Real-World Designed LightGBM Sepsis Prediction

Charlie Nasiadka *jz22966@bristol.ac.uk*
Josh Featherstone *qh22044@bristol.ac.uk*
Josh Hyde *os22435@bristol.ac.uk*
Sonny Cooper *ed22699@bristol.ac.uk*
Josh Everett *td22885@bristol.ac.uk*
Sujin Subanthran *ex22760@bristol.ac.uk*

***Author Contributions*—All members have contributed to this project in an approximately equal manner.**

***Abstract*—This report explores how early detection of the onset of sepsis can be achieved through interpretable data science methods that reflect real-world ICU conditions. Using multivariate, hourly time series data from two hospital datasets comprising more than 40,000 ICU patients, we implemented a complete data pipeline, from cleaning and exploration to model development and visualisation.**

Feature engineering focused on the dynamic nature of sepsis progression, with an emphasis on change-based metrics such as deltas and clinically used medical scores to help track degradation. Statistical tools, including Jensen-Shannon Divergence and Correlation Dendrograms, were applied to guide feature selection, helping visualise mathematical trends and feature clusters.

The final output was a real-time dashboard showing a patient's sepsis risk timeline. This interface continuously estimates patient risk, flagging high-risk intervals to help support proactive clinical decision-making. At the hour of peak risk, a SHAP summary plot is generated to highlight the top five contributing characteristics, offering interpretable case-specific explanations to clinicians. Together, these tools aim to bridge the gap between algorithmic prediction and bedside utility.

I. Introduction

Sepsis is a highly lethal condition in which the body responds improperly to the presence of harmful microorganisms in the blood or other tissues. Its subtle onset and rapid progression pose a major clinical challenge. In many cases, diagnosis is delayed or missed, contributing to high mortality rates of 20–40% [1]. Every hour of delayed diagnosis increases the risk of death by 4–8% [2]. Sepsis also places a significant economic burden on healthcare systems, with annual U.S. costs estimated at approximately $20 billion [3]. As one of the most resource-intensive conditions treated in hospitals, it remains a central clinical research focus. Early identification and timely intervention are critical to improving outcomes, reducing intensive care unit (ICU) stays, and saving lives.

This project is built around the 2019 PhysioNet Computing in Cardiology Challenge, which focuses on the early prediction of sepsis using multivariate time-series data collected from intensive care units (ICUs) [4]. The two datasets provided contain vital signs, laboratory values, and demographic information, with each patient's condition annotated hourly with a binary sepsis label.

The primary motivation behind this work is to provide clinicians with interpretable, real-time insights that support early intervention and improve decision-making. We offer an informative model and dashboard, which clinicians can validate, challenge, and refine using their expertise. While recent advances in medical AI have delivered impressive predictive performance, many of these systems operate as black boxes, limiting clinical trust and adoption. There is growing recognition that medical AI must shift toward transparent and accessible tools to enhance clinical utility [5].

II. Literature Review

The official classification of sepsis is a two-point degradation in SOFA score [6]. In addition to this, during initial research, we found early symptoms of sepsis include a fever, leading to an elevated temperature, along with fluid within the lungs [7]. Combining these ideas, sepsis was a problem that could theoretically be approached from a mathematical and medical viewpoint. We explored the focuses of notable approaches from the 2019 Computing in Cardiology Challenge [8]–[12], which all come at the sepsis problem from a purely machine learning angle. These aim to maximise metrics such as area under curve (AUC) [13] and the utility score provided by the challenge. We realised there was a research gap for a medically focused and clinically applicable project. Our work builds upon previous challenge submissions, aiming to provide an algorithmic prediction and help clinicians in bedside hospital scenarios.

III. Initial Exploration

A. Dataset Overview

This study uses ICU datasets from two independent hospital systems, **Hospital A** and **Hospital B**, each containing time-series patient records with 40 physiological, laboratory, and demographic features sampled hourly. A binary indicator at each time step denotes the presence or absence of sepsis. The datasets are summarised below in *Table 1*.

TABLE I: Dataset Summary Statistics

Characteristic	Hospital A	Hospital B	Combined
Total Patients	20,336	20,000	40,336
Septic Patients	1,790	1,142	2,932
Sepsis Rate (%)	8.80	5.71	7.26
Total Observations	790,215	761,995	1,552,210
Median ICU Stay (hours)	40	38	39
Sepsis Onset Time (avg hours)	50.97	50.78	50.87
Data Completeness (%)	33.40	33.10	33.25

To simulate realistic ICU conditions, the datasets exhibit substantial heterogeneity in data completeness. Vital signs (e.g., heart rate, respiratory rate) are generally complete, whereas laboratory results (e.g., bilirubin, creatinine) are often missing due to infrequent clinical testing. In addition, in the combined dataset, from 40,336 patients, 7.27% were sepsis cases *(SepsisLabel = 1)*, and from the 1,552,210 row only 1.8% were sepsis cases. These both acted to show there was a significant class imbalance between sepsis and non-sepsis data, requiring considerations throughout the model development process.

B. Sepsis Label Smoothening

For sepsis cases, we are given a 6-hour delay between the dataset *sepsisLabel* elevating to one and the actual time of sepsis onset. This transition from *sepsisLabel = 0* to *1* occurs instantly. However, in clinical practice, it is worth noting that exact onset times are uncertain due to diagnosis delays, so a smoothening approach using continuous probabilities would likely better model real-world scenarios (*Equation 1*). While we retained binary labelling for model simplicity, we identified this as a potential area for future improvement.

$$\text{Probability}(t) = \begin{cases} 1, & t \geq t_{\text{sepsis}} \\ \exp\left(-\frac{(t_{\text{sepsis}}-t)^2}{2\sigma^2}\right), & t > t_{\text{sepsis}-6} \end{cases} \quad (1)$$

C. Jensen-Shannon Divergence

Initial checks were required to ensure significant similarity across the two datasets, allowing them to be combined. The Jensen-Shannon Divergence [14] is plotted in *Figure 1*, based on the Kullback–Leibler divergence, a method used to measure the similarity between two probability distributions. Most features showed low JSD values (below 0.1), indicating high alignment and confirming their similarity. Features such as *Temp* and *BaseExcess* had somewhat elevated JSD values, potentially because the hospitals had different frequencies with which they measured these or specific clinical practices, which were reviewed again during feature selection. Ultimately, no features ended up being filtered based on JSD but these checks helped provide confidence in integration.

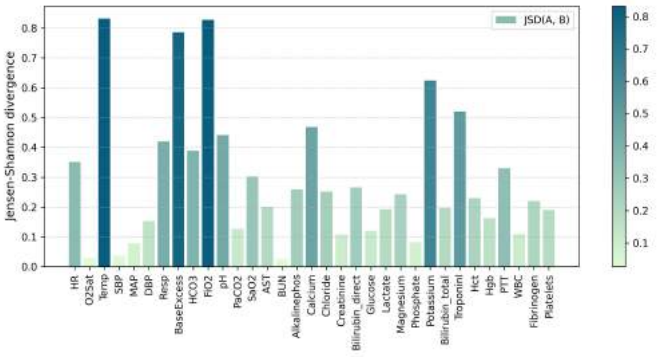

Fig. 1: Jensen-Shannon Divergence - Hospital A and B

D. Dataset Auto-Correlation

For training models, highly correlated features are standardly removed. This is because they contain similar information, so they act to increase model complexity with no real gain in predictive power and can lead to overfitting instead. We plotted a correlation heat map within the dataset in *Figure 2*, focusing on features above 0.25 auto-correlation. We selected this threshold to handle the sensitivity to linear relationships and minimise noise, finding many that fit this criterion. Based on these findings we suspected further checks would need to be carried out to reduce multicollinearity in our dataset.

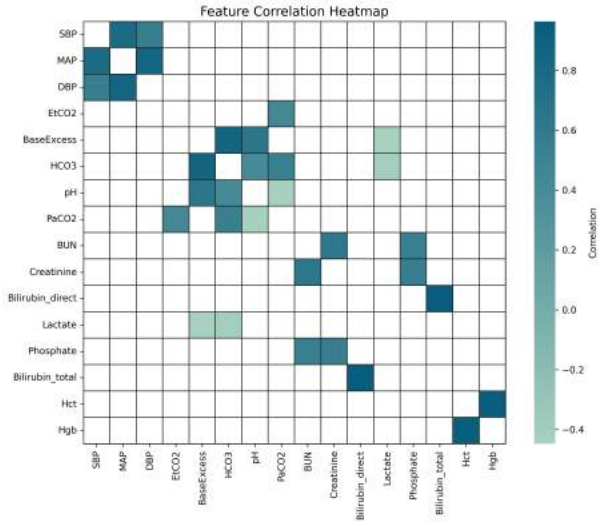

Fig. 2: Feature Correlation Heat Map

Most grid spaces were blank, with most features only highly correlated with specific others. We spotted notable correlation relationships, including examples such as MAP, SBP, and DBP, as well as *Hct* and *Hgb*. A correlation dendrogram in *Figure 3* helped to further explore these relationships, showing hierarchy using Ward distance [15]. This hierarchical clustering showed features we could safely aggregate or remove, keeping the majority of information and pruning redundant features.

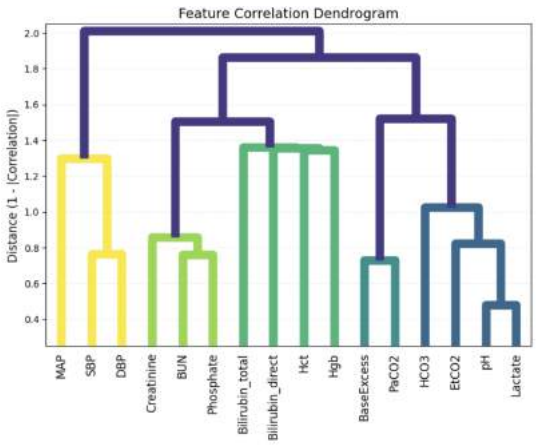

Fig. 3: Feature Correlation Dendrogram

E. Temporal Trends

Finally, we observed the temporal relationship for certain features, noticing specifically that the most continuous features, like heart rate, showed shifted averages and greater volatility for patients with sepsis in comparison. In *Figure 4*, both the average HR and the standard error associated with that value (SEM) [16] are plotted for each value of the intensive care length of stay (ICULOS). The SEM helpfully shows the variability across the population for each time step. The plot shows that sepsis patients with prolonged ICU stays often exhibit extreme heart rates, as well as similar other features, implying instability increases in latter-stage sepsis. This hinted at the usefulness of comparing changes in such features, not just their absolute values, namely delta values discussed later.

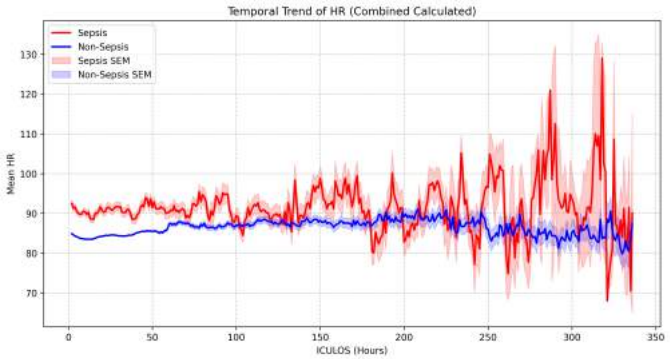

Fig. 4: Heart Rate Temporal Trends

F. Additional Methods Trialled

We also carried out additional exploration methods that, although sometimes suggested in professional practice and lectures, led to no fruitful conclusions. These included the Mann Whitney U-test [17] for correlation; non-continuous means; Box and Whisker Plots [18] and Linear Correlation metrics (Pearson [19], Spearman [20] and Kendall [21]). As an example, this discussed linear correlation for the 'continuous' features is shown in *Figure 5*. The values being no greater in magnitude than 0.05 highlights that the sepsis relationships are more complicated than simply linear. Finally, although in machine learning practice, PCA is often used, the extensive amount of missing data in key features didn't allow for this.

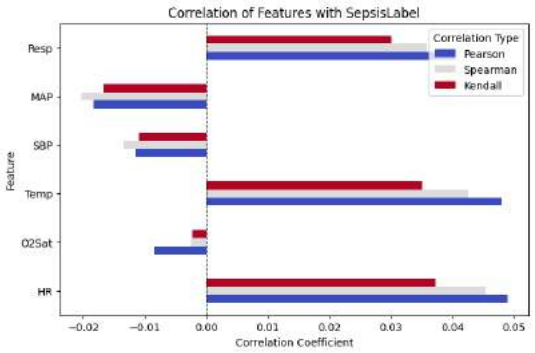

Fig. 5: Feature correlation

IV. Data Preparation

A. Medical Equations

We had suspicions that mathematical relationships underpinned the highly correlated features in the dendrogram (*Figure 3*). After research [22], [23], [24], [25] we found several clinically used mappings defined below in *Equations 2,3,4,5*.

$$MAP = \frac{SBP + 2 \times DBP}{3} \tag{2}$$

$$Hct \approx 2.94 \times Hgb \tag{3}$$

$$pH = 6.1 + \log_{10}\left(\frac{HCO_3}{0.03 \times pCO_2}\right) \tag{4}$$

$$BE \approx HCO_3 - 24.4 + (2.3 \times (Hgb - 7.4)) \tag{5}$$

After identifying these relationships, we needed to establish whether they aligned with our previous data. Comparing equation probability distributions to rows containing all equation features confirmed this trend, as shown in *Figure 6*.

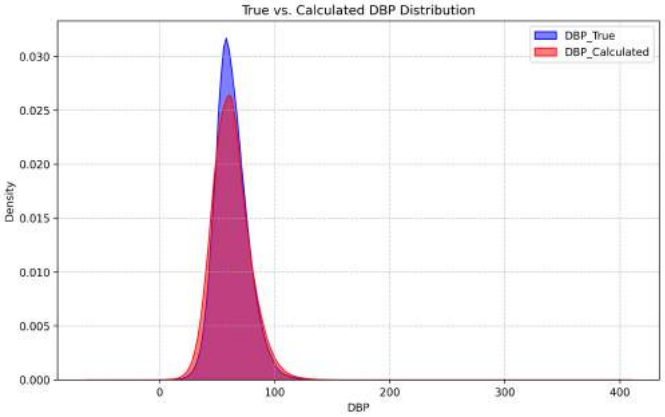

Fig. 6: True vs Imputed Distributions

These allow for clinically grounded imputation, which is far more accurate than statistical methods. In particular, *Equation 2* was a particularly eye-opening discovery, as for **Hospital A**, DBP had a proportionally increased missingness compared to both SBP and MAP, allowing those values to help fill in DBP. The results of clinical equation imputation are displayed below in *Figure 7*, showing marked missingness reductions.

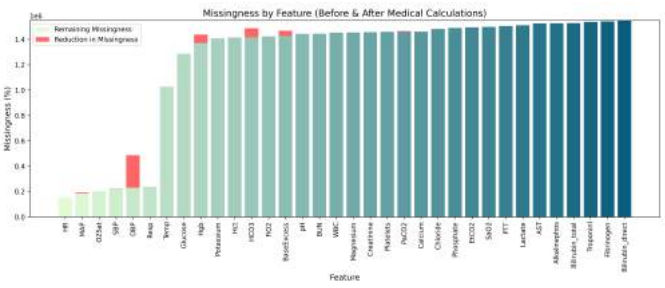

Fig. 7: Medical Missingness Reduction

However, medical imputation had to be handled carefully, as this had the potential to introduce artificial correlation and amplify feature redundancy when training models. In particular, when we initially implemented both XGBoost [26] and LightGBM [27], we discovered that they handle NaN values natively so excessively reducing missingness was unnecessary. It is also often practiced that features with above 95% missing data are discarded; however, because these models rigorously handle NaN values, we opted not to employ this approach.

B. Simple Medical Scores

We also explored feature engineering using established equations deriving new features including Shock Index, Pulse Pressure, SaO_2/FiO_2 ratio, BUN/Creatinine ratio and various clinical flags. An example is shown below in *Equation 6*.

$$ShockIndex = \frac{Heart\ Rate\ (HR)}{Systolic\ Blood\ Pressure\ (SBP)} \tag{6}$$

C. Complex Medical Scores

Alongside these, we harnessed the power of medically used patient health scores, such as SIRS, to help show the state the patient is currently in. Many of these scores are made up of feature considerations we lacked in our dataset, namely qSOFA, SOFA, MEWS, and NEWS2, for which we implemented partial versions of these scores best using the features we had. An example is displayed below in *Equation 7* showing the components for the PartialNEWS2 engineered score.

$$S_{\text{PartialNEWS2}} = S_{\text{Resp}} + S_{\text{O2Sat}} + S_{\text{Temp}} + S_{\text{SBP}} + S_{\text{HR}} + S_{\text{FiO}_2} \tag{7}$$

Further breaking this down, a single component is fleshed out in *Equation 8*, where the margins for the temperature component are demonstrated.

$$S_{\text{Temp}} = \begin{cases} 3, & \text{if Temp} \leq 35.0 \\ 1, & \text{if } 35.0 < \text{Temp} \leq 36.0 \\ 1, & \text{if } 38.0 < \text{Temp} \leq 39.0 \\ 2, & \text{if Temp} > 39.0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

D. Imputation

We explored various imputation methods to deal with the high level of missing data within our dataset. Broadly, forward and backward fill might be more appropriate for the sparse laboratory values. In contrast, linear interpolation was speculated to be more effective for the more complete and continuous vital signs, offering smoother and more physiologically realistic estimates between observed values. In practice, we used a combination of the previously discussed methods called mixed imputation. This assigns forward-fill, backfill, or linear interpolation for each feature individually based on the highest absolute correlation with the sepsis label across the dataset. An example generic depiction of this process is displayed in *Figure 8*.

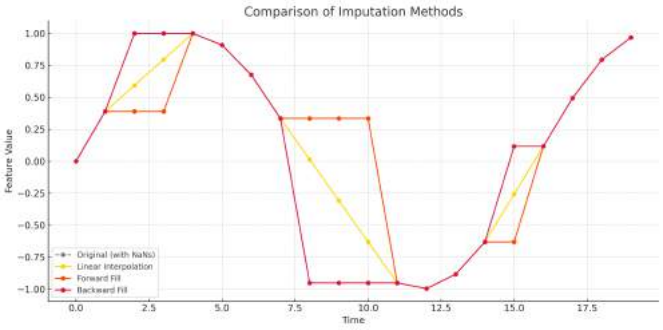

Fig. 8: Mixed Imputation

We also considered machine learning-based imputation methods, such as MiceForest, an approach that uses random forests to perform multiple imputations by chaining equations (MICE) [28]. However, this was ineffective due to the highly sparse data and highly computationally expensive for our large dataset. Additionally, we experimented with Kalman filtering, a time-series imputation technique that estimates missing values based on prior and observed states [29]. Although the dataset exhibits temporal continuity, Kalman filtering was ineffective due to the sparsity and irregular sampling intervals, which limited its ability to reliably estimate missing values.

E. Delta Values

As prior mentioned in *Section 3.E*, we noted that spotting changes in feature values may be as important as the values themselves. This led us to explore the use of delta values, effectively applying a backward sliding window over previous time steps. We only looked to apply this to the seven 'continuous' features up until Temp on the missingness graph (*Figure 7*), as these have enough data points to show regular changes. After exploring many window sizes and combinations for this problem, we found that a single backward window of three time steps (*Equation 9*) best supplemented our approach and improved our model. This improvement was likely due to now being able to catch the quick speed of health deterioration as a patient becomes septic, as mentioned earlier regarding the increased rate of fatality of 4–8% [2] with every hour delay.

$$D_t^{(w=3)} = \text{stat}\left(X_{t-w+1}, \ldots, X_t\right) \tag{9}$$

The official classification of sepsis is a two-point degradation in SOFA score [6], with deltas designed to find these patterns of change that complex scores are unable to capture at a single point in time. We explore many window summary techniques such as Exponentially Weighted Moving averages (EWM) [30], Kurtosis [31], and Second Order Derivatives. However, through experimentation comparing the final model accuracies, we simplified using Mean, Standard Deviation, First Order Derivatives, and Slope [32].

V. Model Metrics

For clarity, we define the following standard terms. To help visualise we have included *Figure 9* by JC Chouinard [33].

True Positives (TP): Sepsis classified as sepsis.
False Positives (FP): Non-sepsis classified as sepsis.
False Negatives (FN): Sepsis classified as non-sepsis.
True Negatives (TN): Non-sepsis classified as non-sepsis.

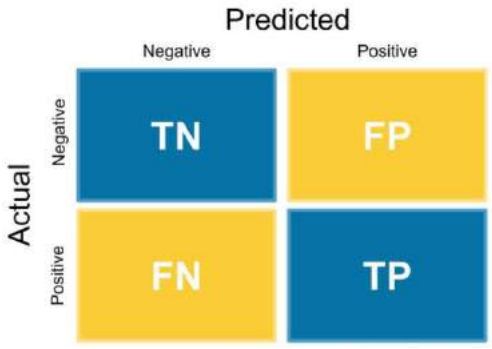

Fig. 9: Confusion Matrix

A. AUC

As mentioned in the literature review, many papers used only AUC and the PhysioNet challenge utility score. AUC, specifically, is a mathematical metric derived from the ROC curve. This plots the true positive rate (TPR) against the false positive rate (FPR) over every decision threshold, for which the area under that curve is then calculated. TPR and FPR equations are included below in *Equation 10*.

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN} \tag{10}$$

In our case, we just wanted to increase the number of correct sepsis predictions (*SepsisLabel = 1*) while not biasing and sacrificing non-sepsis cases (*SepsisLabel = 0*). Although indicative of model performance, AUC acts to overcomplicate this and could not explain easily to clinicians how confidently our model predicts sepsis. We wanted to supplement this with a more straightforward metric.

B. Ignoring Precision

We will now look specifically at precision, whose equation for *SepsisLabel = 0* is shown in *Equation 11*.

$$Precision = \frac{TN}{TN + FN} \tag{11}$$

For our project, due to the significant sepsis class imbalance, precision for the *SepsisLabel = 0* case is always approximately 1. This is because even if a few septic patients are wrongly classified as non-septic (FN), the sheer number of correctly classified non-septic cases (TN) far outweighs this. For this reason, the precision metric is not desired in any of our calculations.

C. F_β Score

Unlike precision, recall *(Equation 12)* compares the correctness of each case individually, not suffering due to the class imbalance. We explored optimising our model's probability threshold for sepsis identification using F_β scoring [34]. This, like the commonly used F1 scoring, balances the outputs between the precision and recall output metrics, and its calculation is also shown in *Equation 13* below. Being able to drive up the β value leans the prediction to prioritise recall over precision.

$$Recall = \frac{TP}{TP + FP} \tag{12}$$

$$F_\beta = \frac{1 + \beta^2 \times Precision}{\beta^2 \times Precision + Recall} \tag{13}$$

D. Custom Recall Score

As F_β score naturally still includes some precision consideration, previously discussed to be non-informative, we decided to build a custom scoring metric for purely recall-related considerations instead. This simultaneously tries to maximise the mean recall across both *SepsisLabel* values while minimising the difference between them to balance equally. This is highlighted in *Equations 14-18*.

$$r_0 = Recall_{SepsisLabel=0} \tag{14}$$

$$r_1 = Recall_{SepsisLabel=1} \tag{15}$$

$$MeanRecall = \frac{r_0 + r_1}{2} \tag{16}$$

$$RecallDiff = |r_0 - r_1| \tag{17}$$

$$RecallScore = MeanRecall - RecallDiff \tag{18}$$

VI. Data Modelling

A. Gradient Boosting Models

We considered a range of modelling approaches, ultimately deciding that tree-based gradient boosting methods would be most appropriate due to their strong performance on tabular data, generally outperforming deep-learning-based methods. The commonly used options, XGBoost and LightGBM, natively handle missing values, which is critical in a medical context where missing data is common. Furthermore, they provide interpretable outputs (such as SHAP scores), which makes them particularly well-suited in a healthcare context, where transparency and interpretability are essential. LightGBM offers faster training speeds, making it well-suited for iterative fine-tuning and rapid experimentation. It also has greater potential predictive accuracy due to its leaf-wise tree growth, though this comes with a risk of overfitting. In contrast, the slower XGBoost offers greater robustness to overfitting, with built-in regularisation strategies that help stabilize learning. Given that both models are strong candidates for the task of sepsis classification, we chose to experiment with both. This allowed us to compare their respective strengths and limitations in the context of our dataset, helping us arrive at a more informed and optimal solution.

B. Class Imbalance

Both models are also able to easily handle class imbalance, allowing for penalisation of misclassification of the minority class (septic patients). This is done by scaling the positive weight in training based on the ratio of positive to negative cases, which is done row-wise and shown in *Equation 19*. Here, the negative samples are the number of patient rows where *SepsisLabel = 0* and vice-versa.

$$scale\_pos\_weight = \frac{NegativeRows}{PositiveRows} = \frac{1218366}{22327} \approx 54.56 \tag{19}$$

C. Training Methodology

Our model training followed the standard 80/20 split between training and testing data. Stratification by sepsis label was used to ensure both sets preserved the original prevalence of sepsis cases. Our XGBoost and LightGBM models were trained using logistic loss due to their practical application in binary classification problems. We optimised hyperparameters with Optuna [35] to fine-tune our model, targeting our custom recall-based scoring metric.

D. Failed Training Considerations

Throughout the final phase of our project, where we refined our modelling approach, we passed through many iterations using different techniques. Firstly was patient-wise upsampling using SMOTE [36], which uses the nearest neighbours for each minority sample (*SepsisLabel = 1*) to generate similar but modified synthetic samples. This, although functional, showed no model improvement, likely due to increases in input complexity and potential overfitting to non-relevant trends in the minority samples. Similarly, patient-wise downsampling also showed significant issues, with the model becoming extremely biased towards *SepsisLabel = 1* and losing predictive power for the majority case. Finally, both Ensemble Methods [37] and K-fold Cross-Validation [38], although taught during many University of Bristol units and often recommended Data Science practice, ended up overcomplicating the simple but effective tree-based gradient boosting methods, leading to reduced performance and increased complexity.

E. ICULOS Inclusion

As a widely disputed topic we observed in the Sepsis Challenge research, our last conversation before training our models was whether the ICULOS feature should be included. This feature denotes the time since a patient has been escalated and placed in intensive care. Since a patient's continued stay in ICU implies they are still extremely unwell, we were concerned whether or not this time-based feature may introduce a model bias towards all longer stay cases or be a data leak. However, given our main application is providing information to help decisions made by real clinicians, who by definition know the duration of a patient's ICU stay, we could instead act to harness the predictive power of ICULOS. In the *Findings* we highlight this improvement's significance.

F. LightGBM Model

As will be seen in later comparisons, LightGBM had minor improvements in metrics across the board compared to XGBoost. Therefore, for completeness, a diagram of this model, taking into account its hyperparameters post Optuna post-processing, is shown in *Figure 10* below. The maximum depth parameter was selected to ensure that complex relationships could be observed without overcomplicating. The low learning rate helped the model converge while not overfitting and adapting to non-existent trends. We combine the output from a sequence of decision tree estimators, resembling the core of gradient boosting [39].

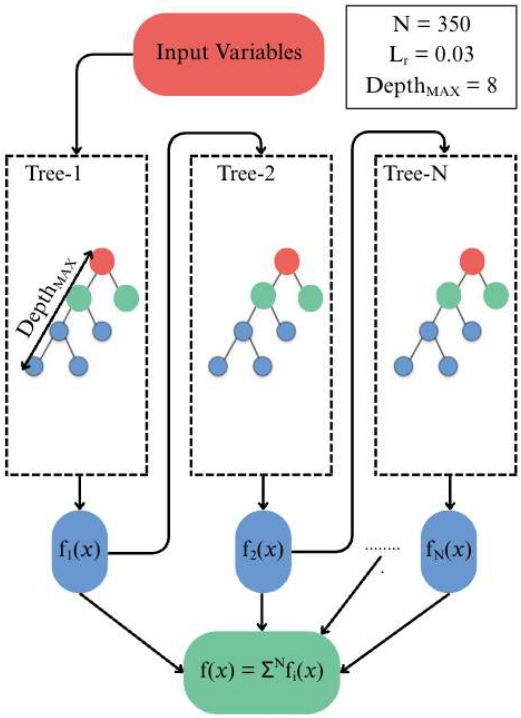

Fig. 10: Optimised LightGBM Model

This model alone allows for robust predictions but performs even better when integrated into a structured and rigorous pipeline, as described in the following section.

VII. Findings

A. Optimised Final Pipeline

The data passed through a distinct sequence of stages in our final pipeline to most effectively train. This began with the raw data dictionary indexed by the patient and ended with the final trained model. Key stages included delta feature engineering, class balancing, and hyperparameter tuning. This pipeline was perfected by rigorous examination of the output metrics, guiding targeted adjustments. The whole pipeline is depicted below in *Figure 11*.

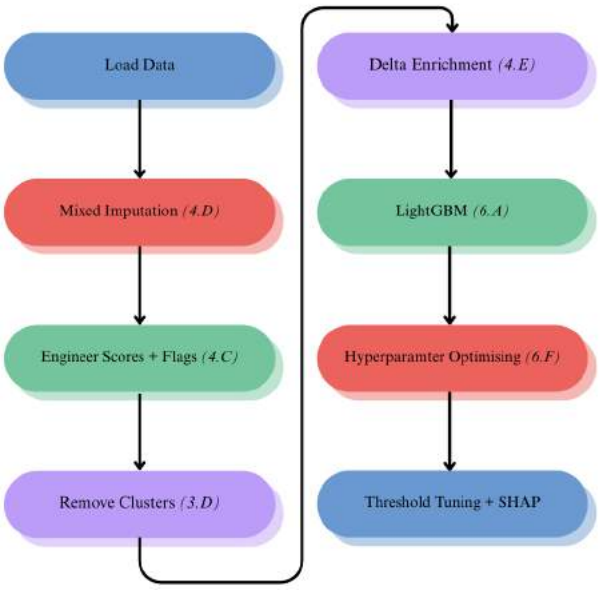

Fig. 11: Final Pipeline

B. Checkpoint Model Metrics

As we progressed through model development, iteratively examining model metrics as previously mentioned in *Section 7.A*, key checkpoints were stored to benchmark our progress, shown below in *Figure 12*.

Model	Recall Score (2.d.p)	AUC (2.d.p)
XGBoost (No ICULOS)	0.74	0.83
XGBoost (ICULOS)	0.77	0.85
LightGBM (No ICULOS)	0.75	0.83
LightGBM (ICULOS)	0.78	0.87
LightGBM (ICULOS + Deltas)	0.82	0.90

Fig. 12: Metric Checkpoints

C. Final AUC and Recall Score

As depicted in *Figure 12*, the final AUC of our system is 0.90, which indicates the model has a strong ability to classify sepsis and non-sepsis cases and carries high mathematical significance. An AUC of 0.50 would imply the model is no better than random guessing. This analysis method also benefits from the visual ROC curve plotted below in *Figure 13*.

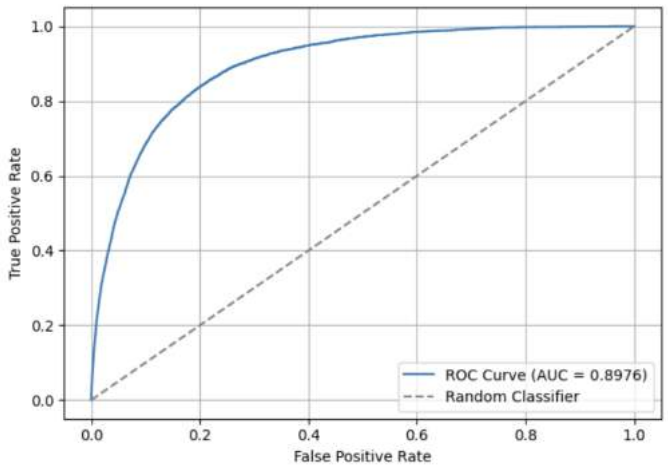

Fig. 13: ROC Curve

Alongside this standard metric, our custom Recall Store reached a maximum of 0.82, equating to a confidence score of **82%** when classifying sepsis cases, which we could quote to clinicians and medical professionals. Individual FP, TP, FN and TN values for our 20% test split that lead to the confidence above are best displayed in a Confusion Matrix [40] depicted in *Figure 14* below.

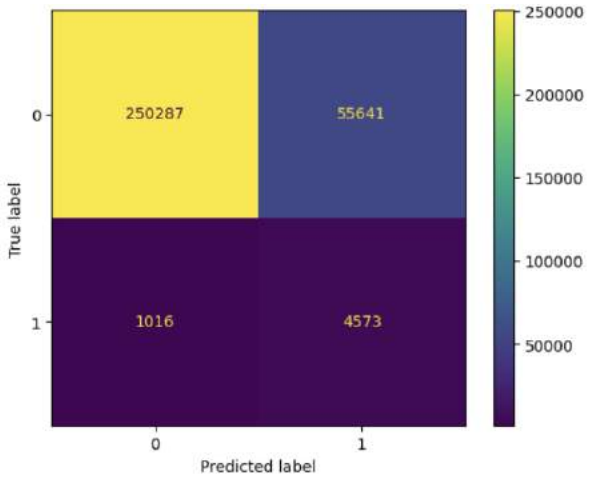

Fig. 14: Confusion Matrix

D. Feature Importance

For our final LightGBM model, many known ways exist to display which features were most contributive to the probability distribution output. Namely, the two most common include the built-in feature importance and the SHapley Additive exPlanations (SHAP) values [41]. In particular, the Beeswarm SHAP depiction carries the greatest interpretability. The vertical axis displays the feature importance ranking, whereas the horizontal shows whether each normalised SHAP value decreased or increased the predicted sepsis probability. This is constructed by adding a single point for each feature across every patient in the test data split and shown below in *Figure 15*.

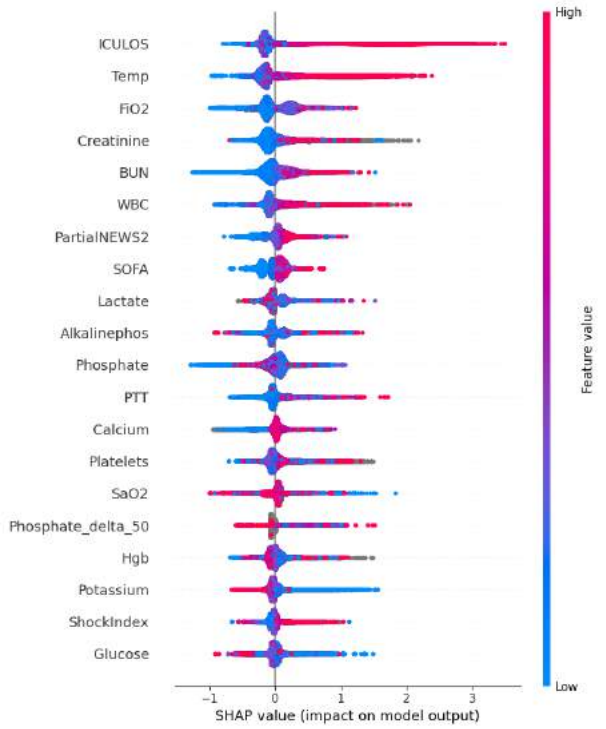

Fig. 15: LightGBM SHAP Scores

It can be seen that, logically, ICULOS is the most impactful feature, with more extended stays implying that the patient is septic. *Temp*, FiO_2, and *Creatinine* level are the next most impactful predictors, also noted as clinically sensible indicators of sepsis, indicating that the model effectively identifies essential relationships in the data. A hallmark sign of sepsis is rising body temperature [42], highlighted by its high positive SHAP value, manifesting with symptoms akin to a fever as the body tries to fight systemic infection. Sepsis can significantly reduce Respiratory function, often even progressing to acute respiratory distress syndrome (ARDS) [43], shown by the predictive nature of high FiO_2, which suggests the patient needs oxygen supplementation. Finally, elevated creatinine levels, shown on *Figure 15* by positive SHAP values, signal acute kidney injury (AKI) [44], one of the main deteriorations spotted by SOFA. These also combine to show the significant destructive nature of sepsis, targeting organ function across the body's core systems.

E. Experimenting With Maximising Sepsis Recall

During model development, we aimed to maintain a balanced recall between septic and non-septic cases to avoid bias or prioritise either. Despite this, we also experimented with maximising recall for *SepsisLabel = 1* cases to 1.0, reflecting ethical considerations in a real-world clinical setting where missing a sepsis case could have hugely serious consequences. However, this could come at the cost of significantly increasing false positives, leading to a lot of unnecessary resource allocation, which would not be feasible for most medical institutions, particularly the UK's NHS [45].

VIII. Visualisation

A. Overview

The primary goal of this project was to develop a dashboard that provides real-time insights into sepsis risk for clinicians while highlighting the key features contributing to each prediction. Harnessing the predictive power of our model and integrating it into a Streamlit [46] application yielded excellent results.

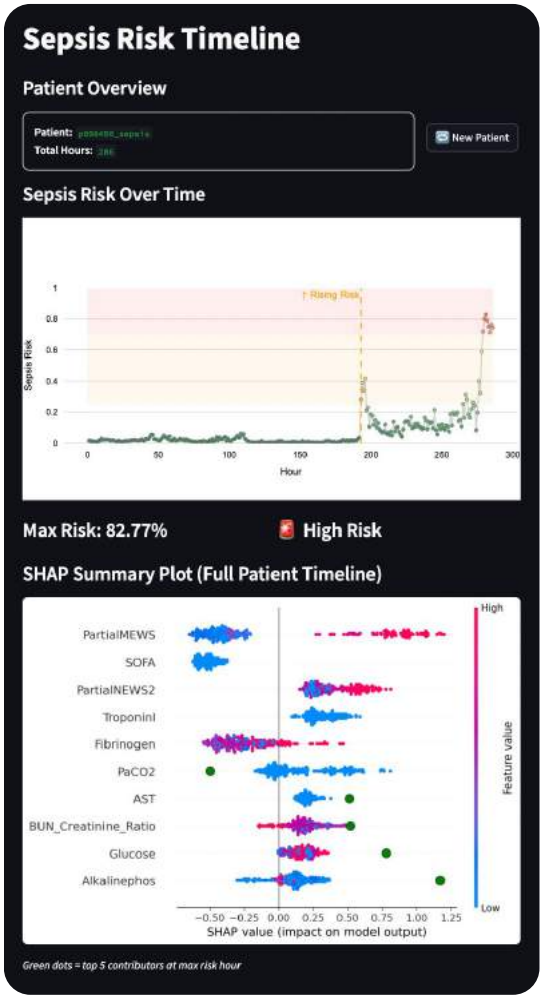

Fig. 16: Live Sepsis Risk Dashboard

B. Dashboard Content

Sepsis Risk Over Time: Displays a time series of the model's sepsis risk predictions based on live patient data, enabling healthcare professionals to quickly assess a patient's condition and detect early warning signs they might not have otherwise. **SHAP Summary Plot:** Enhances interpretability by illustrating the most influential features contributing to the historical and maximum sepsis probabilities. This allows healthcare professionals to cross-reference the model's insights with their expertise, facilitating more informed decision-making and tackling the problem of a black-box system being hard to trust.

IX. Transformers

The structure of the patient's hourly data lends itself to a transformer model [47]. However, unlike textural data and other sequential structures, with which the transformer performs well, this dataset has a high level of missingness, something transformers perform poorly with, and so imputation was required to produce a more friendly dataset at the price of precision. A hybrid method was also attempted (*Section 9.B*), where leaf values were encoded as inputs to a transformer. This would handle the missing values and provide dense representation for the model. Data imbalances were another challenge of this method. Passing in the raw data led to the model playing it safe by overly predicting "not sepsis". This was addressed through weighted loss, ensuring that more attention was paid to the "sepsis" cases. Training this model was computationally costly; time limited the level of hyperparameter tuning that could be done, as new iterations took hours. To combat this, the GPU was leveraged for model training and batch sizes were subsequently increased. Early stopping conditions were put in place, halting training if validation loss did not improve for 5 consecutive epochs, reducing possible overtraining; however, the condition was reached only 38 epochs into training, likely due to the imputation reducing noise and therefore degrading feature quality, leading to a less robust and underfitted model.

A. Stand Alone Transformer

The stand alone transformer performed okay with non-sepsis cases, achieving a recall of 0.750. However, even after implementing weighted loss to discourage the model from 'playing it safe', the recall for sepsis cases remained low at 0.667, probably due to the extensive imputation required to combat the lack of an inbuilt missing data system, smoothing out differences between "non-sepsis" and "sepsis" features and thereby the models ability to detect more subtle indicators of sepsis.

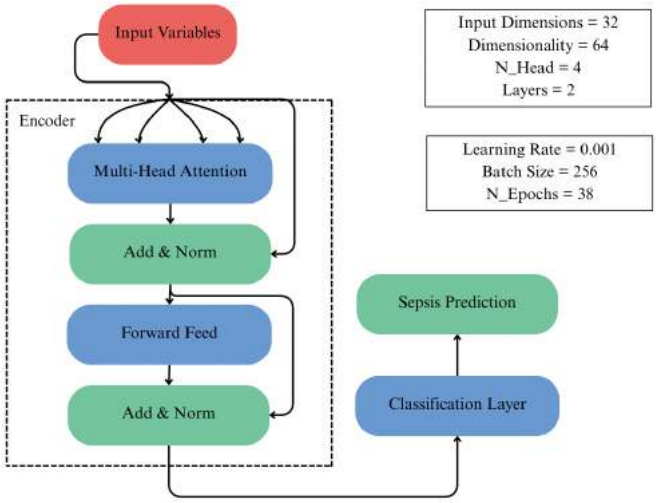

Fig. 17: Transformer Model

B. Hybrid Transformer

Alongside the stand alone model, we attempted a hybrid between the LightGBM and a transformer model. As mentioned previously, transformers do not handle missing data by default. Pretraining a LightGBM model on the data, and then encoding the leaf nodes and passing these nodes, with the corresponding label, to the transformer could increase the metric performance. We thought the transformer might be able to take the key features extracted by the LightGBM model and find hidden trends. Unfortunately, the model did not perform well. This may be because the model overfit to the training data, due to the model being pre-trained on specific LightGBM leaf encodings.

X. Discussion & Conclusion

As we have shown, this project clearly allows us to detect early the onset of sepsis, using methods and providing outputs that reflect real-world ICU conditions. During data preparation, we calculated clinical scores such as SOFA, closely aligning to the tried and tested processes doctors are taught to follow. We were able to effectively select a suitable model to handle the occurrence of missing data, allowing for disparities and mistakes in real doctors using our system, while reading live patient values. Using LightGBM, our model achieved an AUC of 0.90 and a recall of 0.82, showing strong potential for accurate sepsis prediction crucial in the front-line of hospitals. This high accuracy also acts to minimise the false positive rate, taking the load off a stretched health care system and reducing unnecessary treatments. Ultimately, this project delivers a valuable dashboard that aligns with end-users, offering clinicians a reliable, transparent and easily usable system for early sepsis prediction. With a joint focus on predictive accuracy and interpretability, clinicians can rely on our trustworthy machine learning approach to make well informed and crucial decisions for vital patient outcomes.

XI. Future Work

The Sepsis Six is a protocol designed to help treat and improve sepsis survival rate. It includes delivery of broad-spectrum antibiotics one hour before sepsis recognition and often even before infection pathogen confirmation [48]. Although this approach is known to save lives, it can lead to the use of antibiotics in cases that turn out to not actually be bacterial infections. Broad-spectrum antibiotics are known to accelerate the development of a specific anti-biotic resistance, namely Antimicrobial Resistance (AMR) [49]. This is a growing medical dilemma, where overusing specific antibiotics causes a significant reduction in their effectiveness. There is vast potential for our approach to be scaled up and used to help avoid widespread unnecessary empirical treatment, only urgently treating correctly diagnosed septic patients.

References

[1] A. K. Teng and A. B. Wilcox, "A review of predictive analytics solutions for sepsis patients," *Applied Clinical Informatics*, vol. 11, no. 3, pp. 387–398, 2020.

[2] R. D. Kumar A, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Crit Care Med*, vol. 34, no. 6, pp. 1589–1596, 2006.

[3] A. Pfuntner, L. M. Wier, and C. Steiner, "Costs for hospital stays in the United States, 2011," Agency for Healthcare Research and Quality, HCUP Statistical Brief 168, 2013. [Online]. Available: https://www.hcup-us.ahrq.gov/reports/statbriefs/sb168-Hospital-Costs-United-States-2011.pdf

[4] M. A. Reyna, C. S. Josef, R. Jeter, S. P. Shashikumar, M. B. Westover, S. Nemati, G. D. Clifford, and A. Sharma, "Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019," *Critical Care Medicine*, vol. 48, no. 2, pp. 210–217, 2020.

[5] D. Misra, V. Avula, D. M. Wolk, H. A. Farag, J. Li, Y. B. Mehta *et al.*, "Early detection of septic shock onset using interpretable machine learners," *Journal of Clinical Medicine*, vol. 10, no. 2, p. 301, 2021.

[6] M. Singer, C. S. Deutschman, C. W. Seymour *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.

[7] B. K. Harris RL, Musher DM *et al.*, "Manifestations of sepsis," *Arch Intern Med*, vol. 147, no. 11, pp. 1895–1906, 1987.

[8] J. Morrill, A. Kormilitzin, A. Nevado-Holgado, S. Swaminathan, S. Howison, and T. Lyons, "The signature-based model for early detection of sepsis from electronic health records in the intensive care unit," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[9] J. A. Du, N. Sadr, and P. de Chazal, "Automated prediction of sepsis onset using gradient boosted decision trees," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[10] M. Zabihi, S. Kiranyaz, and M. Gabbouj, "Sepsis prediction in intensive care unit using ensemble of xgboost models," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[11] X. Li, Y. Kang, X. Jia, J. Wang, and G. Xie, "Tasp: A time-phased model for sepsis prediction," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[12] J. Singh, K. Oshiro, R. Krishnan, M. Sato, T. Ohkuma, and N. Kato, "Utilizing informative missingness for early prediction of sepsis," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. 1–4.

[13] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[14] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[15] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[16] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 8th ed. Ames, Iowa: Iowa State University Press, 1989.

[17] H. B. Mann and D. R. Whitney, *On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other*. Institute of Mathematical Statistics, 1947, vol. 18, no. 1.

[18] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, no. 1, pp. 12–16, 1978. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00031305.1978.10479236

[19] K. Pearson, *Note on Regression and Inheritance in the Case of Two Parents*. Royal Society, 1895, vol. 58.

[20] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[21] M. G. Kendall, *A New Measure of Rank Correlation*. Oxford University Press, 1938, vol. 30, no. 1/2.

[22] J. E. Hall, *Guyton and Hall Textbook of Medical Physiology*, 14th ed. Elsevier Health Sciences, 2020, mAP = (SBP + 2 × DBP) / 3.

[23] C. A. Burtis, E. R. Ashwood, and D. E. Bruns, "Tietz textbook of clinical chemistry and molecular diagnostics," *Elsevier Health Sciences*, 2012, hct 3 × Hgb is a commonly used approximation in clinical labs; 2.94 is a more precise factor.

[24] K. A. Hasselbalch, "Die berechnung der wasserstoffzahl des blutes aus der freien und gebundenen kohlensäure desselben, und die sauerstoffbindung des blutes als funktion der wasserstoffzahl," *Biochemische Zeitschrift*, vol. 78, pp. 112–144, 1916, classic derivation of the Henderson-Hasselbalch equation.

[25] O. Siggaard-Andersen, "The acid-base status of the blood," *Munksgaard*, 1974, provides the approximation formula for base excess involving HCO3 and hemoglobin.

[26] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794.

[27] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[28] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011. [Online]. Available: https://www.jstatsoft.org/v45/i03/

[29] L. Kleeman, "Understanding and applying kalman filtering." [Online]. Available: https://www.cs.cmu.edu/~motionplanning/papers/sbp_papers/kalman/kleeman_understanding_kalman.pdf

[30] J. D. Hunter, M. Droettboom, T. A. Caswell *et al.*, "Moving average techniques and exponentially weighted moving averages," *Matplotlib Documentation*, 2011, https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.ewma.html.

[31] D. N. Joanes and C. A. Gill, "Comparing measures of sample skewness and kurtosis," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 47, no. 1, pp. 183–189, 1998.

[32] P. Teetor, "Slope calculations in signal analysis," *The R Cookbook*, 2011.

[33] J. Chouinard. (2022) Confusion matrix in scikit-learn. Accessed: 2025-04-15. [Online]. Available: https://www.jcchouinard.com/confusion-matrix-in-scikit-learn/

[34] C. J. van Rijsbergen, *Information Retrieval*, 2nd ed. London: Butterworth-Heinemann, 1979.

[35] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2019, pp. 2623–2631.

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[37] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. CRC Press, 2012.

[38] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," *Encyclopedia of Database Systems*, pp. 532–538, 2009.

[39] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[40] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[41] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2017, pp. 4765–4774.

[42] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.

[43] B. T. Thompson, R. C. Chambers, and K. D. Liu, "Acute respiratory distress syndrome," *New England Journal of Medicine*, vol. 377, no. 6, pp. 562–572, 2017.

[44] J. A. Kellum, N. Lameire, and for the KDIGO AKI Guideline Work Group, "Diagnosis, evaluation, and management of acute kidney injury: a kdigo summary," *Critical Care*, vol. 17, no. 1, p. 204, 2013.

[45] National Health Service (NHS), "The nhs long term plan," 2019, accessed: 2025-04-15. [Online]. Available: https://www.longtermplan.nhs.uk/publication/nhs-long-term-plan/

[46] S. Inc., "Streamlit: The fastest way to build and share data apps," https://streamlit.io, 2023, accessed: 2025-04-15.

[47] S. Kim, J.-H. Sim, and J. C. Park, "Transformers in time-series analysis: A tutorial," *Circuits, Systems, and Signal Processing*, vol. 43, no. 4, pp. 1847–1879, 2024.

[48] R. Daniels, "Surviving the first hours in sepsis: getting the basics right (an intensivist's perspective)," *Journal of Antimicrobial Chemotherapy*, vol. 66, pp. ii11–ii23, 2011.

[49] C. L. Ventola, "The antibiotic resistance crisis: part 1: causes and threats," *Pharmacy and Therapeutics*, vol. 40, no. 4, pp. 277–283, 2015.

Introduction

In our project group, I took on the key role of both leading our organisation and spearheading development, as well as contributing a significant majority to the report writing. Our project aimed to help use clinical data to predict sepsis, especially tailored to making an application with real medical applications.

Personal Contribution

My contributions were widespread across all areas in the Data Science pipeline. During our **Literature Review,** I confirmed the lack of research on developing a medically focused solution to Cardiology Challenge 2019. During **Data Exploration**, my main work was implementing the study into autocorrelation *(Section 3D)*, spotting key clusters in the heatmap and further confirming them in the dendrogram. I also helped plot the Jensen-Shannon Divergence *(Section 3C)* between the hospitals and decided to merge the hospital datasets. Following on from this, as part of **Data Preparation,** I progressed the correlation discussion to spot the key medical equations *(Section 4A)* and their potential use to reduce the total missingness through medical-based imputation, although deciding their correlative nature would reduce model performance. I helped explore key clinical scores such as SOFA *(Section 4C)* and implemented code for feature-based mixed imputation *(Section 4D)*. Another major contribution of mine was developing backwards-looking delta values *(Section 4E)* allowing for the spotting of temporal trends and to spot rapidly changing health conditions. My work with the idea for including **Section 5**, in particular a custom recall score to supplement AUC, allowed us to maximise and balance sepsis and non-sepsis cases. During **Data Modelling,** my work including the decision to not emit ICULOS for our real-world application and the improvement of my initial XGBoost model using LightGBM brought real fruit. The final **Findings** including my proposed pipeline *(Section 7A)* and my use of checkpoints in model metric analysis *(Section 7B)* led us to an outstanding final AUC of 0.90 and the highlighted confusion matrix display. My independent work building the **Visualisation** dashboard culminated and refined our work, providing a real and tangible piece of software that could be developed further for clinical usage. Finally, I have written a strong majority of the final **report**, setting high production and visual standards throughout and constructing a coherent narrative, well supplemented by custom figures I designed.

Project Strengths and Weaknesses

The main strength of our project was the widespread group innovation, with the creation of an idea-sharing culture and lots of brainstorming sessions. The regular inclusion of group check-ins where cross-pollination of ideas could take place was a major factor in the success of this project. There was also clarity in our high level of motivation paired with time-bound goal setting, allowing for effective and quick work progress. However, a weakness was our occasional struggle with task prioritisation with large work efforts occasionally not being best utilised. In future, rigorous planning before any dedicated work session would ensure better time usage.

Relation to Course Material

The course material, from my point of view, had real overhead relevance but somewhat lacked practical links to the final project. With the diverse set of dataset choices and limited lab sessions, I can however really appreciate how tailoring practical sessions to easily cater to all final projects would be impossible. I found the most crucial learning throughout the taught material was the overarching Data Science pipeline. Alongside this, specific material including Quasi Identifiers, Subtleties of Missing Data, Parametric Thinking, Custom Metrics and Virtuous Cycles was most specifically applicable in the end.

Achievements and Lessons Learned

Through this project, I have developed better leadership skills and practised good communication. I have been given a supported environment to trial real agile practices and learn key lessons in how to best delegate work and plan effectively. I believe we as a team have developed an exceptional system that has the potential to be developed into a real-world clinical tool, creatively spinning an ordinarily purely mathematical problem.