University College Dublin
Arcadia University
University of North Carolina at Chapel Hill

AUGMENTING SOFA SCORES FOR SEPSIS PREDICTION WITH MACHINE LEARNING

Feature Augmentation, Explainable AI, and User Interface Implementation

Authors

Kathleen Fort

Dylan Sain
Maria Jorge

Supersivors

Alin Navis, PhD
Catherine Mooney, PhD

July 2024

Contents

Contents i

1 Introduction 2
- 1.1 Definitions . . . 3
- 1.2 Data Pre-processing . . . 3
- 1.3 Previous Results . . . 4
  - 1.3.1 Performance and Bias . . . 4
  - 1.3.2 Imputation Methods . . . 5
  - 1.3.3 Screen Fatigue . . . 6
  - 1.3.4 Interpretability, Explainability, and Usability . . . 6
- 1.4 Gaps . . . 7

2 Methods 8
- 2.1 Pulse Oximeter Bias Correction . . . 8
- 2.2 Fairlearn Package . . . 8
- 2.3 Data Configuration and Pre-Processing . . . 9
- 2.4 Algorithm Implementation and Fine-Tuning . . . 10
- 2.5 Explainable AI Methods . . . 11
- 2.6 User Interface Development . . . 12
- 2.7 Model Enhancement and Data Integration . . . 14
- 2.8 Data Collection . . . 14

3 Materials 15
- 3.1 Software and Tools . . . 15
- 3.2 Data . . . 15
- 3.3 Explainability Tools . . . 15

4 Results 16
- 4.1 Pulse Oximetry Bias . . . 16
- 4.2 Model Performance and User Interface Output . . . 16
- 4.3 User Interface . . . 20
- 4.4 Explainable AI Results . . . 20
- 4.5 Discrepancies in XAI Results . . . 21

5 Discussion and Conclusion **22**

6 Contributions **24**

6.1 Dylan Sain 24

6.2 Kathleen Fort 24

6.3 Maria Jorge 24

1 Introduction

In recent years, machine learning (ML) and artificial intelligence (AI) have become integral to various aspects of life Yang et al., 2023. As ML progresses, there is growing interest in its potential to assist clinicians in medical diagnoses and treatments Eloranta et al., 2022. One notable area of focus is the prediction of sepsis onset.

Sepsis is a condition with one of the highest mortality and morbidity rates in the United States and other high-income countries Singer et al., 2016. The average 30-day mortality rate for sepsis is 24.4%, which increases to 34.7% for septic shock Bauer et al., 2020. Early identification and antibiotic treatment have been shown to improve patient outcomes significantly, with delays of just six hours increasing the mortality rate by 7.6% Yan et al., 2021. Currently, sepsis risk is primarily assessed using diagnostic tools such as the Sepsis-related Organ Failure Assessment (SOFA) Score and the Acute Physiology and Chronic Health Evaluation (APACHE) Islam et al., 2019. Adams et al. highlight machine learning's potential to reduce the global burden of sepsis by analyzing vital signs from a patient, assessing their sepsis risk, and helping clinicians anticipate the need for antibiotics Adams et al., 2022.

A sepsis-predicting ML tool raises concerns about bias affecting the model's performance in clinical settings. Developers and clinicians must identify and mitigate biases to prevent the deployment of ineffective systems. Sources of bias include non-representative data sets, erroneous imputation techniques, and existing prejudices within the healthcare system Fleuren et al., 2020.

One specific source of bias regarding sepsis prediction is erroneous pulse oximeter readings. Clinicians often use pulse oximeters to read a patient's arterial blood oxygen saturation (SpO2). However, studies on pulse oximeter outputs have found that these readings have higher error rates when used on individuals with darker skin tones due to higher levels of melanin in the skin. For pulse oximeter readings to be used in machine learning data, programmers must address the bias toward patients of color Jamali et al., 2022 Sjoding et al., 2020.

This paper will utilize the BOLD dataset to predict a patient's 24-hour deterioration or improvement. It includes an investigation of important features through feature selection and XAI. It utilizes three different machine learning models and various

feature selection tools to bring about a new understanding of SOFA scores and the prediction of patient's 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 paper 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.

1.1 Definitions

Predicting sepsis is challenging in part due to the lack of unique identifying biomarkers. Additionally, the definition of sepsis varies between hospitals. Evans et al. explains how, according to an international task force, sepsis is broadly defined as a life-threatening organ dysfunction caused by the body's abnormal response to infection. This task force has identified the SOFA score as the best indicator for predicting in-hospital mortality caused by sepsis Evans, 2018, which evaluates a patient's respiration, coagulation, cardiovascular, renal, liver, and central nervous system functions. Additionally, Adams et al. found that a patient's risk of sepsis can be most efficiently and accurately assessed using the quick SOFA (qSOFA) score, which evaluates blood pressure, respiratory rate, and Glasgow Coma Scale score Adams et al., 2022.

In cases where sepsis progresses, a patient may develop septic shock, defined as a subset of sepsis where circulatory, cellular, and metabolic abnormalities are associated with a higher risk of mortality compared to sepsis alone Evans, 2018. Despite these standardized definitions, their integration into clinical practice remains inconsistent. Moreover, gaps remain in the definitions of sepsis, such as the lack of a precise definition of "infection" Evans, 2018.

By extension, comparing different algorithms is challenging due to variations in the timing of sepsis detection. Sepsis prediction can occur before sepsis develops or at the onset of sepsis. Furthermore, models can either predict sepsis continuously (right-leaning algorithms) or at a specific point in time (left-leaning algorithms) Fleuren et al., 2020. Additionally, depending on the model, sepsis identification can detect sepsis, severe sepsis, or septic shock. In the ICU, models are more likely to identify sepsis or severe sepsis, while in hospitals, the focus is often on septic shock Yan et al., 2021.

1.2 Data Pre-processing

One challenge with creating a sepsis-detection algorithm is working with imperfect data on sepsis cases. Often, medical data sets, especially when they focus on time-series data, have large amounts of missing data. Past studies have used simplistic approaches to address this issue, for example, dropping features with missing values more than 20% Luming Zhang, 2022 Nianzong Hou, 2020. A more sophisticated approached used is

linear interpolation Mahmud et al., 2020 Scherpf et al., 2019 Liu et al., 2022. However, while linear interpolation performs well for values temporally close, it fails for values farther apart in time. Other approaches involve utilizing missing values in the data, as, in theory, it could reveal a pattern in the data collection Mahmud et al., 2020 Futoma et al., 2017 Roussel et al., 2019. Some studies expand upon this theory, utilizing methods such as Gaussian Processes Futoma et al., 2017 or models like XGBoost Nianzong Hou, 2020 to make missing values have less impact on the model.

Another challenge that arises in data relating to sepsis cases is the disporportionate ratio of 'positive' to 'negative' values Mahmud et al., 2020. Even in large datasets, there are much more non-sepsis patients than patients with sepsis. This causes the ML model to be biased towards non-sepsis patients.

When fine-tuning a ML model, feature selection must be considered. Past models have utilized various tests such as ANOVA, KS(Kolmogorov-Smirnov), Mann Whitney U, Chi-squared, and Fisher's exact to determine the features that optimize model performance Mahmud et al., 2020 Wang et al., 2021 Luming Zhang, 2022 Liu et al., 2022 Nianzong Hou, 2020 Chang et al., 2024 Lucas M. Fleuren, 2020. These tests are significant, as the lack of feature selection tests intensifies the black box effect in machine learning models Futoma et al., 2017 Scherpf et al., 2019 Roussel et al., 2019.

1.3 Previous Results

1.3.1 Performance and Bias

The most common metric evaluation for machine learning models is the Area Under the ROC Curve (AUC). However, as it is difficult to assess sepsis prediction with precision-based accuracy metrics, a more sophisticated way of evaluating the model 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 user-given criteria.

Among algorithms for sepsis prediction, XGBoost and Random Forest models have consistently demonstrated high performance. These models have shown significant advantages in predicting sepsis in ICU patients, exhibiting higher accuracy (ACC) and concordance index (c-index) values compared to other models. However, a systematic review of models for sepsis prediction by Adams et al. (2022) found that most exhibited a high risk of bias, while the remaining models had an unknown risk of bias. These risks were attributed to factors such as small sample sizes, missing data, and the misinterpretation of complex data Adams et al., 2022.

One source of bias in these algorithms is incorporation bias, which often occurs when investigational predictors are also determinative factors in defining the outcome Prasad et al., 2023. This could involve using diagnostic data to predict the likelihood of a patient developing sepsis. This presents bias because such tests are typically ordered only if a clinician suspects the patient is deteriorating. A similar case can occur when

using the administration of antibiotics as an indicator of sepsis risk. Training a model on this data risks categorizing patients as "low risk" simply because no diagnostic tests or antibiotics were ordered for them, causing both the model and the clinician to overlook those most at risk of delayed antibiotic treatment due to vague symptoms and no obvious vital sign abnormalities. A study reviewing 107 predictive algorithms found that none accounted for "informative observations," where the presence of a diagnostic observation was non-random and driven by clinician concern. This oversight not only increases the risk of misdiagnosis but also exacerbates the likelihood of a "diagnostic deadlock," where a clinician's suspicions are reinforced by the model, leading to further delays in diagnosis Prasad et al., 2023.

One strategy to mitigate bias through data integration involves using "bland" data during model training. Prasad et al. investigated the effectiveness of employing "bland" clinical data to counteract bias within ML models. Their study revealed that the bland model exhibited subpar performance metrics, including elevated rates of false positives and poor sensitivity Prasad et al., 2023.

A subsection of assessing bias in the model is ensuring that the model is fair Chang et al., 2024. The majority of sepsis-prediction algorithms have been trained on a dataset witha majority white subjects, where people of color are under-represented. One study found that there is a positive correlation between fairness and better predictability. When creating a machine learning model, especially in a medical context, it is critical that the model can accurately predict for all ages, races, ethnicities, genders, and social statuses.

1.3.2 Imputation Methods

Another avenue through which bias can infiltrate AI algorithms is via data imputation, the process by which missing data is handled within the model. Several studies have explored the use of the K-nearest neighbors (KNN) approach in conjunction with the SMOTE oversampling technique to address issues related to imbalanced datasets Su et al., 2021 Böck et al., 2024 He et al., 2023. Research conducted by Memon et al. demonstrated that KNN imputation achieved superior precision in predicting missing values for categorical variables compared to methodologies such as Random Forest, Sequential Hot-Deck, and Multiple Imputation by Chained Equations Memon et al., 2023.

Furthermore, Su et al. endorsed the application of KNN imputation specifically within sepsis prediction models, while acknowledging the potential efficacy of alternative methods such as stochastic regression and tree-based models for future comparative analyses Su et al., 2021. In another investigation, Baniasadi et al. employed a two-step imputation strategy involving linear interpolation combined with a sample-weighted AdaBoost model to manage missing data effectively, enhancing model robustness and generalizability relative to other approaches Baniasadi et al., 2021.

1.3.3 Screen Fatigue

Studies on screen fatigue have highlighted that, while electronic interfaces in clinical settings achieve high ratings in terms of data accuracy, precision, and processing efficiency, they also provoke significant dissatisfaction among users. This dissatisfaction stems from the perceived effort required, frequent interruptions, and frustration associated with navigating these interfaces Hilty et al., 2022.

A meta-analysis by Dunn Lopez et al. emphasized that electronic usage in clinical settings can lead to fatigue, including emotional exhaustion, weariness, and challenges in maintaining engagement. Emotional repercussions such as anger, irritability, and stress were frequently reported, reflecting the adverse psychological impact of prolonged electronic interface interactions. Furthermore, there exists a noticeable absence of standardized assessments, monitoring mechanisms, or intervention protocols specifically tailored for technology in clinical workplaces beyond initial onboarding or training processes. These factors contribute to heightened complexity within the already stressful and high-pressure healthcare environment Lopez et al., 2018.

Eye strain from prolonged screen use in clinical settings contributes significantly to fatigue. Krupinski et al. found that increased screen time leads to symptoms like blurred vision and difficulty focusing. They also noted a rise in clinical errors after extended screen reading, especially when screens were viewed at close distances Krupinski et al., 2010.

One of the primary drivers of screen fatigue is the extensive time spent using technology. Tutty et al. found that for every hour physicians spent on direct patient care, they spent nearly three additional hours interacting with electronic health records. This study underscored the critical role of leadership in engaging physicians during technology implementation and emphasized the importance of workflow design in the UI. It advised implementation teams to conduct thorough testing before and after implementation using scenarios aimed at minimizing clinical burden Tutty et al., 2019.

In extension to screen fatigue, alarm fatigue is a phenomenon that can increase clinician stress in the hospital and decrease both efficiency and the quality of patient care Futoma et al., 2017 Scherpf et al., 2019 Lucas M. Fleuren, 2020. According to one study, 63.4% of alerts from hospital monitors and alarms were canceled by the nurses who received them Futoma et al., 2017. A sepsis-prediction algorithm must minimize the number of false positives to avoid the consequences of alarm fatigue.

1.3.4 Interpretability, Explainability, and Usability

Ensuring the interpretability of AI models by clinicians, developers, and patients is crucial. As models become more complex to achieve higher performance and accuracy, there is a decline in explainability and transparency. Post hoc explanations are necessary to interpret outputs from ML algorithms rather than relying solely on inherent explainability. These methods include analyzing learned features and assessing feature importance and interactions. Two valuable methods for explaining AI in clinical

contexts are Local Interpretable Model-Agnostic Explanations (LIME) and the use of Shapley values (SHAP) Velden et al., 2022.

As AI usage expands in medical settings, clinical validation emerges as a fundamental requirement. Clinical validation ensures that AI systems can effectively perform in real-world scenarios, focusing on prediction accuracy and minimizing error rates Amann et al., 2020. The explainability of AI tools is crucial for resolving disagreements between clinicians and AI by allowing clinicians to evaluate the model's recommendations. Moreover, explainability allows users to identify model flaws and provide feedback to improve training and performance Amann et al., 2020. Explainable AI (XAI) techniques, such as assessing feature importance, can be utilized in these scenarios Di Martino et al., 2022.

1.4 Gaps

For widespread implementation of sepsis prediction algorithms in clinical settings, a standard definition of sepsis is essential to facilitate accurate comparisons of different prediction models and septic survival rates, as well as for identifying potential biases within these models Bauer et al., 2020. Moreover, there exists a notable absence of a universally accepted checklist for evaluating the quality of diagnostic machine-learning research within medical contexts. Establishing such standards for assessing AI can mitigate the risk of biases in models Lopez et al., 2018. Additionally, there is frequently insufficient transparency regarding sepsis prediction models, with many developers of high-performing algorithms withhold essential information such as the training dataset and model specifics. Enhancing transparency is crucial for the widespread adoption of sepsis prediction models in hospital settings Yan et al., 2021.

Another gap in the implementation of sepsis prediction algorithms is the lack of data on sepsis patients in lower-income countries. This limits the use of sepsis prediction data and therefore sepsis prediction models to population levels Fleuren et al., 2020.

Furthermore, there is a pressing need to develop models that rely on noninvasive or minimally invasive indicators. Many existing models evaluated in research depend on invasive procedures to achieve accurate predictions Adams et al., 2022. For example, recent comprehensive studies suggest that arterial blood gas (ABG) results obtained through invasive methods may offer more accurate indicators of sepsis than previously understood Fleuren et al., 2020. To address this challenge, our research group aims to mitigate bias by improving the accuracy of Pulse-Oximeter results as an alternative to ABG testing. This approach seeks to enhance the reliability and accessibility of sepsis prediction without the need for invasive procedures.

2

Methods

2.1 Pulse Oximeter Bias Correction

One way to correct pulse oximeter bias is by a ML algorithm that can correct the erroneous SpO2 readings from patients by re-calibrating the inaccurate R-value based methods Venkat et al., 2019 Ren et al., 2022. The proposed model overcomes the bias limitations by the traditional R-value-based calibration method through a machine learning model using various time and frequency domain features Venkat et al., 2019 Ren et al., 2022. The model was trained and tested using the clinical data collected from 95 subjects with SpO_2 levels varying from 81-100%. An XGBoost Regressor Model was used with a 4-fold cross-validation grid search to classify hyperparameters Chen et al., 2016. The model yielded an R2 of 21.8% to 67.6% and an RMSE of 2.01% to 2.63% in (SaO2, SpO2) across racial and ethnic subgroups Chen et al., 2016. The best improvement was obtained for black patients. Overall, the designed model reduced (SaO2, SpO2) error, i.e. bias in pulse oximetry.

2.2 Fairlearn Package

Fairlearn was integrated into the ML model to ensure fairness in the algorithm. Fairlearn is an open-source Python library developed by Microsoft that provides tools to understand, evaluate, and improve the fairness of machine learning models by examining the disparities in model performance across different demographic groups. Fairlearn uses fairness metrics to evaluate how demographics (race, gender, age, etc.) are treated by a machine learning model. These metrics explain disparities in predictions and outcomes Tallgren et al., 2009. Integrating several mitigation algorithms designed to deal with unfairness in ML models helps reduce bias and ensure more equitable outcomes across all demographic groups. Fairlearn allows users to perform detailed model assessments by analyzing how different groups are affected by the model's predictions, including tools to examine the trade-offs between fairness and accuracy. Finally, Fairlearn is designed to work seamlessly with popular machine learning frameworks like Visual Studio Code, making it easy to integrate into existing workflows.

Our team used Fairlearn's Equalized Metrics package, which uses Equalized Opportunity and Equalized odds. Equalized Opportunity is when a model's true positive rate is equal across different groups, and Equalized Odds require that both the true positive rate (TPR) and the false positive rate (FPR) be equal across different groups.

2.3 Data Configuration and Pre-Processing

It is inferred that the BOLD data set has a great amount of misleading SpO2 levels due to the inherent bias in pulse oximeter results. To use this data set, an algorithm in the Visual Studio Code programming platform will be integrated to correct the bias seen in the data and ensure the accuracy of the ML model results.

To ensure the machine learning model is fair when analyzing data among demographics, the model will need to use a fairness package. Our team selected the Fairlearn package created by Microsoft due to Fairlearn's efficient and straightforward programming and its pre-installed packages for data processing and bias analyzation.

Our research team used the BOLD dataset to train the machine learning model. 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, there is extensive missing data within BOLD (see Figure 2.1).

Initially, we removed specific patients due to missing critical prediciton values, decreasing the size of the dataset from $49,000$ to approximately $30,000$. Furthermore, any row with over 25% of its values missing was dropped.

After these pre-processing steps, the dataset still presented a large amount of missing values. To mitigate this deficiency in the data, our team experimented with various imputation techniques, from linear interpolation to moving averages or a full KNN imputation. Out of these imputation techniques, KNN imputation performed the best.

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

After data imputation, we were able to use the dataset for more extensive modeling. The use of smaller KNN subsets was crucial in reducing bias within the data, as it allowed the application of medical knowledge, specifically how one feature might affect the other. For example, there is a negative correlation between a patient's hemoglobin and fibrinogen levels. Despite this, they exist in two different lab subsets: CBC and COAG. Armed with this knowledge, the KNN imputation strategy is able to accurately predict the missing values.

Through data analysis, Methemoglobin and Carboxyhemoglobin levels were found to have a large number of missing values, as they were only ordered by specific hospitals. These features were dropped to avoid the potential of introducing hospital bias to the

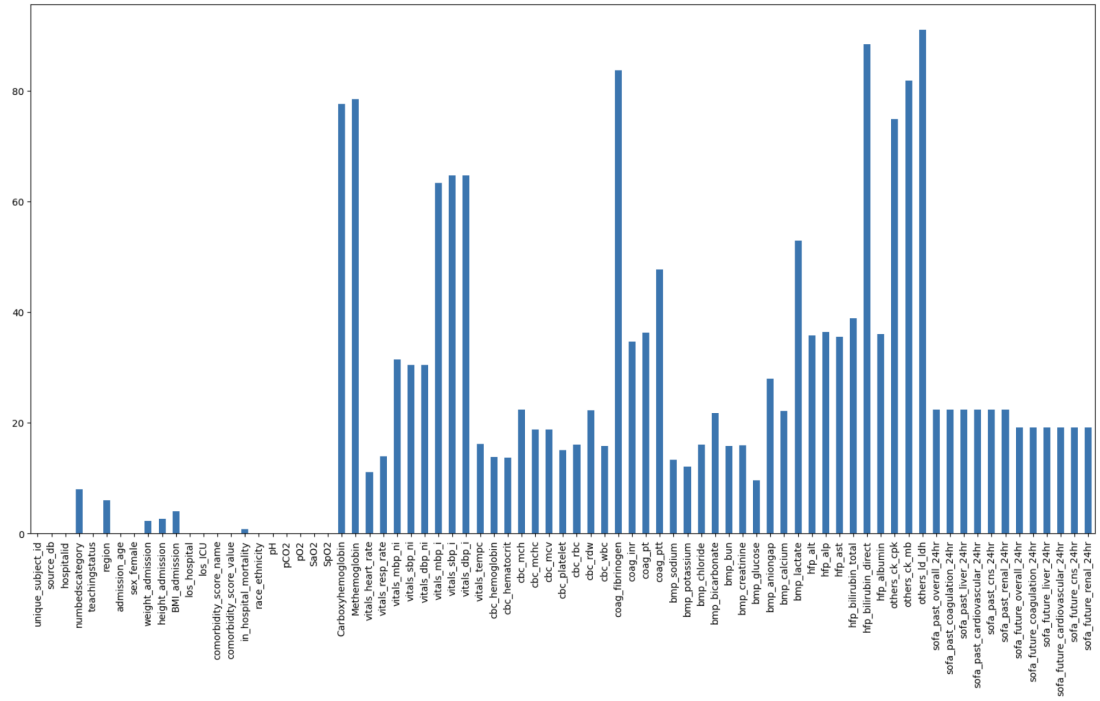

Figure 2.1: *A figure containing all features in BOLD on the X axis and the percentage (%) of missing values on the Y axis*

model. We also had to adjust the comorbidity score as the dataset included two scales, Charlson and Elixhauser. This was accomplished using a min/max scaling. One-hot encoding was used on gender and other text-based columns.

2.4 Algorithm Implementation and Fine-Tuning

Based on the tabular data, the binary classification class, and previous work on the subject, our team used three models to predict patient status after 24 hours: an XGBoost model, a Random Forest classifier, and a simple multi-layer Perceptron. These models were chosen due to their ability to analyze tabular data, in addition to performing well in classification tasks. Furthermore, the XGBoost and Random Forest models have high levels of explainability.

We ran a series of Optuna trials for hyperparameter tuning on a 60, 20, 20 train/test/validation split. The Bayesian technique provided by these trials 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 the number 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 the number of estimators, max depth, minimum sample spilt, and minimum samples per leaf.

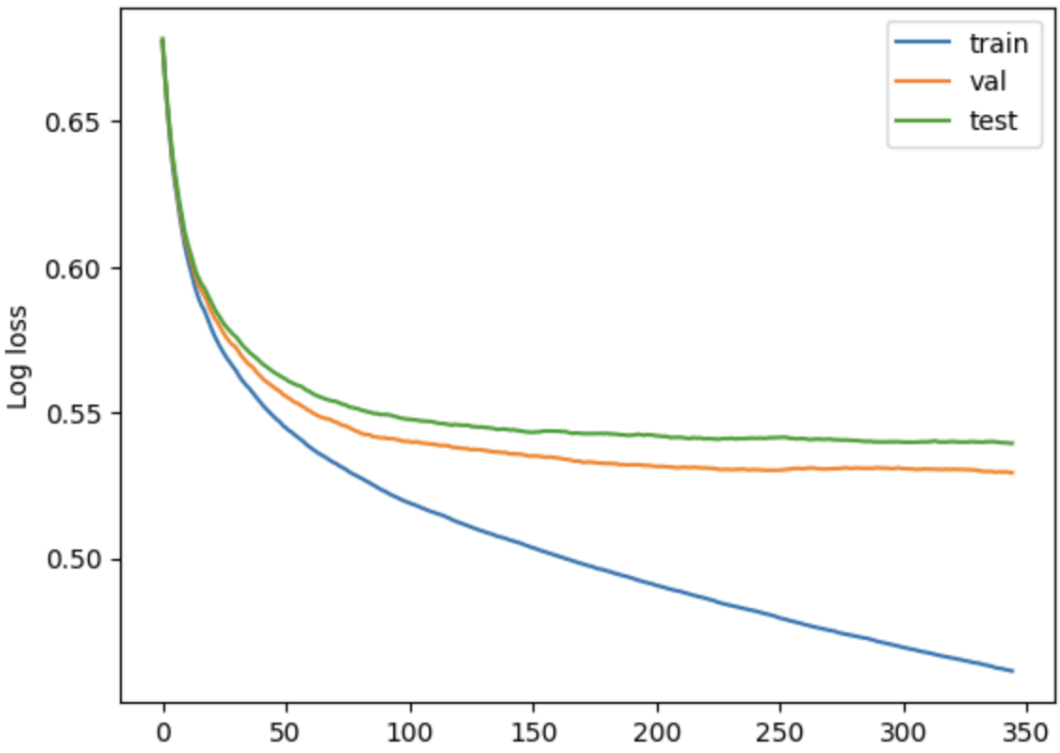

Figure 2.2: *A figure representing the log loss of a XGBoost model over the training process.*

The MLP was tuned for first-layer neurons, second-layer neurons, learning rate, activation functions, solver, and an alpha value.

Each trial was set to optimize the log-loss of the validation set. A graph of the training can be seen in Figure 2.2. The model was also fine-tuned for feature engineering to balance maximalizing predictive power and minimalizing the number of user inputs. While the Optuna trials utilized all 62 features, it was critical to reduce this number to create a user-friendly interface. Using a technique called recursive feature elimination (RFE) and feature importance plots, we created a ranking of the top features based on their scores and contributions toward a model's F1 score. This ranking was used to choose the top features, and also as an indicator of what additional labs could be used to augment SOFA scores in predicitons.

2.5 Explainable AI Methods

To achieve local explainability in our predictive model, we employed Shapley Additive Explanations (SHAP). SHAP was selected due to its high performance and ability to provide detailed local explainability. 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 2.3).

Figure 2.3: *SHAP Local Explainability for a patient x*

In addition to SHAP, we used Permutation Feature Importance (PFI) and Local Interpretable Model-Agnostic Explanations (LIME) to achieve global model explainability (see Figures 2.4, 2.5, and 2.6). Although SHAP consistently outperforms other methods, employing a variety of techniques allowed us to identify consistent patterns in feature importance across different explainability methods.

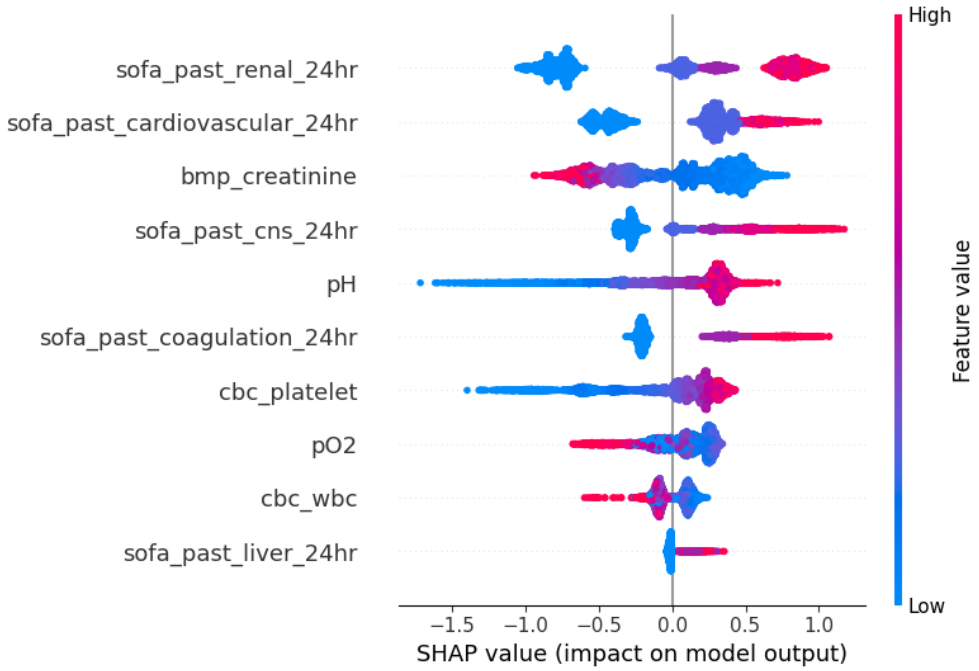

Figure 2.4: *SHAP Global Explainability for the machine learning model*

2.6 User Interface Development

The user interface was developed using the Django framework, selected for its robust back-end admin interface and scalability. The website code was written in Python using Visual Studio Code 4. To enhance the interface, we utilized Django packages, Bootstrap for styling, and JavaScript for interactivity. Plotly was chosen for visualizing SHAP values due to its compatibility with Django.

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. The view computes additional variables and constructs a data frame with the patient's information. This data frame is then passed into the predictive model, and the

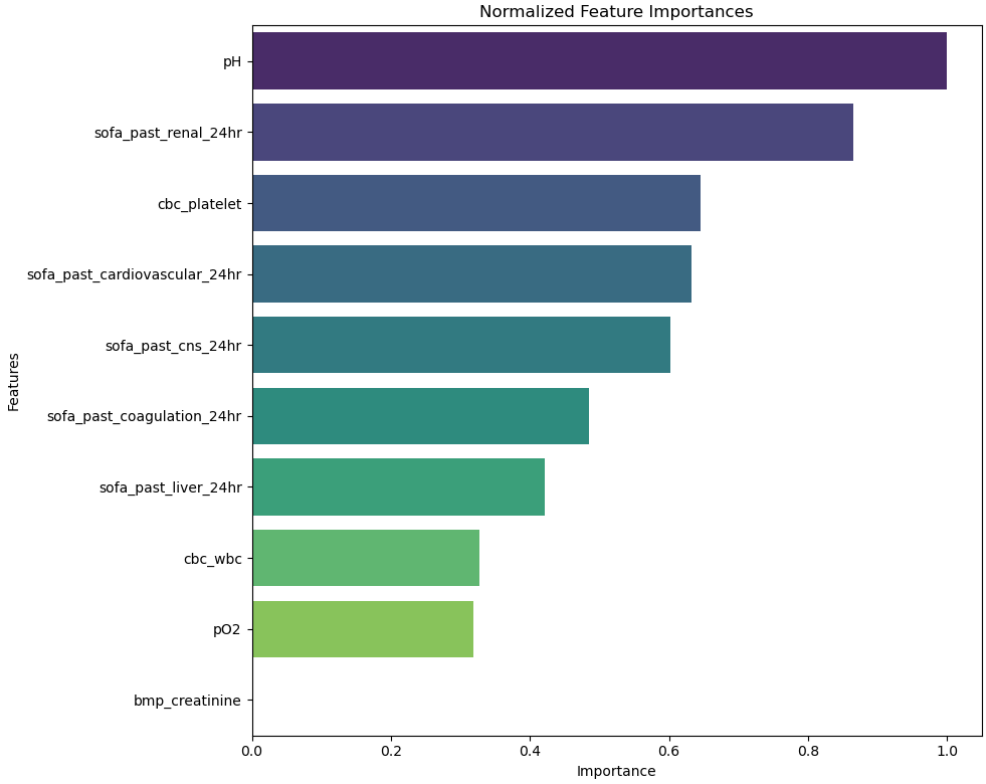

Figure 2.5: *LIME Global Explainability for the machine learning model*

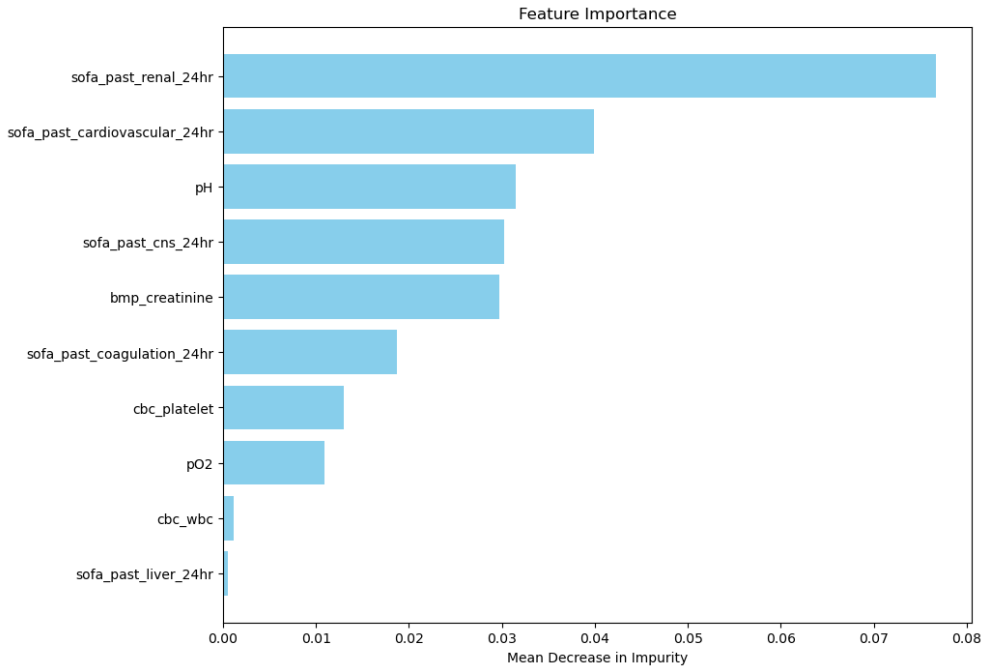

Figure 2.6: *PFI Global Explainability for the machine learning model*

predicted output is stored as a variable. Additionally, the data frame is used to calculate SHAP values and the expected value (baseline risk).

A force plot is created with the SHAP values, scaled to our prediction range (0 to 1), and converted into HTML for display to the user. A context dictionary containing the prediction, SHAP values, vitals data, and page title is passed to the rendered template. Upon successful computation, the user is redirected to a results page displaying the prediction and SHAP visualization.

2.7 Model Enhancement and Data Integration

To enhance the predictive power of our model, we incorporated additional laboratory data variables, including white blood cell count, partial pressure of oxygen (pO2) levels, and blood pH measurements. We also integrated raw data on creatinine and platelet count levels alongside the Sequential Organ Failure Assessment (SOFA) categories for renal and coagulation scores.

2.8 Data Collection

Our user interface is designed to collect patient demographic information (specifically race/ethnicity and gender) exclusively for retrospective bias and performance analysis. These demographic variables are not utilized in the predictive modeling process.

Instead of directly collecting SOFA coagulation and renal scores from users, the interface gathers raw data on platelet count and creatinine levels, respectively. These values are stored as individual variables and are used to compute the patient's SOFA scores for coagulation and renal function. Both the raw data and the computed SOFA scores are then passed into the predictive model.

3

Materials

3.1 Software and Tools

- Visual Studio Code 4 (code editor)
- Python (programming language)
- Fairlearn Package
- Django (web framework)
- Bootstrap (CSS framework)
- JavaScript (scripting language)
- Plotly (data visualization library)

3.2 Data

- BOLD, a blood-gas and oximetry linked dataset (Data set includes eICU Collaborative Research Database, MIMIC-III Clinical Database, MIMIC-IV) [PHYSIONET]
- UI: Patient demographics (race/ethnicity, gender)
- UI: Laboratory data (white blood cell count, pO2 levels, blood pH, creatinine, platelet count)
- UI: Imputed SOFA scores (respiration, central nervous system, and liver)
- UI: Computed SOFA scores (coagulation and renal)

3.3 Explainability Tools

- Shapley Additive Explanations (SHAP, global and local methods)
- Permutation Feature Importance (PFI, global methods)
- Local Interpretable Model-Agnostic Explanations (LIME, global methods)

By integrating these materials and methods, our team aimed to develop a robust and interpretable predictive model, enhancing the understanding of individual patient risks and improving overall model performance.

4

Results

4.1 Pulse Oximetry Bias

After the alayzation of pulse oximetry bias in the data, it was observed that the fault lies in the device design, as engineers overlooked the effect that melanin would have on UV light measurements. Therefore, the data needed to be analyzed and corrected for any of the foreseen bias. By using a data processing algorithm, it was seen that most of the patients exhibit healthy SpO2 levels (see Figures 4.1, 4.2, 4.3). When applying the Fairlearn Equalized metrics package to analyze the results of the model, a table was generated to observe the accuracy of the results (see Figure 4.4). See Figure 4.5 for exponential gradient metrics results.

4.2 Model Performance and User Interface Output

Using the most efficient models and their given hyperparameters provided from the Optuna trials, these models were tested with the test set. The models achieved scores just above 70% (see Table 4.1). 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 (improvement) classes, the confusion matrix didn't display any significant model bias.

The most significant features were selected for the final model based on feature importance determined by RFE results (see a sample of features in Table 4.2). Using this table, our team trained various XGBoost models on an increasing number of features. Figure 4.6 displays the culminated results of these trials. An F1 score of nearly 70% was achieved with only 5 to 10 features. Our model's performance analysis revealed that

	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 4.1: *Reported F1 Score for the various full models*

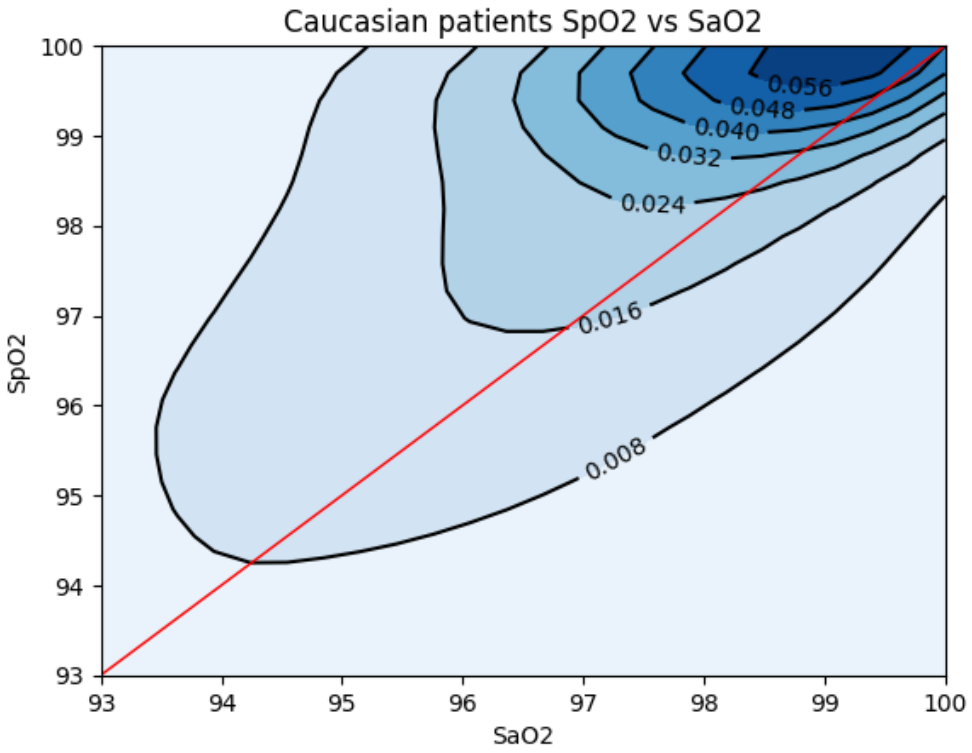

Figure 4.1: *Density plot of caucasian patients, SpO2 vs SaO2*

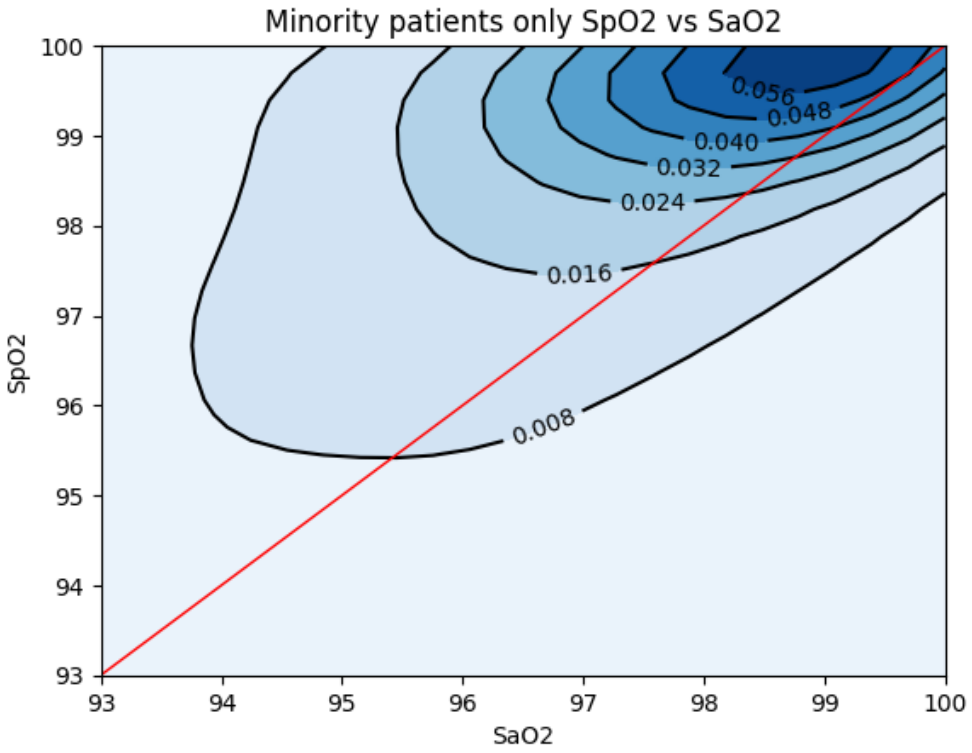

Figure 4.2: *Density plot of minority patients, SpO2 vs SaO2*

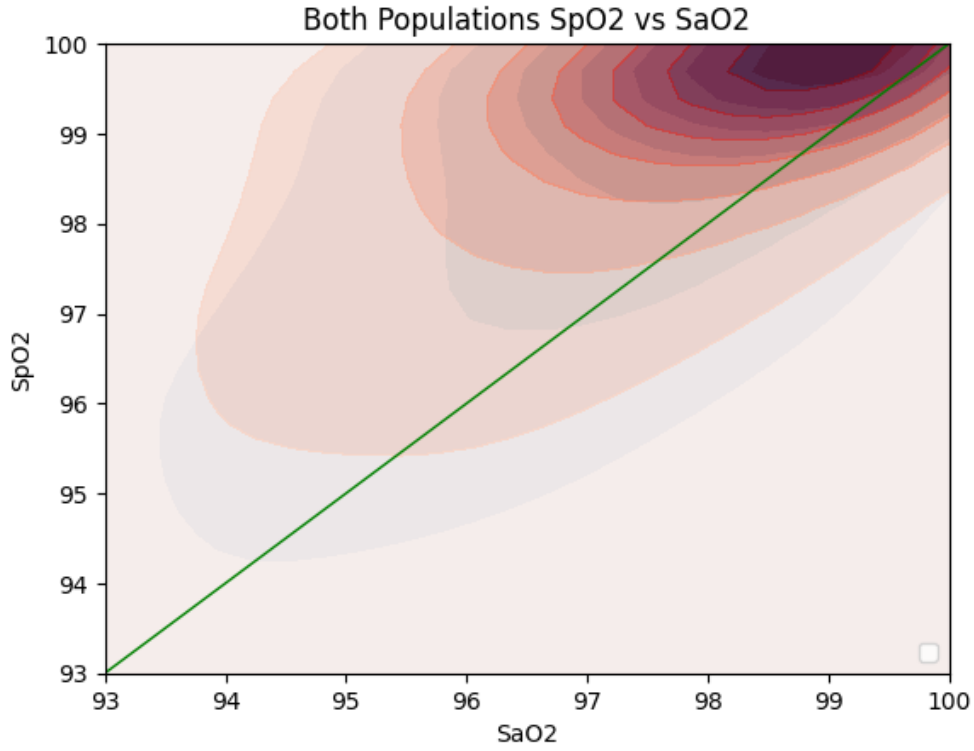

Figure 4.3: *Overlapped populations, SpO2 vs SaO2*

		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 4.4: *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 4.5: *Exponential gradient metrics results*

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 4.2: *Average selected and average score for various features using RFE*

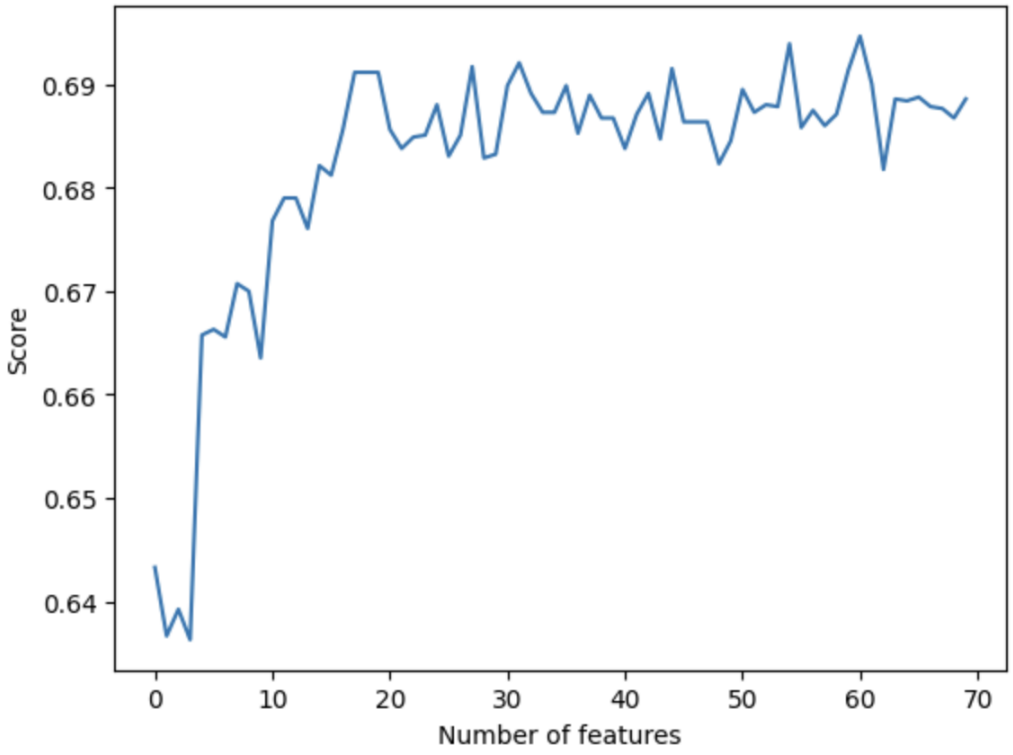

Figure 4.6: *A comparison of a XGBoost model with increasing number of features based on the RFE results.*

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, including the SOFA score subdivisions and critical raw laboratory data.

4.3 User Interface

After feature input in the user interface, the data is processed in the final model. The user is then redirected to a results page, displaying 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.

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

4.4 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.

4.5 Discrepancies in XAI Results

We observed some discrepancies in the explainable AI outputs that can be attributed to data imperfections. For instance, in our 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.

5

Discussion and Conclusion

A healthy SpO2 range for a possible sepsis patient falls in the range of 95%-100%. Since there is no clear distinction between racial demographics in the data, and all the data are within the safe SpO2 ranges, there is no bias seen from pulse oximeters. This suggests that there is no significant bias in the BOLD dataset, and the need to correct erroneous readings for patients with darker skin tones does not apply to this specific dataset.

When comparing the False Positivity Rates (FPR) and True Positivity Rates (TPR) results, 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 ML models.

The ML models trained on the BOLD dataset achieved 70% accuracy. While this is a promising score for an initial approach to the problem of sepsis prediction, it fell short of our goal of 80%. This accuracy was set with the goal of outperforming current clinical metrics in ICUs. For this reason, the model trained by our team can be used as a tool for clinicians to augment conventional sepsis prediction metrics, but should not be used alone to predict a patient's status within 24 hours upon admission to the hospital.

The insufficiencies in the accuracy of our model can be attributed to several factors. XGBoost and random forest, while good baseline machine learning models, are rather simplistic when compared to large neural networks. Additionally, the MLP used by our team was comprised of only two hidden layers and a small number of parameters. A larger network and more sophisticated activation functions may produce better results.

It is also possible that the dataset the model was trained on contributed to the poor accuracy of the model. Certain biases exist inherently in the data. A particularly concerning area of bias is the lack of diversity in the age of the subjects, as the majority were aged 50 and above. This disparity causes the data to be skewed toward older

populations.

Additionally, as data is collected upon the admission of the patient to the ICU, critical data for sepsis prediction is missing. This is an inherent limitation of instantaneous sepsis prediction, and it may be impossible with the current tools and data available to achieve accuracy near or above 95%.

The results of the global explainability analysis and RFE plots provide clinicians with valuable insights into the significance of various features in predicting sepsis. 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. For instance, clinicians could place greater emphasis on the patient's renal performance, which was ranked as more substantial, compared to the patient's liver performance, which was deemed less significant.

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 SOFA subdivision, ordering more frequent tests, or administering antibiotics or other medications proactively.

The observed discrepancies in the data, such as the contradictory indications between renal scores and creatinine levels, raise concerns about the model's performance. These issues highlight the need for a more comprehensive dataset to improve model accuracy and resolve inconsistencies. A more extensive dataset would not only address these discrepancies but also enhance the overall performance of the model.

Future steps for this project include implementing a more extensive and generalized model to achieve higher model performance and accuracy. This could involve the use of a larger MLP and the exploration of more extensive neural networks. Subsequently, a thorough bias analysis is critical to ensure the model's fairness and reliability. Furthermore, 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.

6

Contributions

6.1 Dylan Sain

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.

6.2 Kathleen Fort

Created a User Interface prototype to collect patient data for ML procesing and provide an interpretable output using local SHAP XAI plots and global SHAP, LIME, and PFI graphs.

6.3 Maria Jorge

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.

Bibliography

Adams, R. et al. (July 21, 2022). *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).

Amann, Julia et al. (Nov. 30, 2020). "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective". In: *BMC Medical Informatics and Decision Making* 20.1, 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).

Baniasadi, Atefeh et al. (Jan. 2021). "Two-Step Imputation and AdaBoost-Based Classification for Early Prediction of Sepsis on Imbalanced Clinical Data". In: *Critical Care Medicine* 49.1, 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).

Bauer, Michael et al. (May 19, 2020). "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, 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).

Böck, M. et al. (2024). *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).

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

Chen, Tianqi and Carlos Guestrin (Aug. 2016). "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], pp. 785–794. DOI: 10.1145/2939672.2939785. URL: http://arxiv.org/abs/1603.02754 (visited on 07/22/2024).

Di Martino, F. and F. Delmastro (Oct. 26, 2022). *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).

Eloranta, Sandra and Magnus Boman (Aug. 2022). "Predictive models for clinical decision making: Deep dives in practical machine learning". In: *Journal of Internal Medicine* 292.2, 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).

Evans, Tom (Apr. 1, 2018). "Diagnosis and management of sepsis". In: *Clinical Medicine* 18.2, 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).

Fleuren, Lucas M. et al. (Mar. 1, 2020). "Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy". In: *Intensive Care Medicine* 46.3, 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).

Futoma, Joseph et al. (Aug. 2017). "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, pp. 243–254. URL: https://proceedings.mlr.press/v68/futoma17a.html.

He, YiRan et al. (Dec. 1, 2023). "A machine-learning approach for prediction of hospital mortality in cancer-related sepsis". In: *Clinical eHealth* 6, 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).

Hilty, D. M. et al. (May 25, 2022). *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).

Islam, Md. Mohaimenul et al. (Mar. 1, 2019). "Prediction of sepsis patients using machine learning approach: A meta-analysis". In: *Computer Methods and Programs in Biomedicine* 170, 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).

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

Krupinski, Elizabeth A. et al. (Sept. 1, 2010). "Long Radiology Workdays Reduce Detection and Accommodation Accuracy". In: *Journal of the American College of Radiology* 7.9, 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).

Liu, Shuhui et al. (2022). "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, pp. 4258–4269. DOI: 10.1109/JBHI.2022.3171673.

Lopez, Karen Dunn and Linda Fahey (June 1, 2018). "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, 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).

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

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

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

Memon, Shaheen MZ., Robert Wamala, and Ignace H. Kabano (Jan. 1, 2023). "A comparison of imputation methods for categorical data". In: *Informatics in Medicine Unlocked* 42, 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).

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

Prasad, Varesh et al. (Nov. 1, 2023). "Diagnostic suspicion bias and machine learning: Breaking the awareness deadlock for sepsis detection". In: *PLOS Digital Health* 2.11. 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).

Ren, Shuangxia et al. (May 2022). "Machine learning based algorithms to impute PaO2 from SpO2 values and development of an online calculator". en. In: *Scientific Reports* 12.1. 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).

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

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

Singer, M. et al. (Feb. 23, 2016). *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).

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

Su, Longxiang et al. (June 28, 2021). "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. 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).

Tallgren, M., M. Bäcklund, and M. Hynninen (2009). "Accuracy of Sequential Organ Failure Assessment (SOFA) scoring in clinical practice". en. In: *Acta Anaesthesiologica Scandinavica* 53.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1399-6576.2008.01825.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).

Tutty, M. A. et al. (Apr. 9, 2019). *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).

Velden, Bas H. M. van der et al. (July 1, 2022). "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis". In: *Medical Image Analysis* 79, 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).

Venkat, Swaathi et al. (July 2019). "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, pp. 482–485. DOI: 10.1109/EMBC.2019.8856434. URL: https://ieeexplore.ieee.org/abstract/document/8856434 (visited on 07/22/2024).

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

Yan, M. Y., L. T. Gustad, and Ø Nytrø (Dec. 13, 2021). *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).

Yang, Zhenyu, Xiaoju Cui, and Zhe Song (Sept. 27, 2023). "Predicting sepsis onset in ICU using machine learning models: a systematic review and meta-analysis". In: *BMC Infectious Diseases* 23.1, 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).

University College Dublin
Arcadia University
University of North Carolina at
Chapel Hill

Augmenting SOFA Scores for Sepsis Prediction with Machine Learning

Kathleen Fort

Dylan Sain

Maria Jorge

July 2024