment the SOFA scores with additional labs that the model found important.

4.2 Ensuring Fairness

To be able to enforce accuracy in the machine learning model, a fairness model needs to be integrated into the program. The chosen fairness model is called Fairlearn.

In a medical setting demographic parity, which aims to ensure that the proportion of positive outcomes is the same across different demographic groups, isn't commonly used. Fairlearn however uses a special package called Equalized Metrics. In the package the model will use Equalized Opportunity and Equalized odds. Equalized Opportunity is when a model's true positive rate is equal across different groups. It makes sure that individuals who belong to different demographic groups have an equal chance of receiving a positive outcome if they truly qualify for it. Equalized Odds requires that both the true positive rate (TPR) and the false positive rate (FPR) be equal across different groups. This ensures that the model's predictions are equally accurate for all demographic groups. In machine learning, particularly in binary classification, FPR (False Positive Rate) and TPR (True Positive Rate) are critical metrics used to evaluate a classification model's performance. TPR, also known as Sensitivity or Recall, is the proportion of positives correctly identified by the model. A high TPR indicates that the model is good at identifying positive instances. FPR is the proportion of negatives incorrectly identified as positive by the model. FPR measures how often the model incorrectly identifies negative instances as positive. A low FPR indicates that the model makes fewer false positive errors.

4.3 Explainable AI Methods

Shapley Additive Explanations (SHAP) was employed to achieve local explainability in the predictive model. This process was important to ensure the clinicians can visualize how the data is being analyzed on a patient by patient basis. The SHAP explainer analyzes the model using interventional feature perturbation, which marginalizes out feature values to reflect interventional probabilities. This method preserves relationships between features and avoids generating unrealistic data points. SHAP values were derived from the vital signs data collected from users, and an expected value was extracted to serve as a baseline risk for sepsis development. This baseline is crucial for comparing individual patient risks (see Figure 6).

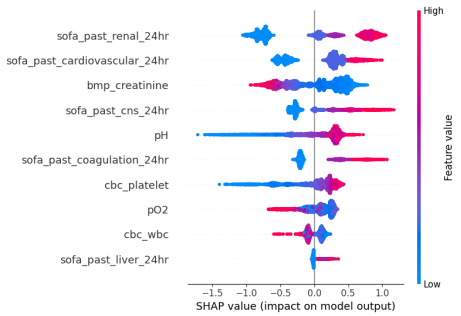

Figure 7: SHAP Global Explainability for the machine learning model

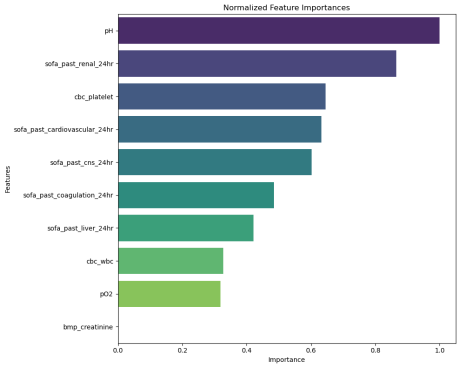

Figure 8: LIME Global Explainability for the machine learning model

In addition to SHAP, Permutation Feature Importance (PFI) and Local Interpretable Model-Agnostic Explanations (LIME) were used to achieve global model explainability (see Figures 7, 8, and 9). Although SHAP consistently outperformed other methods, employing a variety of techniques allowed us to identify consistent patterns in feature importance across different explainability methods.

4.4 User Interface Development

An asynchronous web interface was developed to handle Plotly visualizations. User information is collected through a Django-based form, which saves the data to a Django database. This data is then passed into the predictive model, and the predicted output is shown to the user as well as the SHAP values and the expected value (baseline risk).

A force plot is created with the SHAP values, scaled to the prediction range (0 to 1). A context dictionary containing the prediction, SHAP values, vitals data, and page title is also shown.

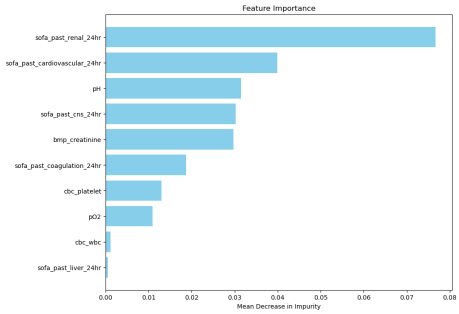

Figure 9: PFI Global Explainability for the machine learning model

	XGBoost	Random Forest	Simple MLP
Train	78.6	84.2	72.2
Validation	72.2	70.7	71.1
Test	72.9	71.3	71.2

Table 1: Reported F1 Score for the various full models

4.5 Model Enhancement and Data Integration

From the RFE experiments a handful of extra features including white blood cell count, partial pressure of oxygen (pO2) levels, and blood pH measurements were integrated into the final model. The web interface also integrated raw data on creatinine and platelet count levels alongside the Sequential Organ Failure Assessment (SOFA) categories for renal and coagulation scores.

4.6 User Interface Data Collection

The user interface is designed to collect patient demographic information, specifically race or ethnicity, and gender, exclusively for retrospective bias and performance analysis. These demographic variables are not utilized in the predictive modeling process. This is potentially due to limits within the dataset.

5 Results

5.1 Model Performance and Feature Selection

With the top models and their hyperparameters chosen from the optuna trials, they were finally tested against the test set. The F1 scores, shown in Table 1, show that the models were able to get F1 scores just above 70%. Note as well that both the XGBoost and the Random Forest model had significantly higher F1 scores on the training set. This is due to a possible overfit of the train set. Despite a large difference between positive (deterioration) and negative

Feature Name	Average Selected	Average Rank
PH	1.0	1.0
pO2	1.0	1.0
cbc_platelet	1.0	1.0
bmp_creatinine	0.96	1.04
bmp_wbc	0.96	1.16
bmp_lactate	0.88	1.20
coag_ptt	0.88	1.20
vitals_sbp_ni	0.84	1.44
coag_pt	0.76	2.16
weight_admission	0.72	2.12

Table 2: Average selected and average score for various features using RFE

(improvement) classes a confusion matrix did not show any significant difference in the model choosing one class over the other.

The feature importance was also instrument in finding a final model. Using the RFE results and sorting them by most and least important, see a sample of features in table 2, the top features were decided. Using this table various XGBoost models were trained on increasing number of features. This culminated in figure 10, where a F1 score of almost 70% could be achieved with only 5-10 features. After this jump in accuracy, the model tapered off at around 70%. This is probably due to the added features adding very little in terms of information gain for the model to use. This allows the creation of a model with a minimal number of features for a clinician to input before getting a fairly accurate prediction of a patient's status over 24 hours.

5.2 Fairness

When applying the Fairlearn Equalized metrics package to analyze the results of the model, a table was generated to easily look at the accuracy of the results. Figure 11 and figure 12 show the equalized odds accuracy results and the exponential gradient metrics respectively.

5.3 User Interface Output

Based on the RFE results and the model performance analysis revealed that integrating the SOFA score subdivisions and additional laboratory data significantly enhanced predictive accuracy for sepsis. The user interface operates by collecting this comprehensive set of data, which includes the SOFA score subdivisions and critical raw laboratory data. This data is processed through the model, and the user is redirected to a results page. The results page displays the model's output, an interpretation of the output (indicating whether the patient's condition will likely improve or deteriorate), and a SHAP visualization of the patient-specific output.

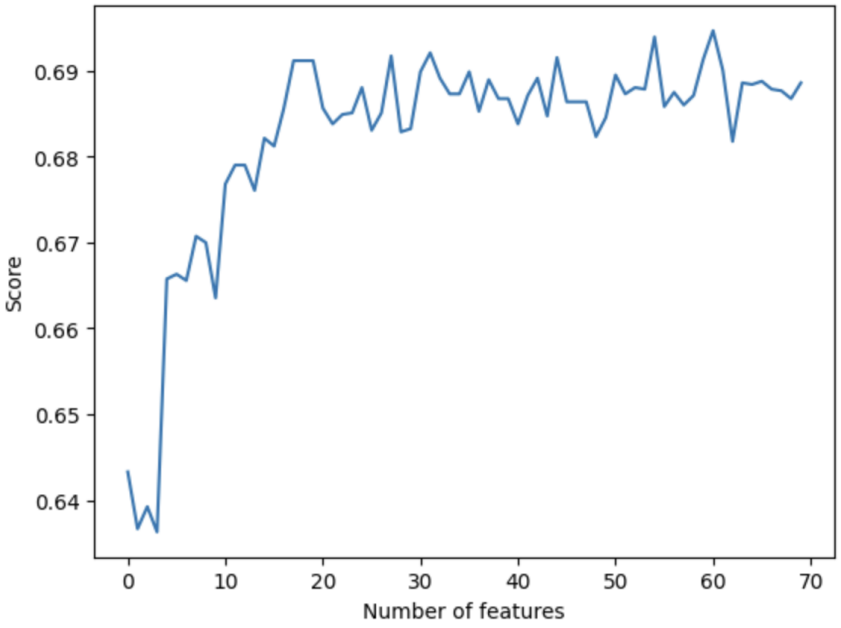

Figure 10: A comparison of a XGBoost model with increasing number of features based on the RFE results.

		accuracy		fpr		tpr	
		train	test	train	test	train	test
race_ethnicity	sex_female						
0	0	0.823529	0.736842	0.047619	0.500000	0.615385	0.909091
	1	0.840000	0.850000	0.129032	0.100000	0.789474	0.800000
1	0	0.797619	0.777778	0.174419	0.208333	0.768293	0.766667
	1	0.794118	0.704545	0.102941	0.217391	0.691176	0.619048
2	0	0.811804	0.711864	0.184154	0.262411	0.807425	0.688312
	1	0.814010	0.721429	0.170507	0.195489	0.796954	0.646259
3	0	0.824468	0.750000	0.217617	0.258621	0.868852	0.758621
	1	0.810398	0.777778	0.250000	0.333333	0.868263	0.855072
6	0	0.824658	0.710407	0.206718	0.288136	0.860058	0.708738
	1	0.821218	0.697183	0.203187	0.317073	0.844961	0.716667
7	0	0.822075	0.727032	0.185868	0.280672	0.830633	0.735568
	1	0.802806	0.714823	0.210209	0.275132	0.816956	0.704471

```
equalized odds (test): 0.40
accuracy (test): 0.72
```

Figure 11: Equalized odds accuracy results

		accuracy		fpr		tpr	
		train	test	train	test	train	test
race_ethnicity	sex_female						
0	0	0.852941	0.736842	0.190476	0.500000	0.923077	0.909091
	1	0.900000	0.700000	0.096774	0.300000	0.894737	0.700000
1	0	0.815476	0.796296	0.197674	0.166667	0.829268	0.766667
	1	0.867647	0.681818	0.147059	0.260870	0.882353	0.619048
2	0	0.802895	0.718644	0.207709	0.269504	0.814385	0.707792
	1	0.803140	0.739286	0.193548	0.203008	0.799492	0.687075
3	0	0.821809	0.767241	0.191710	0.241379	0.836066	0.775862
	1	0.816514	0.786325	0.225000	0.333333	0.856287	0.869565
6	0	0.815068	0.705882	0.204134	0.305085	0.836735	0.718447
	1	0.813360	0.697183	0.187251	0.317073	0.813953	0.716667
7	0	0.818690	0.727915	0.186152	0.282353	0.823907	0.739292
	1	0.805761	0.715360	0.204537	0.280423	0.816956	0.711014

```
equalized odds (test): 0.33
accuracy (test): 0.72
```

Figure 12: Exponential gradient metrics results

The global XAI results are also made accessible through the user interface, providing users with a comprehensive understanding of feature importance and model predictions.

5.4 Explainable AI Results

Utilizing three different methods of global Explainable AI (XAI) — Shapley Additive Explanations (SHAP), Permutation Feature Importance (PFI), and Local Interpretable Model-Agnostic Explanations (LIME) — allowed us to identify consistent patterns in feature importance. Across all three methods, the patient's SOFA renal score, SOFA cardiovascular score, and pH level were consistently ranked as highly important features. Conversely, SOFA liver scores, white blood cell count, and pO2 levels consistently ranked as less important.

5.5 Discrepancies in Explainable AI Results

Some discrepancies were observed in the explainable AI outputs that can be attributed to data imperfections. For instance, in the local AI analyses, creatinine levels often appeared as a protective feature while renal scores were highlighted as concerning factors. This is contradictory since renal scores are directly influenced by the patient's creatinine levels. These inconsistencies point to underlying data issues that need further investigation to improve model reliability and interpretation accuracy.

6 Discussion and Conclusion

The results of the machine learning models were all around 70% accuracy. While this is good for an initial approach to the problem it did not hit the goal of over 80% accuracy for it to be appropriate for use in the ICU. That being said the

majority of models that work in sepsis prediction utilize hours if not days of data while this paper's models used only a snapshot in time. Overall the results of the models can be used as a tool for clinicians to use, but should not be used as a golden rule for a patients status after 24 hours.

Poor accuracy can be attributed to a number of possible factors. XGBoost and random forest, while good baseline machine learning models, are rather simplistic when compared to large neural networks. The MLP that was used in this paper, while is a neural network, was only comprised of two hidden layers and a small number of parameters. Its possible that a larger network and more sophisticated activation functions might produce better results. However it is possible that the poor accuracy could be attributed directly to the data gathered. Even with the data from three different datasets, certain biases existed in the data, such as the majority of the people being from the ages 50 and above. This can cause some issues with the prediction. Another possibility might be from the problem itself. Due to the nature of a snapshot in time, a lot of information about the status of the patient is missing. Due to this it might be impossible with the current tools and data to get predictions into the 95% or above.

The results from the RFE and XAI plots show how the model itself ignores certain aspects that clinicians thought crucial while emphasising others not included in the SOFA scores. Currently, all subdivisions of the Sequential Organ Failure Assessment (SOFA) scores are weighted equally, each receiving a score from 0 to 4. However, by analyzing the Explainable AI (XAI) model results, clinicians can prioritize more critical features over less significant ones. Values such as PH and PO2 are both highly important for the accuracy of the models, and yet they are not included in the SOFA scores at all. Additional a patient's renal score is ranked higher than its liver performance, and yet SOFA ranks them all equally. It's possible that the model is using information that is critical for prediction a patient's status after 24 hours without clinicians knowing their importance. This information could be used to help clinicians make critical decisions in the future.

Local explainable AI results are particularly useful for clinicians as they provide detailed insights into individual patient factors. By examining local XAI, clinicians can identify which factors are concerning and which are protective. This information enables clinicians to tailor treatment management plans based on the specific risks identified for each patient. For example, if a patient is identified as high risk for deterioration, the clinician can increase the frequency of tests for that patient. Moreover, clinicians can pinpoint areas of concern and take proactive measures to mitigate risks or maintain closer observation. Potential management strategies include consulting specialists for the subdivision, ordering more frequent tests, or administering antibiotics or other medications proactively.

When measuring equalized odds directly, the model is simply assessing the current state of the model's fairness. In contrast, using the exponentiated gradient method actively modifies the model to satisfy the equalized odds constraints, which can lead to changes in both True Positive Rate (TPR) and False Posi-

tive Rate (FPR) across groups. The equalized odds test result for the original model is 0.40. However, the result for the exponentiated gradient method is 0.33. This discrepancy (0.40 vs. 0.33) indicates that the exponentiated gradient method successfully reduced the disparity between groups compared to the original model, leading to a lower equalized odds metric. Comparing the FPR and TPR results for both metrics, the exponential gradient method shows higher results in both FPR and TPR tests. The higher FPR in the exponential gradient metrics may indicate that the algorithm increased FPR for some groups to balance the TPR, leading to more equalized outcomes across groups. This is a common trade-off when trying to satisfy fairness constraints. The differences in results between direct equalized odds metrics and the exponentiated gradient method are due to the active intervention of the latter to satisfy fairness constraints, leading to adjusted TPR and FPR rates across groups. Understanding these trade-offs is crucial for interpreting fairness outcomes and making informed decisions about deploying fair machine learning models.

In general, working on more extensive models as well as making a generalized model are both things that should be included in future work. Making a larger MLP and exploring other neural network or other more traditional machine learning or statistical approaches to the problem should be explored. A larger or more sophisticated model might be able to pick out patterns or small differences between the patients that the more simplistic models did not. Exploring more of the RFE and XAI results is also crucial to understanding the problem. While this many require a clinician's input, figuring out if there is a reason why values like PH are important while others are not should be of great interest to future work done in this project. Its possible SOFA and other scores are missing, overlooking, or highlighting the wrong information, and machine learning techniques like these might be able to help uncover this information.

Future steps for this project include conducting a thorough bias analysis to ensure the model's fairness and reliability. Additionally, the user interface needs to be tested with clinicians to gather feedback on its usability and practicality in a clinical setting. This feedback will be crucial for refining the interface and ensuring it meets the needs of healthcare professionals if implemented in practice.

7 Contributions

Dylan: Worked on the data analysis with feature selection using RFE. Built and tuned the machine learning models as well as choosing the best working for the final website. Kathleen: Created a User Interface prototype to be used by clinicians to collect patient data to be analyzed by the machine learning model and provide an interpretable output, along with local SHAP XAI plots and global SHAP, LIME, and PFI graphs. Maria: Analyzed the bias found in pulse oximetry results, explained the design flaws, and facilitated the implementation of a fairness metrics package in the machine learning model to interpret the program's results.

References

[1] Julia Amann et al. "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective". In: *BMC Medical Informatics and Decision Making* 20.1 (Nov. 30, 2020), p. 310. ISSN: 1472-6947. DOI: 10.1186/s12911-020-01332-6. URL: https://doi.org/10.1186/s12911-020-01332-6 (visited on 07/22/2024).

[2] Atefeh Baniasadi et al. "Two-Step Imputation and AdaBoost-Based Classification for Early Prediction of Sepsis on Imbalanced Clinical Data". In: *Critical Care Medicine* 49.1 (Jan. 2021), e91. ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000004705. URL: https://journals.lww.com/ccmjournal/abstract/2021/01000/two_step_imputation_and_adaboost_based.28.aspx (visited on 07/22/2024).

[3] Michael Bauer et al. "Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019— results from a systematic review and meta-analysis". In: *Critical Care* 24.1 (May 19, 2020), p. 239. ISSN: 1364-8535. DOI: 10.1186/s13054-020-02950-2. URL: https://doi.org/10.1186/s13054-020-02950-2 (visited on 07/22/2024).

[4] Michaela Brenner and Vincent J. Hearing. "The Protective Role of Melanin Against UV Damage in Human Skin". en. In: *Photochemistry and Photobiology* 84.3 (2008). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-1097.2007.00226.x, pp. 539–549. ISSN: 1751-1097. DOI: 10.1111/j.1751-1097.2007.00226.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-1097.2007.00226.x (visited on 07/22/2024).

[5] Chia-Hsuan Chang, Xiaoyang Wang, and Christopher C. Yang. *Explainable AI for Fair Sepsis Mortality Predictive Model*. 2024. arXiv: 2404.13139 [cs.LG]. URL: https://arxiv.org/abs/2404.13139.

[6] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. arXiv:1603.02754 [cs]. Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. URL: http://arxiv.org/abs/1603.02754 (visited on 07/22/2024).

[7] *complex case of EHRs: examining the factors impacting the EHR user experience — Journal of the American Medical Informatics Association — Oxford Academic*. URL: https://academic.oup.com/jamia/article/26/7/673/5426085 (visited on 07/22/2024).

[8] Glynda Rees Doyle and Jodie Anita McCutcheon. "5.3 Pulse Oximetry". en. In: (Nov. 2015). Book Title: Clinical Procedures for Safer Patient Care Publisher: BCcampus. URL: https://opentextbc.ca/clinicalskills/chapter/5-2-pulse-oximetry-2/ (visited on 07/22/2024).

[9] Sandra Eloranta and Magnus Boman. "Predictive models for clinical decision making: Deep dives in practical machine learning". In: *Journal of Internal Medicine* 292.2 (Aug. 2022), pp. 278–295. ISSN: 0954-6820, 1365-2796. DOI: 10.1111/joim.13483. URL: https://onlinelibrary.wiley.com/doi/10.1111/joim.13483 (visited on 07/22/2024).

[10] Tom Evans. "Diagnosis and management of sepsis". In: *Clinical Medicine* 18.2 (Apr. 1, 2018), pp. 146–149. ISSN: 1470-2118. DOI: 10.7861/clinmedicine.18-2-146. URL: https://www.sciencedirect.com/science/article/pii/S1470211824017354 (visited on 07/22/2024).

[11] *Explainable AI for clinical and remote health applications: a survey on tabular and time series data — Artificial Intelligence Review*. URL: https://link.springer.com/article/10.1007/s10462-022-10304-3 (visited on 07/22/2024).

[12] John R. Feiner, John W. Severinghaus, and Philip E. Bickler. "Dark Skin Decreases the Accuracy of Pulse Oximeters at Low Oxygen Saturation: The Effects of Oximeter Probe Type and Gender". en-US. In: *Anesthesia & Analgesia* 105.6 (Dec. 2007), S18. ISSN: 0003-2999. DOI: 10.1213/01.ane.0000285988.35174.d9. URL: https://journals.lww.com/anesthesia-analgesia/fulltext/2007/12001/dark_skin_decreases_the_accuracy_of_pulse.4.aspx (visited on 07/23/2024).

[13] Lucas M. Fleuren et al. "Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy". In: *Intensive Care Medicine* 46.3 (Mar. 1, 2020), pp. 383–400. ISSN: 1432-1238. DOI: 10.1007/s00134-019-05872-y. URL: https://doi.org/10.1007/s00134-019-05872-y (visited on 07/22/2024).

[14] Joseph Futoma et al. "An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection". In: *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez et al. Vol. 68. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 243–254. URL: https://proceedings.mlr.press/v68/futoma17a.html.

[15] Erich Gnaiger et al. "Control of mitochondrial and cellular respiration by oxygen". en. In: *Journal of Bioenergetics and Biomembranes* 27.6 (Dec. 1995), pp. 583–596. ISSN: 1573-6881. DOI: 10.1007/BF02111656. URL: https://doi.org/10.1007/BF02111656 (visited on 07/23/2024).

[16] YiRan He et al. "A machine-learning approach for prediction of hospital mortality in cancer-related sepsis". In: *Clinical eHealth* 6 (Dec. 1, 2023), pp. 17–23. ISSN: 2588-9141. DOI: 10.1016/j.ceh.2023.06.003. URL: https://www.sciencedirect.com/science/article/pii/S2588914123000126 (visited on 07/22/2024).

[17] Md. Mohaimenul Islam et al. "Prediction of sepsis patients using machine learning approach: A meta-analysis". In: *Computer Methods and Programs in Biomedicine* 170 (Mar. 1, 2019), pp. 1–9. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2018.12.027. URL: https://www.sciencedirect.com/science/article/pii/S016926071831602X (visited on 07/22/2024).

[18] Md. Mohaimenul Islam et al. "Prediction of sepsis patients using machine learning approach: A meta-analysis". In: *Computer Methods and Programs in Biomedicine* 170 (Mar. 1, 2019), pp. 1–9. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2018.12.027. URL: https://www.sciencedirect.com/science/article/pii/S016926071831602X (visited on 07/22/2024).

[19] Haya Jamali et al. "Racial Disparity in Oxygen Saturation Measurements by PulseOximetry: Evidence and Implications". eng. In: *Annals of the American Thoracic Society* 19.12 (Dec. 2022), pp. 1951–1964. ISSN: 2325-6621. DOI: 10.1513/AnnalsATS.202203-270CME.

[20] *Journal of Medical Internet Research - Findings and Guidelines on Provider Technology, Fatigue, and Well-being: Scoping Review*. URL: https://www.jmir.org/2022/5/e34451 (visited on 07/22/2024).

[21] Elizabeth A. Krupinski et al. "Long Radiology Workdays Reduce Detection and Accommodation Accuracy". In: *Journal of the American College of Radiology* 7.9 (Sept. 1, 2010), pp. 698–704. ISSN: 1546-1440. DOI: 10.1016/j.jacr.2010.03.004. URL: https://www.sciencedirect.com/science/article/pii/S1546144010001341 (visited on 07/22/2024).

[22] Shuhui Liu et al. "Dynamic Sepsis Prediction for Intensive Care Unit Patients Using XGBoost-Based Model With Novel Time-Dependent Features". In: *IEEE Journal of Biomedical and Health Informatics* 26.8 (2022), pp. 4258–4269. DOI: 10.1109/JBHI.2022.3171673.

[23] Karen Dunn Lopez and Linda Fahey. "Advocating for Greater Usability in Clinical Technologies: The Role of the Practicing Nurse". In: *Critical Care Nursing Clinics of North America*. Human Factors and Technology in the ICU 30.2 (June 1, 2018), pp. 247–257. ISSN: 0899-5885. DOI: 10.1016/j.cnc.2018.02.007. URL: https://www.sciencedirect.com/science/article/pii/S0899588518300091 (visited on 07/22/2024).

[24] Charlotte L. Zwager Lucas M. Fleuren Thomas L. T. Klausch. "Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy". In: *Springer Link* (2020). DOI: https://doi.org/10.1007/s00134-019-05872-y.

[25] Fengshuo Xu Luming Zhang Tao Huang. "Prediction of prognosis in elderly patients with sepsis based on machine learning (random survival forest)". In: *Springer Link* (2022). DOI: https://doi.org/10.1186/s12873-022-00582-z.

[26] Fahim Mahmud, Naqib Sad Pathan, and Muhammad Quamruzzaman. "Early detection of Sepsis in critical patients using Random Forest Classifier". In: *2020 IEEE Region 10 Symposium (TENSYMP)*. 2020, pp. 130–133. DOI: 10.1109/TENSYMP50017.2020.9231011.

[27] Shaheen MZ. Memon, Robert Wamala, and Ignace H. Kabano. "A comparison of imputation methods for categorical data". In: *Informatics in Medicine Unlocked* 42 (Jan. 1, 2023), p. 101382. ISSN: 2352-9148. DOI: 10.1016/j.imu.2023.101382. URL: https://www.sciencedirect.com/science/article/pii/S2352914823002289 (visited on 07/22/2024).

[28] Michael Moor et al. "Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review". In: *Frontiers in Medicine* 8 (May 28, 2021). Publisher: Frontiers. ISSN: 2296-858X. DOI: 10.3389/fmed.2021.607952. URL: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2021.607952/full (visited on 07/22/2024).

[29] Lu He Nianzong Hou Mingzhe Li. "Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost". In: *Springer Link* (2020). DOI: https://doi.org/10.1186/s12967-020-02620-5.

[30] Varesh Prasad et al. "Diagnostic suspicion bias and machine learning: Breaking the awareness deadlock for sepsis detection". In: *PLOS Digital Health* 2.11 (Nov. 1, 2023). Ed. by Luis Filipe Nakayama, e0000365. ISSN: 2767-3170. DOI: 10.1371/journal.pdig.0000365. URL: https://dx.plos.org/10.1371/journal.pdig.0000365 (visited on 07/22/2024).

[31] *Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis — Nature Medicine*. URL: https://www.nature.com/articles/s41591-022-01894-0 (visited on 07/22/2024).

[32] Shuangxia Ren et al. "Machine learning based algorithms to impute PaO2 from SpO2 values and development of an online calculator". en. In: *Scientific Reports* 12.1 (May 2022). Publisher: Nature Publishing Group, p. 8235. ISSN: 2045-2322. DOI: 10.1038/s41598-022-12419-7. URL: https://www.nature.com/articles/s41598-022-12419-7 (visited on 07/22/2024).

[33] Chanu Rhee and Michael Klompas. "Sepsis trends: increasing incidence and decreasing mortality, or changing denominator?" In: *Journal of Thoracic Disease* 12.Suppl 1 (Feb. 2020), S89–S100. ISSN: 2072-1439. DOI: 10.21037/jtd.2019.12.51. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7024753/ (visited on 07/21/2024).

[34] Benjamin Roussel, Joachim Behar, and Julien Oster. "A Recurrent Neural Network for the Prediction of Vital Sign Evolution and Sepsis in ICU". In: *2019 Computing in Cardiology (CinC)*. 2019, Page 1–Page 4. DOI: 10.22489/CinC.2019.082.

[35] Matthieu Scherpf et al. "Predicting sepsis with a recurrent neural network using the MIMIC III database". In: *Computers in Biology and Medicine* 113 (2019), p. 103395. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2019.103395. URL: https://www.sciencedirect.com/science/article/pii/S0010482519302720.

[36] *Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review — Journal of the American Medical Informatics Association — Oxford Academic*. URL: https://academic.oup.com/jamia/article/29/3/559/6460282 (visited on 07/22/2024).

[37] Michael W. Sjoding et al. "Racial Bias in Pulse Oximetry Measurement". eng. In: *The New England Journal of Medicine* 383.25 (Dec. 2020), pp. 2477–2478. ISSN: 1533-4406. DOI: 10.1056/NEJMc2029240.

[38] Ivana Srzić, Višnja Nesek Adam, and Darinka Tunjić Pejak. "SEPSIS DEFINITION: WHAT'S NEW IN THE TREATMENT GUIDELINES". In: *Acta Clinica Croatica* 61.Suppl 1 (June 2022), pp. 67–72. ISSN: 0353-9466. DOI: 10.20471/acc.2022.61.s1.11. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9536156/ (visited on 07/23/2024).

[39] Longxiang Su et al. "Early Prediction of Mortality, Severity, and Length of Stay in the Intensive Care Unit of Sepsis Patients Based on Sepsis 3.0 by Machine Learning Models". In: *Frontiers in Medicine* 8 (June 28, 2021). Publisher: Frontiers. ISSN: 2296-858X. DOI: 10.3389/fmed.2021.664966. URL: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2021.664966/full (visited on 07/22/2024).

[40] *Superhuman performance on sepsis MIMIC-III data by distributional reinforcement learning — PLOS ONE*. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0275358 (visited on 07/22/2024).

[41] M. Tallgren, M. Bäcklund, and M. Hynninen. "Accuracy of Sequential Organ Failure Assessment (SOFA) scoring in clinical practice". en. In: *Acta Anaesthesiologica Scandinavica* 53.1 (2009). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.13 6576.2008.01825.x, pp. 39–45. ISSN: 1399-6576. DOI: 10.1111/j.1399-6576.2008.01825.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1399-6576.2008.01825.x (visited on 07/22/2024).

[42] *The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) — Critical Care Medicine — JAMA — JAMA Network*. URL: https://jamanetwork.com/journals/jama/fullarticle/2492881 (visited on 07/22/2024).

[43] Bas H. M. van der Velden et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis". In: *Medical Image Analysis* 79 (July 1, 2022), p. 102470. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102470. URL: https://www.sciencedirect.com/science/article/pii/S1361841522001177 (visited on 07/22/2024).

[44] Swaathi Venkat et al. “Machine Learning based SpO2 Computation Using Reflectance Pulse Oximetry”. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. ISSN: 1558-4615. July 2019, pp. 482–485. DOI: 10.1109/EMBC.2019.8856434. URL: https://ieeexplore.ieee.org/abstract/document/8856434 (visited on 07/22/2024).

[45] J. L. Vincent et al. “The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine”. eng. In: *Intensive Care Medicine* 22.7 (July 1996), pp. 707–710. ISSN: 0342-4642. DOI: 10.1007/BF01709751.

[46] Dong Wang et al. “A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients”. In: *Frontiers in Public Health* 9 (2021). ISSN: 2296-2565. DOI: 10.3389/fpubh.2021.754348. URL: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2021.754348.

[47] John G. Webster. *Design of Pulse Oximeters*. en. Google-Books-ID: eQh1DQtvowUC. CRC Press, Oct. 1997. ISBN: 978-1-4200-5079-0.

[48] Zhenyu Yang, Xiaoju Cui, and Zhe Song. “Predicting sepsis onset in ICU using machine learning models: a systematic review and meta-analysis”. In: *BMC Infectious Diseases* 23.1 (Sept. 27, 2023), p. 635. ISSN: 1471-2334. DOI: 10.1186/s12879-023-08614-0. URL: https://doi.org/10.1186/s12879-023-08614-0 (visited on 07/22/2024).