Group 7: Predicting Sepsis Risk during In-patient admissions

Ashwani Kumar Sharma,[1] Tu Kha Huynh,[1] Amy Wing Tung Hung,[1]
Alexander Rhys Turner,[1] Nyamtsetseg Nergui,[1] and Abdul Momen Usmani[1]

[1]*Department of Engineering and Mathematical Sciences,*
University of Western Australia, Perth, Western Australia, Australia

(Dated: July 14, 2025)

I. INTRODUCTION

Sepsis remains a formidable challenge in modern healthcare, notorious for its vague early symptoms, rapid progression, and high mortality rates. The primary objective of this project, initiated in collaboration with Royal Perth Hospital (RPH) and the Health in a Virtual Environment (HIVE), is to develop a robust machine learning (ML) model to aid in the early detection and management of sepsis. Given sepsis' ambiguous clinical presentation, current reliance on manual diagnostics and clinical judgment has led to inconsistent risk assessments, potential misdiagnoses and critically, delays in lifesaving interventions.

The synthesised RPH and MIMIC-III data used in this project ranges from structured to unstructured, is high-dimensional, evolves over irregular intervals, and captures diverse patient trajectories and medical pathways. Specifically, it contains charted pathology events such as demographics, vital signs, laboratory tests, and vital status; initial admission and overall diagnosis in the form of ICD-9 diagnostic codes; administered drugs; and written clinician evaluations covering the patient admission length of stay (see section A for an entity-relationship-attribute diagram). Given the escalated risk of inpatient fatality for each delayed hour of treatment [12], we consider only patient pathology and admission data available at early stages of admission: demographic information; comorbidities and risk factors; and test results available from patient lab results. With consideration for the paramount importance of interpretability and transparency in healthcare predictions, and the accuracy of model predictions, our investigation is anchored around the efficacy of three distinct algorithmic methodologies: traditional ML techniques, Long Short-Term Memory networks (LSTMs), and survival analysis strategies.

Our approach has several key benefits: optimising hospital resource allocation by automating initial data interpretation stages, improving care quality through standardised sepsis assessment criteria, and potentially reducing healthcare costs via early, accurate diagnoses. Furthermore, the model paves the way for advanced developments, including real-time risk assessment and the discovery of novel sepsis biomarkers or patterns through employed learning algorithms. This project, thus, not only stands to bolster healthcare efficiency and patient outcomes at RPH but also contribute significantly to the global scientific discourse on sepsis understanding and management.

II. METHODS

A. Data preparation

TABLE I: Summary of characteristics of MIMIC-III dataset

	Sepsis	Non-Sepsis
Number of admission (n)	5,406	53,570
Age (years, median)		
<18	0	0
18+	68	65
Female (n, %)	2,403 (44.5%)	23,623 (44.1%)
Mortality Rate (n, %)	3,292 (60.9%)	19,294 (36.0%)
Admission length of stay (days, median)	10.07	6.23
Time for first blood test since admission (hours, mean)	0.84	2.41

All patients with sepsis related ICD-9 codes in the dataset were identified and enrolled in this study and summary statistics were calculated (Table I). Considering our clients' emphasis on adult populations and the inherent biological differences that distinguish them from younger patients, we excluded all patients under the age of 18 from our study [22]. Patients with unreasonable age (above 100) were replaced with the median age. To handle unique admission identifiers (HADM_ID) missing in lab-events data, we joined the admissions data with lab-event tables by subject ID, and checked whether recorded chart time was within the admission time and the discharge time for each patient. Admissions with no recorded lab-events, or admissions with missing HADM_ID values were excluded. 50,766 unique admissions met the criteria for this study (see 1).

Considering the variations in ward admission times and test results across admissions, a standard reference time was defined in order to structure the data suitably for modelling approaches. This reference time was determined by the earlier of: time of admission; time the first chart result was documented. If a sepsis patient is not treated within six hours of hospital admit time, their mortality rate increases by 9% [16]. As such, the scope of our project prioritises early hours of patient data post reference time (up to 8).

Given the class imbalance present in our data (10.9% sepsis patients), we utilised stratified sampling for training data preparation for all modelling approaches to help address bias and ensure model generalisability to unseen data. Training and test sets were generated with proportions of 80% and 20% respectively. Given the differing units of measurements, we standardise the input data.

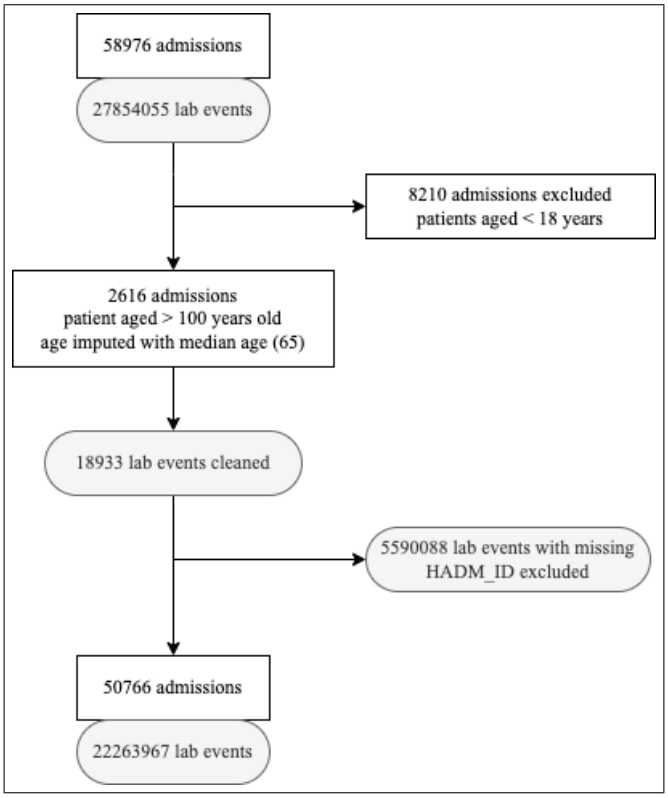


FIG. 1: Exclusion Criteria

B. Feature Selection

Selecting meaningful features for healthcare prediction requires establishing causal inference with the disease [18]. Given the ambiguity of early sepsis biomarkers, the feature selection process was crucial, and performed in iterative stages:

Literature Review - Initial analysis and extraction of features (lab-test measurements, comorbidites and high-risk groups) according to current, reputable research [6], [3].

Theoretical Plausibility - We rely on the expertise of clinical collaborators to confirm relevant features. Furthermore, considering temporal data leakage bias, we avoid including microbiology results or prescribed medications, both of which are typically ordered after sepsis onset.

High-risk Groups - Through statistical and visual analyses of demographic information, we identified that the risk of sepsis increases with age 2. Additionally, studies in the medical field have revealed that females have a higher susceptibility to sepsis [4]. Consequently, we included age and gender to ensure that these high-risk groups were adequately represented.

Lab Test Derangements - Using Neo4j, the abnormal flags of lab-test results were counted for all admissions, as well as specifically for sepsis admissions. Based on popularity index, lab-events showing derangement's in sepsis cases that were recorded within 10 hours after admission were extracted. Furthermore, the fluid type of these items for sepsis patients were confirmed to be predominantly from blood tests (96%), aligning with studies of hematologic changes in sepsis [10].

Availability of Clinical Measurement - The early data collection methods of RPH were assessed through an analysis of the temporal dynamics of lab results. The results indicate that blood fluid types are the most frequently taken (3) and earliest available measurements available for all patients (4).

Feature Engineering - The data supplied by RPH utilises ICD-9 codes for clinical diagnosis, a system now deemed outdated in light of the transition to ICD-10 in 2015 [2] and the introduction of the more refined Sepsis-3 definition in 2016 [20]. Critically, both rely on the use of the Sequential Organ Failure Assessment (SOFA) to quantitatively describe the development of organ dysfunction [23]. To avoid historical bias, we engineered the SOFA score as a feature.

Identifying Comorbidities - Comorbidities may predispose individuals to an increased risk of developing sepsis. Centrality metrics in Neo4j were used to identify comorbidities most commonly associated with sepsis. Abnormal biomarkers for these comorbidities were extracted.

Establishing Causality - Causality implies that one condition directly influences the occurrence of the other, not just that they happen to occur together. To establish causality with sepsis, comorbid conditions and their associated lab-test results (with abnormal count greater than 1000) were queried against the RPH local language model, fine-tuned to medical data [15] (see Table II).

TABLE II: Summary of most frequent comorbidity and important features

Variable	Sepsis	Non-Sepsis
Frequent Comorbidity (n, %)		
Septicemia NOS	3,102 (59.9%)	604 (1.3%)
Septic shock	2,442 (47.2%)	143 (0.3%)
Acute kidney failure NOS	2,106 (40.7%)	7,006 (15.4%)
Urin tract infection NOS	1,329 (25.7%)	5,218 (11.5%)
Pneumonia, organism NOS	980 (18.9%)	3,823 (8.4%)
Important Features (n, %)		
Hemoglobin	5,114 (98.8%)	42,535 (93.3%)
Red Blood Cells	5,109 (98.7%)	43,107 (94.6%)
Glucose	5,108 (98.7%)	43,091 (94.5%)
Hematocrit	5,094 (98.4%)	42,724 (93.7%)
Calcium, Total	4,868 (94.1%)	30,868 (67.7%)

Admissions with an age less than 18 are excluded from the calculation.

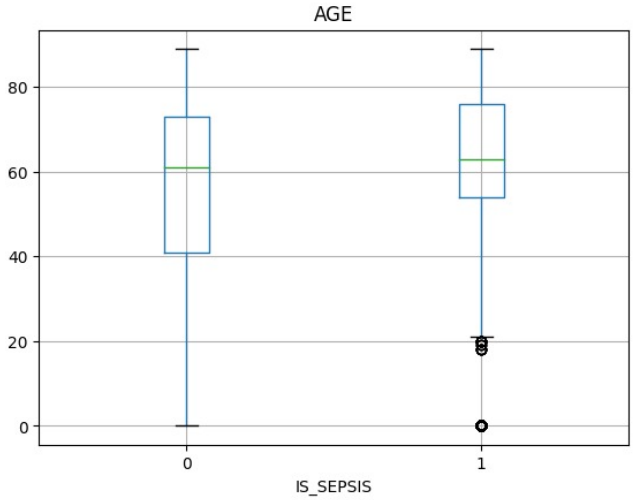


FIG. 2: Distribution of Age

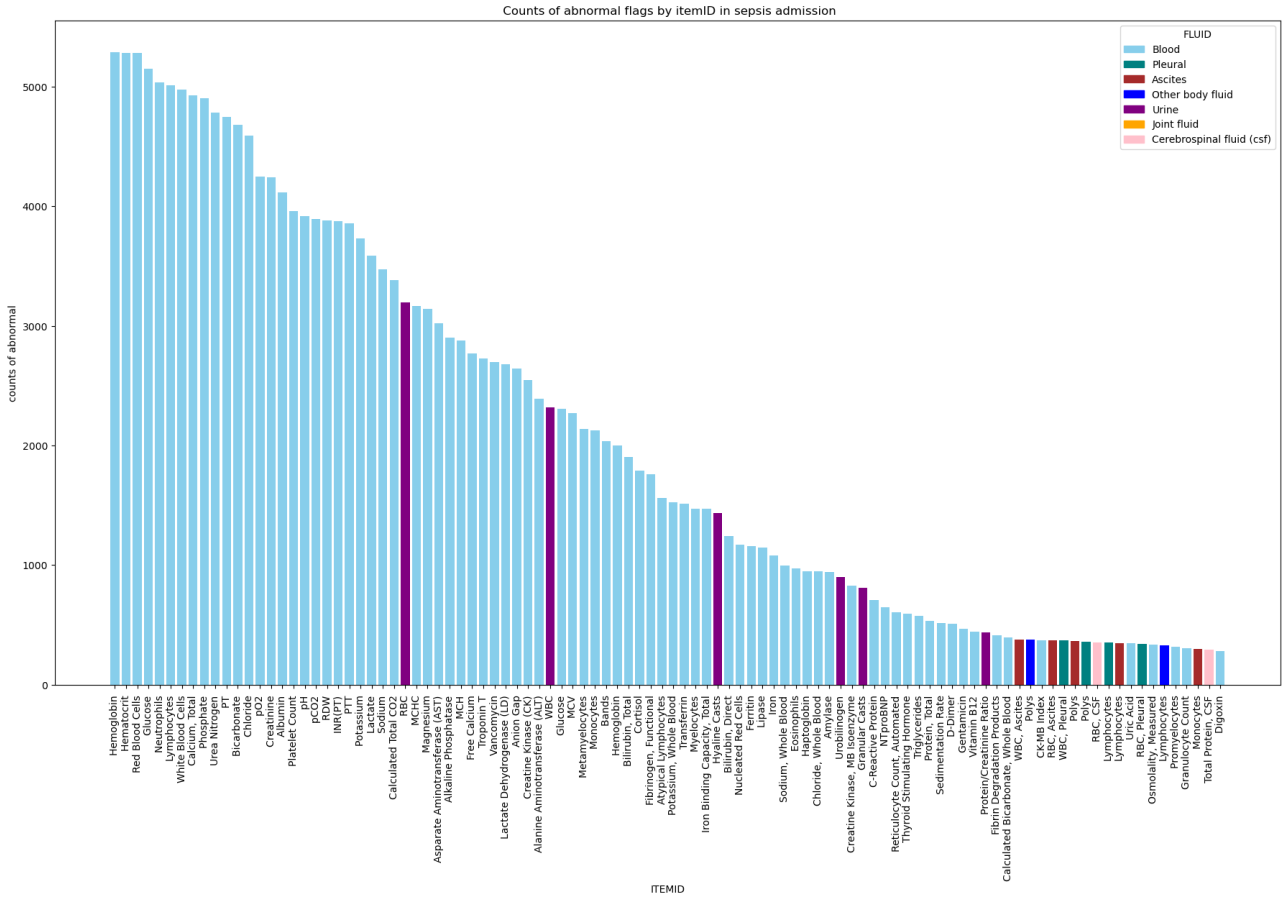


FIG. 3: Frequency of lab test by fluid type (considering entire admission time)

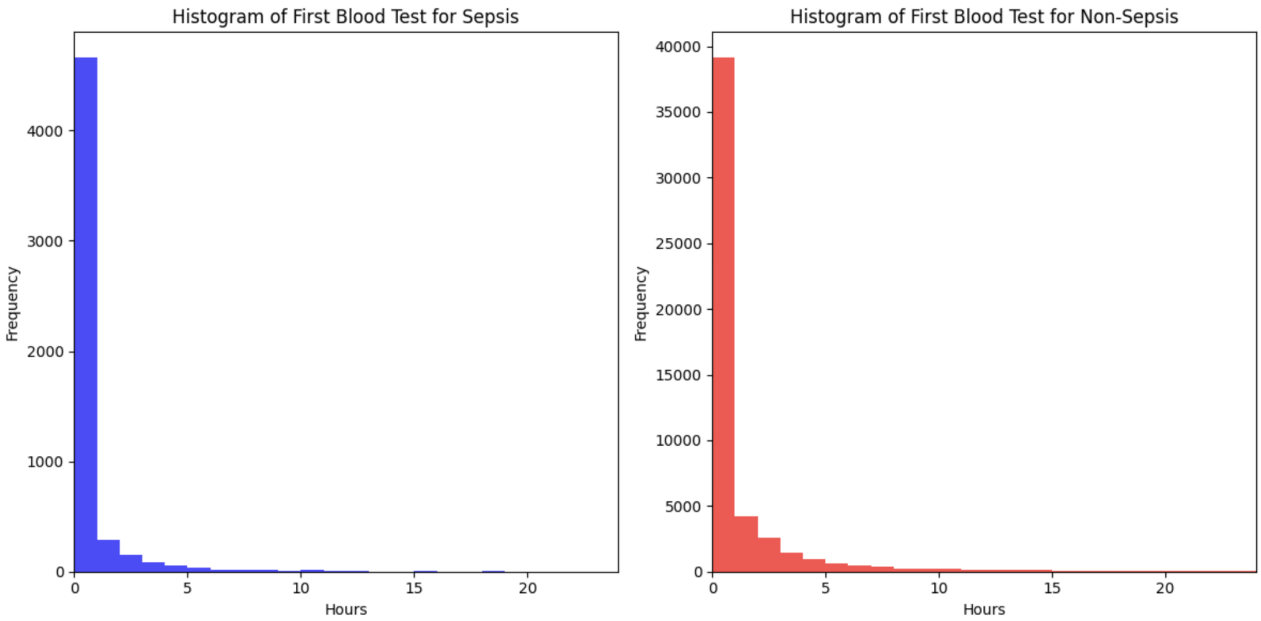


FIG. 4: Blood test results are available in the early admission hours

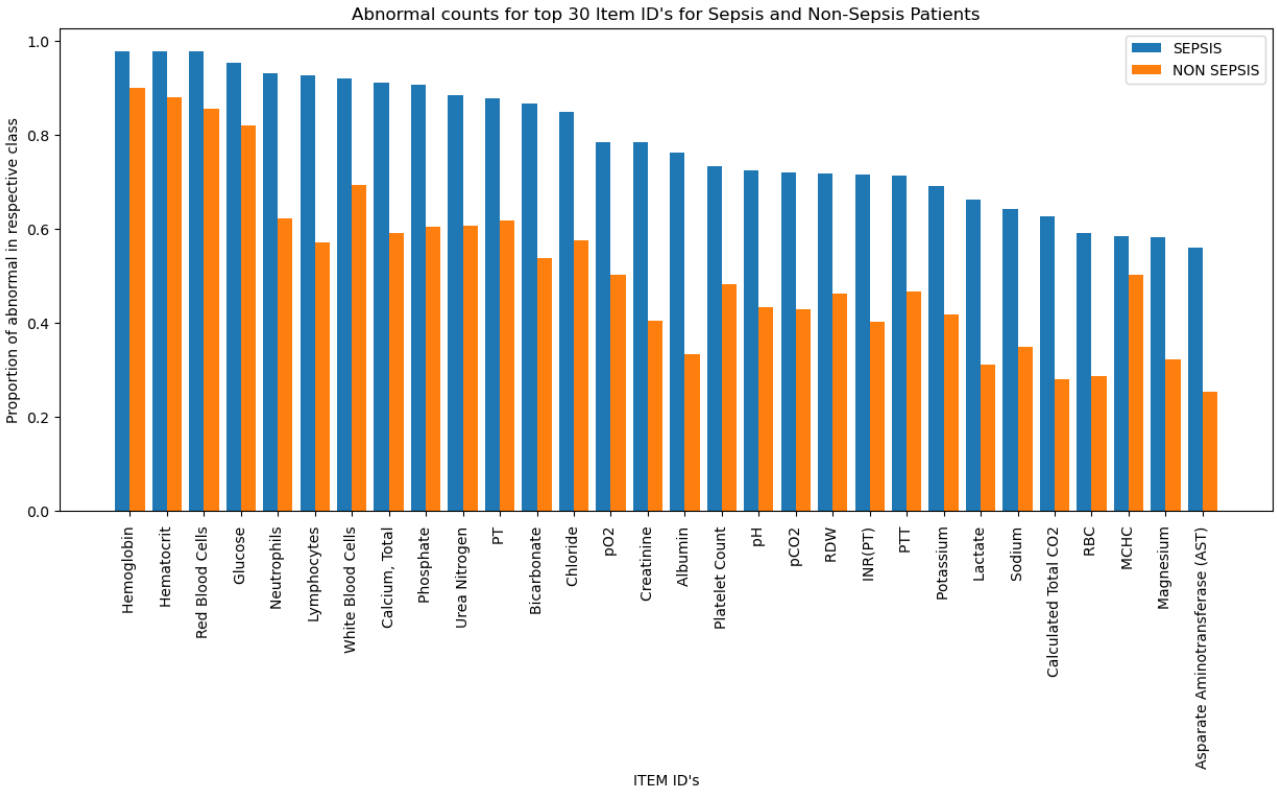


FIG. 5: Proportion of abnormal lab tests by sepsis label (considering features identified as important by Neo4j)

C. Machine Learning

1. Traditional ML

The selection of candidate algorithms were dependent on their strengths in the context of healthcare, notably the ability to: handle variability; reduce overfitting; handle high dimensionality data; be interpretable by design. After careful review of prior sepsis research using early-stage admission data, logistic regression [8] [24], random forest [21], and gradient boosting classifier [26] algorithms were selected.

Forward filling was employed for data extraction at various time windows. In cases where test results were absent during the data extraction phase, a designated masking value was used. With the limitations of the static model structure, we identify an appropriate time window to extract patient lab events data by analysing the time at which the first test results were available for each patient (6, 7).

The primary objective of traditional ML models in this project is to predict sepsis as early as feasible using the fewest features whilst maintaining a high performance level. To determine the optimal combination of features and timing, we trained models across multiple configurations: three algorithms, five time windows (t0 to t4), and four different feature set sizes (top 10 to top 40 features). In total, 60 distinct models were evaluated. For each of these time windows, the percentage of missing data was assessed and evaluated against model performance.

Owing to the imbalanced class proportions, we adjust weights for model training to prevent bias to majority class. Appropriate hyperparameters were selected for each model to reduce overfitting (III, IV, V).

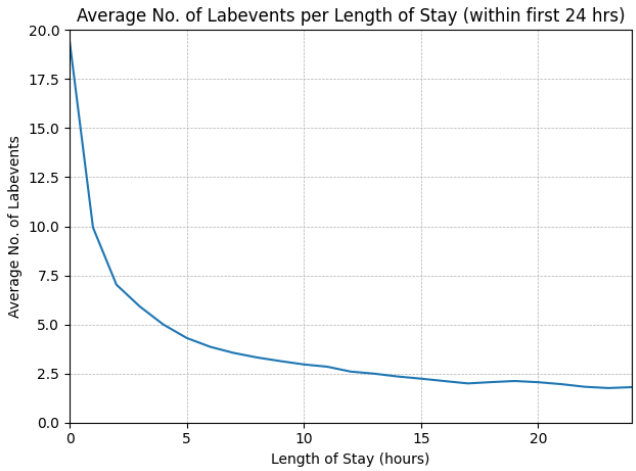


FIG. 6: Average number of lab events per length of stay for all admissions

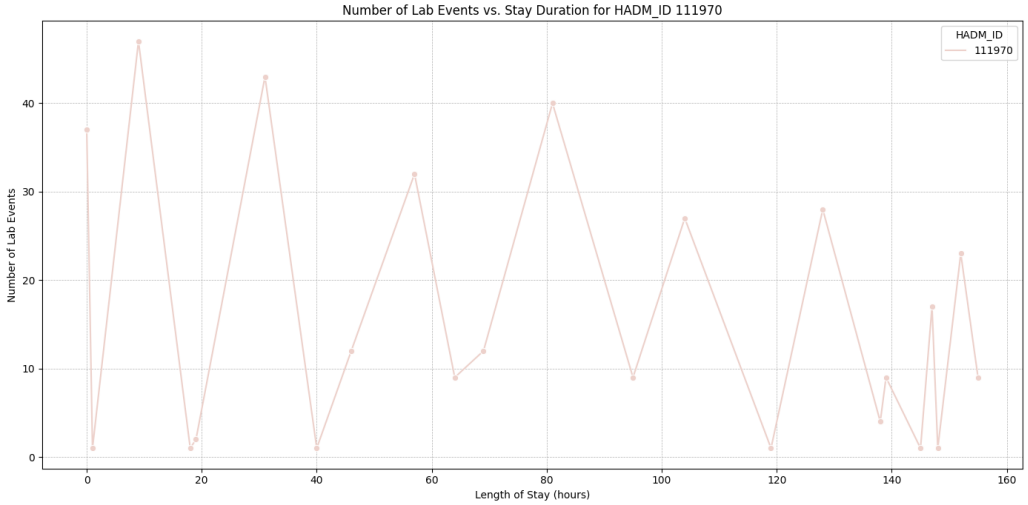


FIG. 7: Number of lab events per length of stay for a single admission

TABLE III: LR Model parameters

Hyperparameter	Value
Maximum iteration	10000000000
C	0.1
Penalty	None

TABLE IV: RF Model parameters

Hyperparameter	Value
Max_depth	7
Min_samples_leaf	25
Min_samples_split	250

TABLE V: GB Model parameters

Hyperparameter	Value
Number of Estimators	100
Learning Rate	0.1
Max_depth	3

2. Deep Learning

The LTSM networks handle sequences of varying lengths by maintaining a memory state and using gates to capture information from earlier time steps and carry it forward through the sequence, making them effective in sepsis prediction [13]. With the healthcare data, where patient records have varying admission durations and time intervals between observations, LSTMs capture long-range dependencies and patterns in sequences of patient data without requiring padding or truncation. Furthermore, LSTMs solve the vanishing gradient problem by controlling information flow through time and allowing the network to learn long-range dependencies without gradients disappearing. The use of additive interactions within the memory cells helps ensure that gradients can flow more freely, making it possible for LSTMs to capture and learn from important patterns in long sequences of data. In sepsis prediction, where early signs might be spread across multiple time steps, mitigating the vanishing gradient problem is critical for effective learning and accurate predictions.[19]

The input data - The input time series data was collected at time from 0 to 8 hours post reference time and is organised chronologically. Missing values were imputed using mean aggregation. Since the input data has irregular intervals, it is handled using Ragged tensors, allowing flexibility in predicting the onset of sepsis

at various time points. Ragged tensors employ nested lists to represent admissions, each with a shape of <Timestep, fixed number of features>. While the number of features remains fixed for each timestep, timesteps can vary between admissions.

The model - The LSTM model consists of two LSTM layers with 64, and 64 units, followed by one dense layers with 2 units. The Rectified Linear Unit (ReLU) and sigmoid activation functions are used in the LSTM and dense layers allowing the model to learn complex patterns and relationships in data and to ensures that the sum of the output probabilities is equal to 1. A dropout rate of 0.25 is added between each layer to prevent overfitting. The model output is an array of 2 proportions for predicted target non-sepsis and sepsis.

Hyperparameter tuning is performed on learning rate and model structure. The exponential decay learning rate scheduler is used to decrease the learning rate after each epoch in order to avoid overshooting the global minimum of the loss function and aid with model convergence. The model parameters are displayed in Table VI.

TABLE VI: LSTM Model parameters

Neural Networks Parameter	Values
Activation of the dense layer	sigmoid
Optimizer	ADAM
Loss function	Binary cross entropy
Learning rate	Exponential decay learning rate with initial learning rate at 0.001
Early stopping	Validation accuracy
Number epoch	50

3. Survival Analysis

Survival analysis models were considered based on their ability to: effectively handle censored data; incorporate time-varying covariates; and provide an interpretable and continuous risk assessment over time, supporting timely and proactive clinical decisions [7]. If a subject has not died (experienced the event) by the end of his/her follow-up, we say that his/her survival time is censored. Notably, the Targeted Real-time Early Warning System (TREWS)–the most promising, successfully deployed alert system for sepsis–utilises a regularised Cox proportional hazards model [11]. In this study, we consider several survival modelling methods: Cox Proportional Hazards (coxph), Cox Proportional Hazards with Elastic Net regularisation (coxnet), Random Survival Forest (RSF), and Gradient Boosted (GB) Survival.

In our study, we assume the principle of independence, meaning the survival times of individual participants do not influence one another. Additionally, we operate under the assumption of non-informative censoring. The subjects who are censored have the same underlying risk of reaching the study endpoint as those who experienced the event.

Patients who had sepsis but did not die were excluded. Admission length of stay was calculated as the difference between discharge time and admit time. For the survival models, status (indicating whether the actual survival time was observed or it was censored) and length of stay are structured in an array. This corresponds to the time of death (if Status == True), or the last time that person was seen (if Status == False).

All candidate models were fit using input data extracted at time windows t0, t2, t4 and t8. A Cox's proportional hazard's model was fit on the different time data. For the coxnet model, elastic penalty is set to 0.9. The best hyperparameters were identified for the RSF model (VII) and the GB survival model (VIII) using GridSearchCV.

TABLE VII: RSF Model parameters

Hyperparameter	Value
Minimum Sample Split	5
Minimum Samples Leaf	5
Number of Jobs	-1

TABLE VIII: GB Survival Model parameters

Hyperparameter	Value
Number of Estimators	100
Learning Rate	1
Max_depth	5

D. Performance Measures

1. Traditional Machine Learning and Deep Learning

To assess the effectiveness of each of these modelling techniques, appropriate evaluation metrics (balanced accuracy, precision, recall, F1-score, and AUROC) were used as performance metrics. Stratified k-fold cross-

validation was employed to ensure that each fold is a good representative of the overall dataset, providing a more reliable estimate of its generalisability.

2. Survival Analysis

The performance of survival models is evaluated on basis of concordance index (C-index) and IBS score. C-index provides a measurement for how well a model can rank individuals by their survival times. IBS score provides a calibration measurement; an overall calculation of the model performance at all available times. Predicted survival probability of individual patients over time are plotted, which support patient prioritisation in the clinical environment.

III. RESULTS

A. Traditional Machine Learning Results

Initial evaluations based on balanced accuracy indicated that models trained using the top 40 features at the t3 time window perform best. To validate these findings, we conducted further tests using 5-fold cross-validation in two specific scenarios: comparing performance across different time windows while keeping the number of features constant at 40; and evaluating how performance changes with varying feature set sizes, specifically focusing on the t4 time window.

From the 5-fold cross-validation results, we observed the performance gains plateaued beyond the t2 time window and after considering the top 20 features. Notably, the combination of the t2 time window and top 20 features was identified as the most streamlined and effective for predicting sepsis. This conclusion was considered to be more trustworthy compared to the initial assessments as, in 5-fold cross-validation, the dataset is partitioned into five distinct subsets. For each iteration, one subset is employed as the validation set while the remaining are used for training. After cycling through all subsets, the performance metrics are averaged, providing a more comprehensive and consistent measure of the model's performance. This averaged outcome, derived from multiple iterations, ensures a reduction in potential biases or anomalies associated with any particular data split, making it inherently more robust and reliable than single-pass evaluations.

Models trained in t2 with 20 features were then evaluated with more important performance matrices as shown in TABLE IX, FIGURE 8, and FIGURE 9. Generally, we observe Random Forest model performed the best among the three traditional ML algorithms with the highest balanced accuracy (0.747), recall (0.758), F1 score (0.411) and satisfactory AUROC (0.828).

TABLE IX: Traditional ML models Performance (20 features and t2)

	Random Forest	Logistic Regression	Gradient Boosting
Balanced Accuracy	0.747	0.720	0.586
Precision	0.282	0.258	0.603
Recall	0.758	0.724	0.189
F1 Score	0.411	0.381	0.288
AUROC	0.828	0.776	0.838

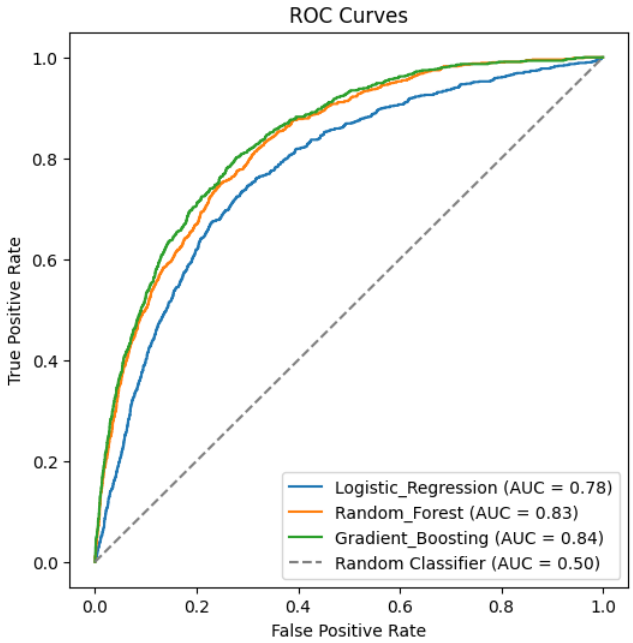
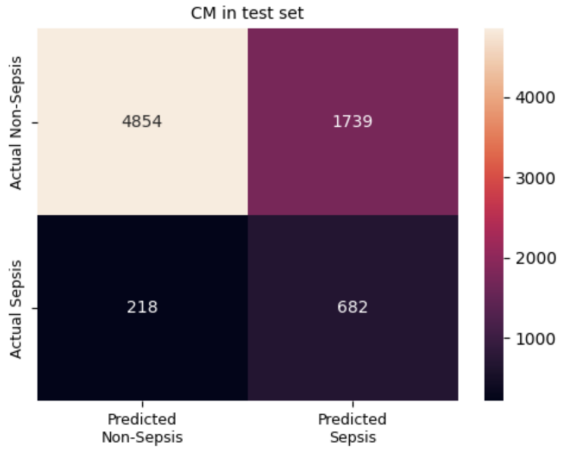
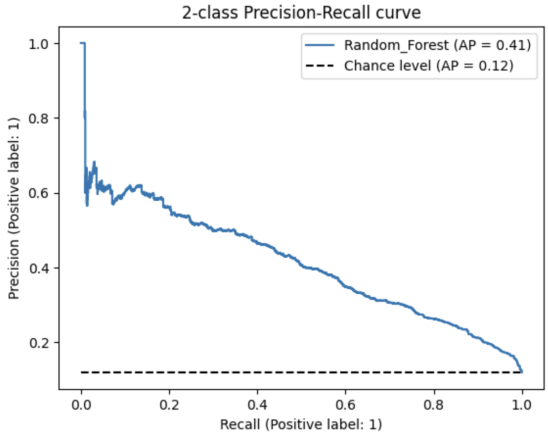


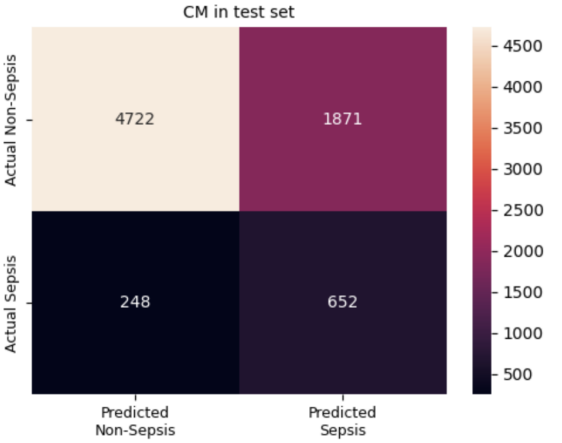
FIG. 8: Traditional ML models at t2 and top20 features



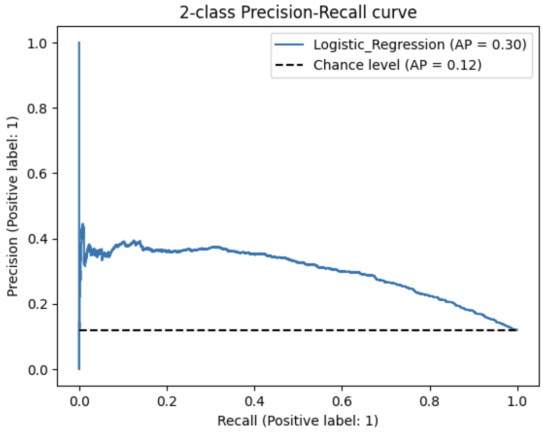
(a) Confusion Matrix (Random Forest)



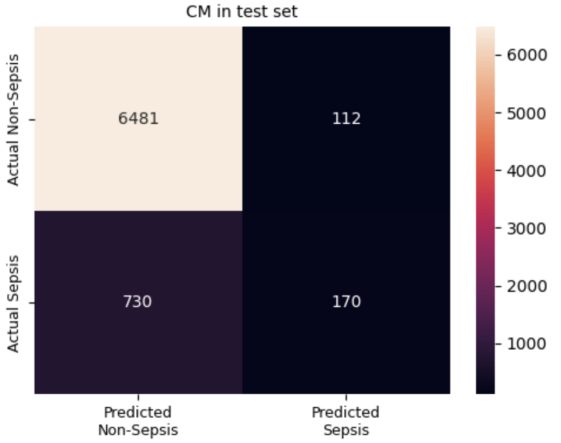
(b) Precision Recall Curve (Random Forest)



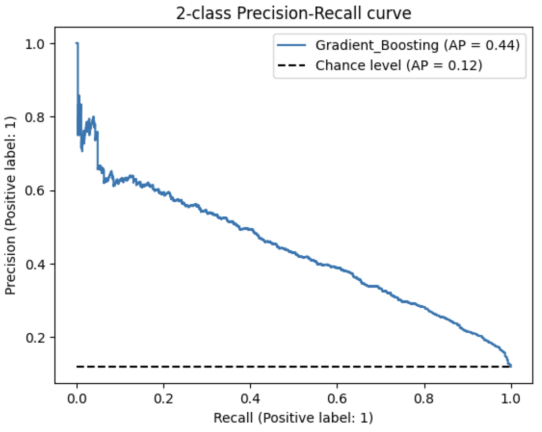
(c) Confusion Matrix (Logistic Regression)



(d) Precision Recall Curve (Logistic Regression)



(e) Confusion Matrix (Gradient Boosting)



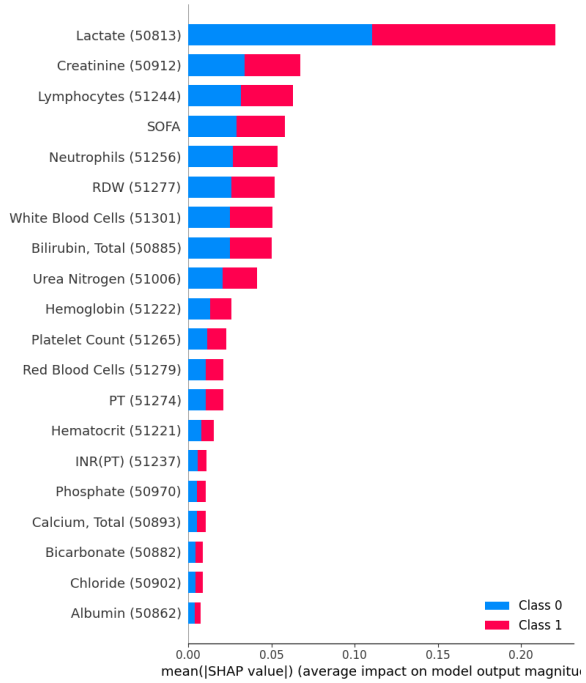
(f) Precision Recall Curve (Gradient Boosting)

FIG. 9: Plots for Traditional ML model performance

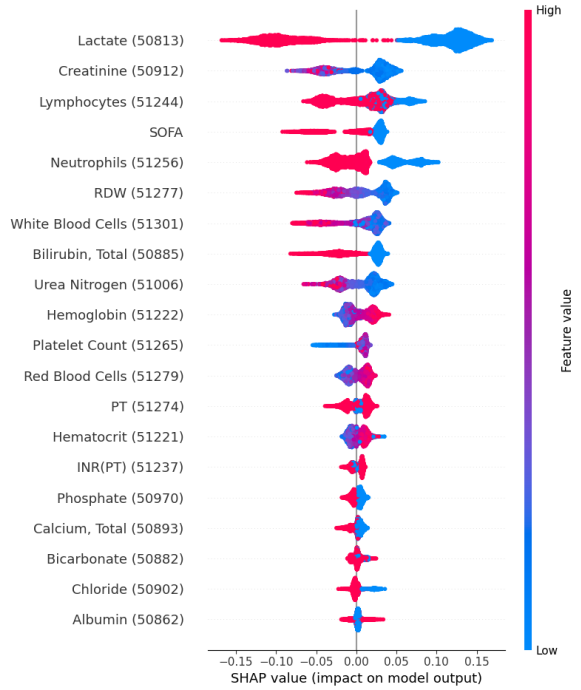
Feature importance was also investigated among traditional ML models using SHAP. From TABLE X, we observe Lactate, Neutophils, Lymphocytes, White Blood Cells, and Total Bilirubin were consistently ranked as the top 10 important features in all three traditional ML models. Given Random Forest is the best performing traditional ML model, we included its SHAP results (both global and individual feature importance) in this report for better illustration in FIGURE 10 and FIGURE 11.

TABLE X: Top 10 Global Features in Traditional ML models by SHAP

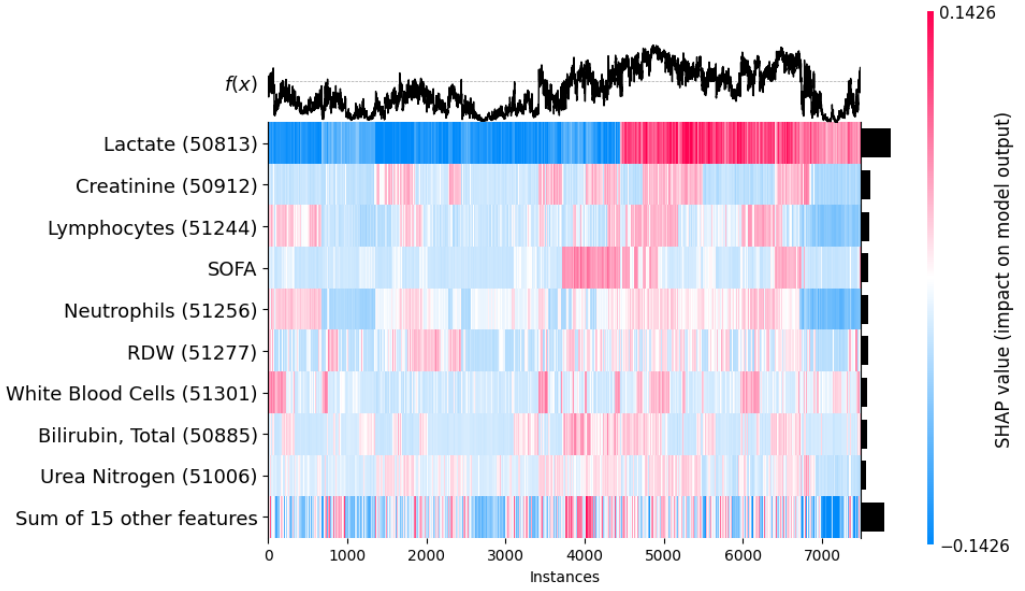
Rank	Random Forest	Logistic Regression	Gradient Boosting
1	**Lactate**	**Neutophils**	**Lactate**
2	Creatinine	**Lymphocytes**	**Neutophils**
3	**Lymphocytes**	RDW	**White Blood Cells**
4	SOFA	Red Blood Cells	**Lymphocytes**
5	**Neutophils**	**Lactate**	RDW
6	RDW	INR(PT)	Creatinine
7	**White Blood Cells**	PT	**Bilirubin, Total**
8	**Bilirubin, Total**	**Bilirubin, Total**	SOFA
9	Urea Nitrogen	**White Blood Cells**	Hemoglobin
10	Hemoglobin	Bicarbonate	PT



(a) Bar chart for Top ranked features

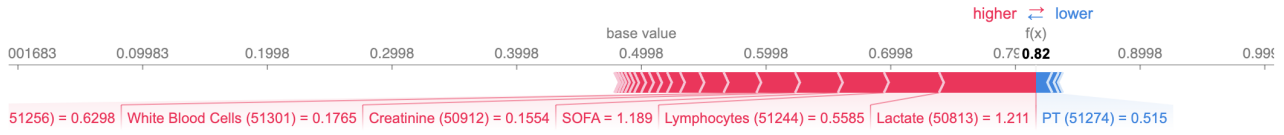


(b) Summary plot for Top ranked features

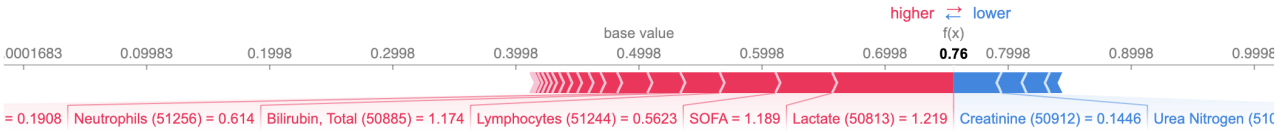


(c) Heatmap for Top ranked features

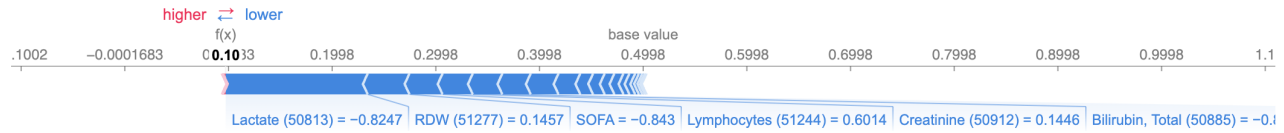
FIG. 10: Global Feature Importance in SHAP (Random Forest)



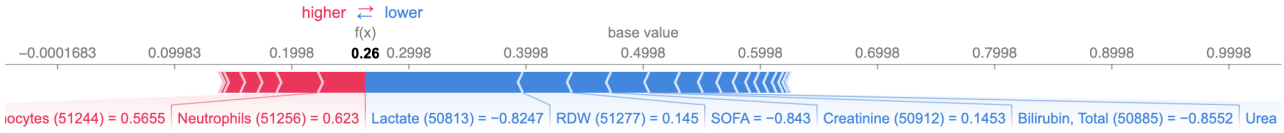
(a) Sepsis Admission Example 1



(b) Sepsis Admission Example 2



(c) Non Sepsis Admission Example 1



(d) Non Sepsis Admission Example 2

FIG. 11: Individual Feature Importance in SHAP (Random Forest)

B. Deep Learning

In the MIMIC-III dataset, our primary focus is on laboratory event results presented in a time series format. To utilise the advantages of LSTM networks for time series event data, we need to establish the time step, measured in hours since admission, at which a patient exhibits all abnormal signs in laboratory results but has not yet received medical treatment. Specifically, we are examining the initial 8 hours following admission to determine the most effective time step for predicting sepsis. This approach is aimed at training the model using event results before medical treatment is administered. In terms of features, top 20 features confirmed through both traditional ML models and feature selection step are used. The ROC curves and performance of the model at each timestep are presented in FIGURE 12 and TABLE XI

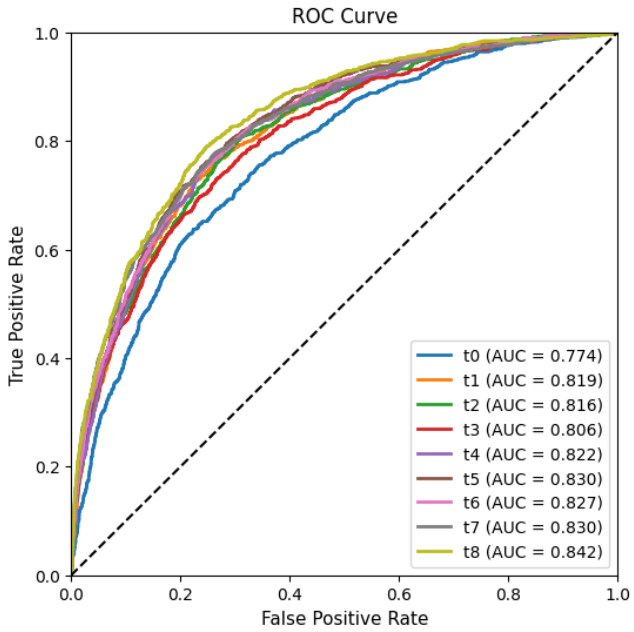
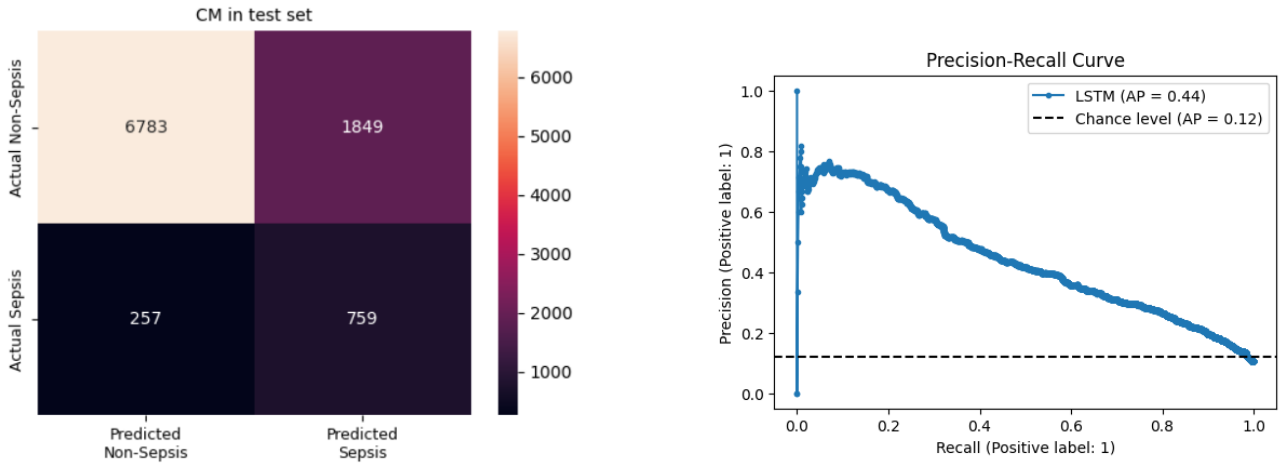


FIG. 12: Deep Learning models at t from 0 to 8 and top 20 features

TABLE XI: LSTM Models Performance on Test Set (20 features and t0 to 8)

	t0	t1	t2	t3	t4	t5	t6	t7	t8
Balanced Accuracy	0.703	0.743	0.738	0.726	0.743	0.754	0.753	0.752	0.766
Precision	0.248	0.295	0.279	0.272	0.3	0.281	0.287	0.289	0.291
Recall	0.662	0.698	0.701	0.673	0.676	0.732	0.718	0.711	0.747
F1 Score	0.361	0.414	0.399	0.387	0.416	0.406	0.41	0.411	0.419
AUCROC	0.774	0.819	0.816	0.806	0.822	0.83	0.827	0.83	0.842

From these results, the model trained on t8 data performs best with an AUC of 84.2% and a balanced accuracy of 76.6%. Given the flexibility of the input sequence length, this model is applicable for predicting observations at any timestep, not limited to t8. Thus, it can be used for early sepsis prediction at any time (t). For this model, sepsis admissions are correctly predicted in nearly 75% of cases, with an Average Precision of 44% (see FIGURE 13).



(a) LSTM Confusion Matrix

(b) LSTM Precision Recall Curve

FIG. 13: Plots for LSTM model performance

In the LSTM model, we utilize TimeSHAP (Time Series Shapley Values) as a model-agnostic recurrent explainer to elucidate feature importance. We generate global explanations by applying TimeSHAP to all sepsis admissions, employing a tolerance of 1.25e-6 for the temporal coalition pruning algorithm. Top 5 most important features extracted from the LSTM model are Lymphocytes, Neutophils, RDW, Urea Nitrogen, and Platelet Count. The highlighted features in Table XII are those that hold a high rank in the traditional models.

TABLE XII: Top Global Features in LSTM model by TimeSHAP

Rank	Important Features
1	**Lymphocytes**
2	**Neutophils**
3	RDW
4	Urea Nitrogen
5	Platelet Count
6	**White Blood Cells**
7	Creatinine
8	**Bilirubin, Total**
9	SOFA

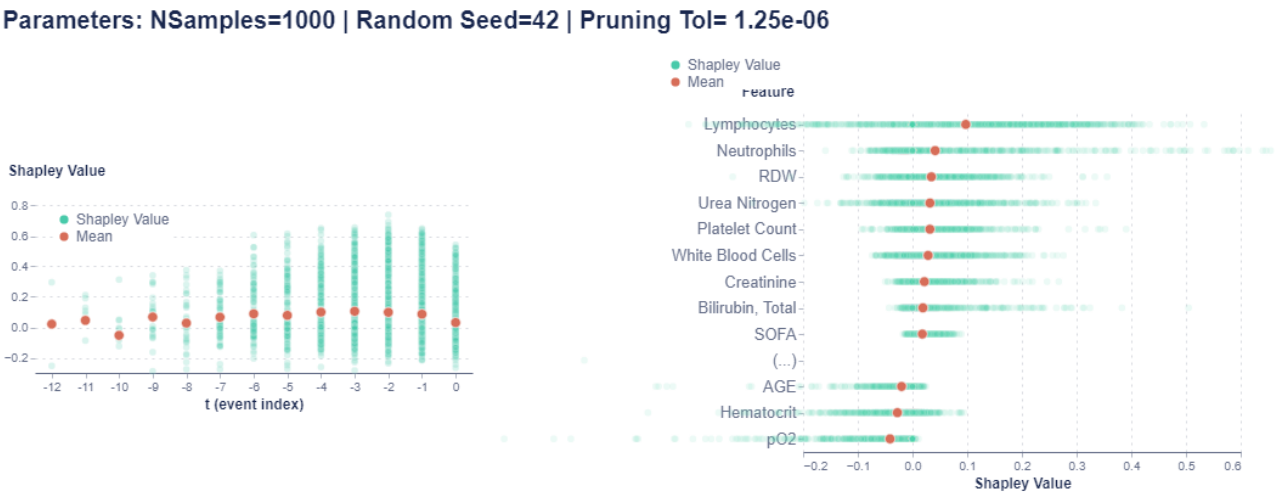


FIG. 14: Global Feature Imprtance in TimeSHAP

Four individual admissions are selected to observe local feature explanations, including two sepsis cases and two non-sepsis cases. Local feature explanations are computed using a tolerance of 0.025 and an nsample of 32,000. As demonstrated in FIGURE 15, Lymphocytes, Neutrophils, and Urea Nitrogen are among the top 5 features for both sepsis cases.

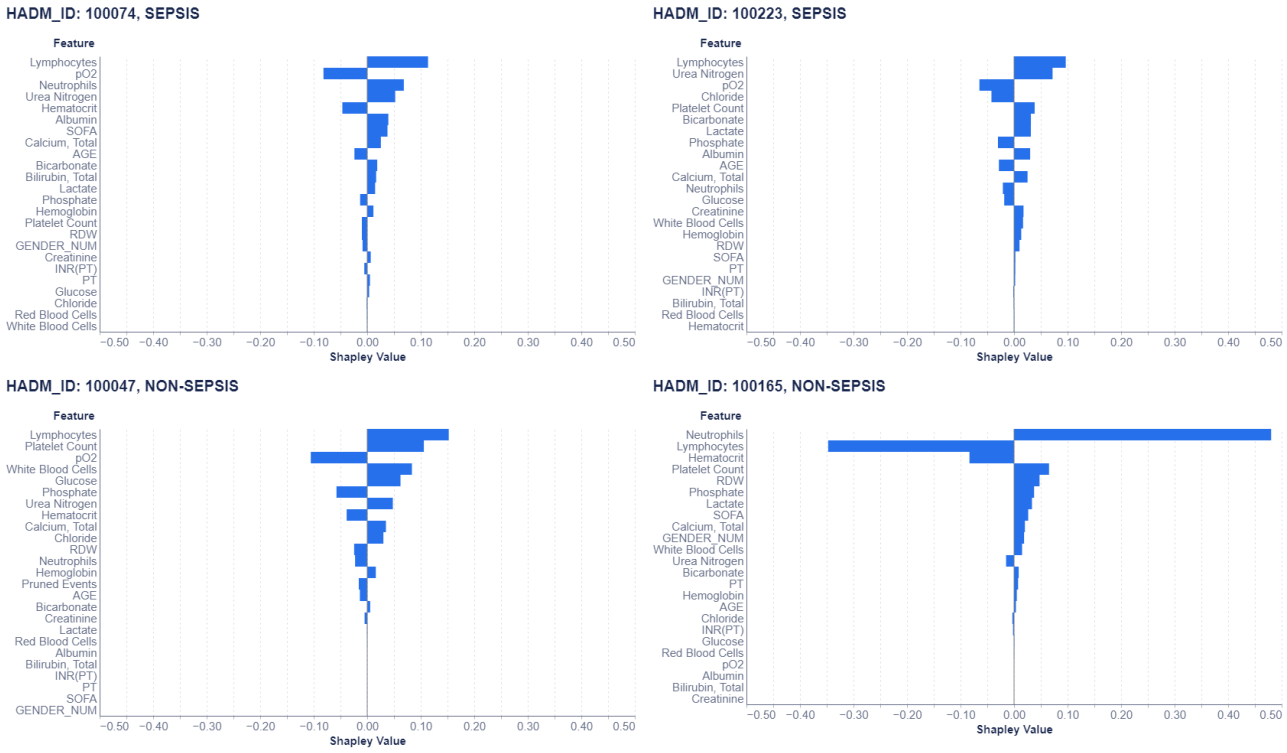


FIG. 15: Individual Feature Importance in TimeSHAP)

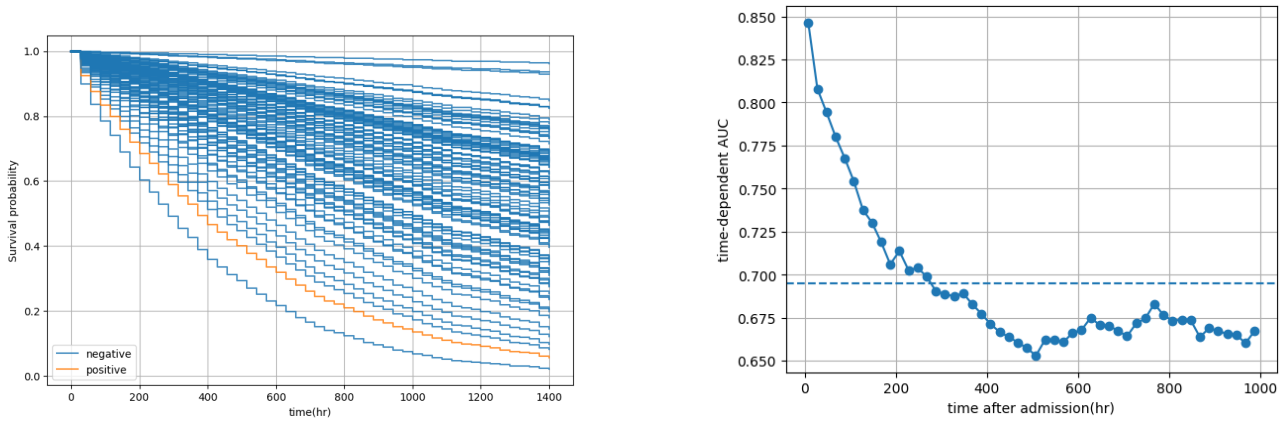
C. Survival Analysis

TABLE XIII: Survival model results

Model	t0		t2		t4		t8	
	C-index	IBS score	C-index	IBS score	C-index	IBS score	C-index	IBS score
coxph	0.679	0.120	0.715	0.116	0.728	0.110	0.742	0.107
coxnet	0.677	0.119	0.714	0.116	0.717	0.111	0.741	0.107
RSF	0.728	0.116	0.763	0.112	0.772	0.104	0.795	0.102
GBM	0.733	0.128	0.759	0.124	0.768	0.112	0.798	0.111

The primary object of survival analysis is to predict the survival probabilities of a patient over any time t after their admission. Based on the survival model results XIII, the RSF and GBM models generally outperform the coxph and coxnet models in terms of both predictive accuracy and calibration, especially at later time points. This might be due to their ability to capture complex, non-linear relationships and interactions between variables, which are common in clinical data. The performance of all models improves over time, with the most significant improvements typically seen between t0 and t2. This suggests that the additional data available at later time points significantly enhances the model's ability to predict sepsis. Despite the RSF model's slightly superior performance, the GBM model's higher C-index and competitive IBS score suggest it's also a strong candidate for sepsis prediction. Overall, we would go for RSF model as best model among all other survival models.

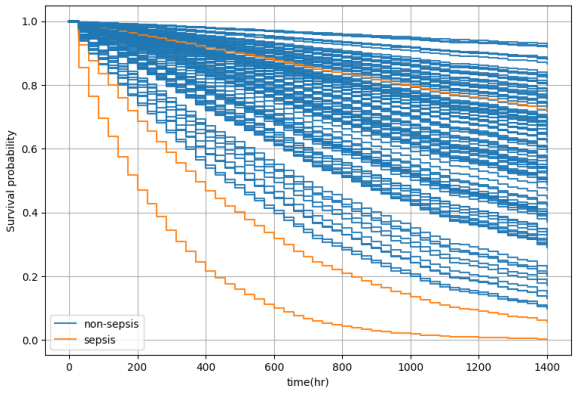
The survival probability curves provide insight into a patient's survival chances at time t after their admission. We can also have a look at the Survival probability curves over time, predicted by the respective best models and time-dependent AUC curves at t8 dataset. we can see that in RSF and GB Survival models, high risk patient has 40% reduction in their survival probability between 200 and 300 hours in comparison to Coxph and Coxnet moel where it gets around 300 hours. Also the time-dependent AUC curves of all models shows same kind of pattern, which is decreasing initially and then becomes gradual after 400 hours.However the RSF and GB AUC curves have higher mean AUC(around 0.74) in comaprison to COXph and Coxnet (around 0.68).



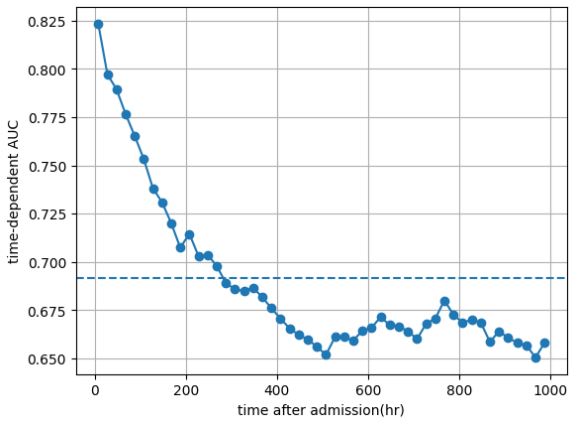
(a) Survival probability curves

(b) Time-dependent AUC

FIG. 16: Plots of coxph model

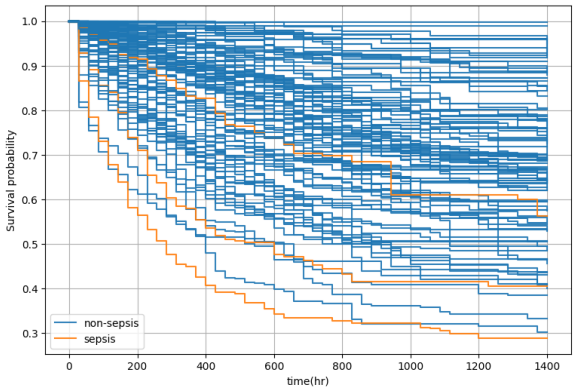


(a) Survival probability curves

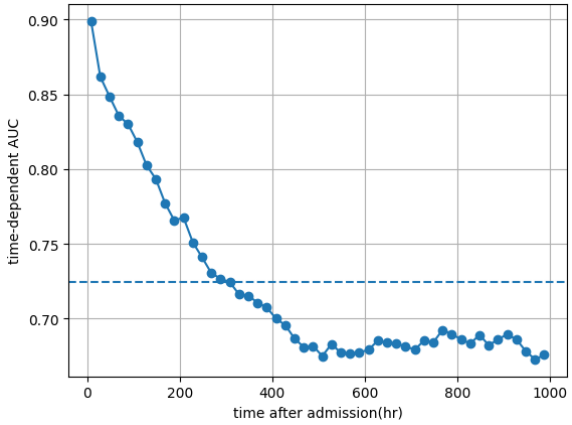


(b) Time-dependent AUC

FIG. 17: Plots of coxnet model

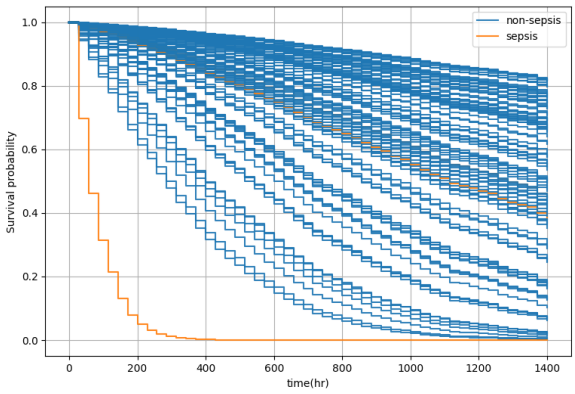


(a) Survival probability curves

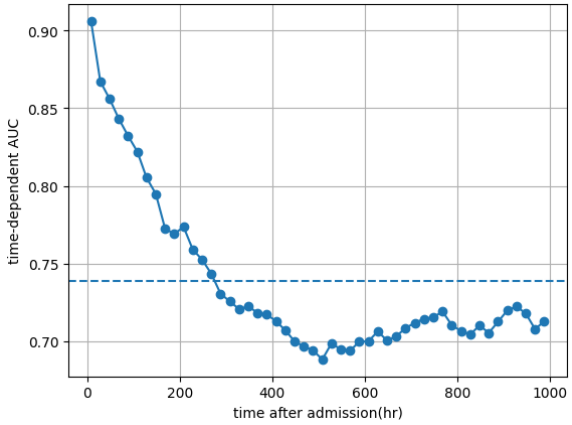


(b) Time-dependent AUC

FIG. 18: Plots for Random Survival forest model



(a) Survival probability curves



(b) Time-dependent AUC

FIG. 19: Plots for GB survival model

TABLE XIV: Top Features playing role in RSF model

Rank	Important Features
1	AGE
2	pH (50820)
3	Bands (51144)
4	PO2 (50821)
5	Lactate (508133)
6	Calculated Total CO2 (50804)
7	White Blood Cells (51301)
8	Neutrophils (51256)
9	pCO2 (50818)
10	Troponin T (51003)

Features importance for RSF model (best survival model choosen) also determined using permutation_importance. The top features which have major role in determining the survival probability and affecting the model's performance metrices(C-index and IBS score) are found to be AGE,PH,Bands,PO2,Lactate,Calculated Total CO2, WBC, Neutrophils, pCO2,Troponin T (shown in table XIV).

IV. DISCUSSION

In determining the optimal machine learning model for sepsis prediction across the above tested methods, our evaluation was based on two fundamental criteria: (i) predictive capabilities and (ii) model transparency and interpretability.

A. Predictive Capabilities

A comparison between Traditional ML models and the LSTM model is presented in TABLE XV and FIGURE 20. The LSTM model displays superior performance across most metrics, including Balanced Accuracy, F1 Score, and AUROC, compared to the traditional machine learning models. The LSTM model offers a significant advantage when dealing with time series data due to their ability to capture and model complex temporal dependencies of the patient trajectory. The Random Forest model, for instance, commendably holds its ground, it achieves competitive performance levels, especially notable given its reduced data input requirements (refer to metrics of RF at t2 compared to LSTM at t8). This highlights its efficiency and potential applicability in scenarios with constrained data availability.

The prevalence of missing values is intrinsic to working with early admission patient data, where clinical measurements are gathered at irregular intervals. Approaches used to handle missing time series information reflect a compromise between methodological rigor and data completeness. Focusing on preserving the integrity of the data, masking was used for traditional ML methods. This method doesn't mitigate the lost information, and as such, masking values may impact the predictive performance and interpretation of results. Conversely, the mean imputation utilised for training the LSTM model may introduce bias, making potentially erroneous assumptions of the data. Given that no model guarantees absolute accuracy in sepsis prediction, it gives rise to the significance of our subsequent consideration - model transparency and interpretability.

Considering the effects of alert fatigue on clinicians[25], the relatively low precision of our models is a concern; many false alarms may be made in sepsis identification. This being said, the potentially dangerous nature of false negatives, our high recall is preferable.

TABLE XV: Models Performance (Traditional ML vs Deep Learning)

	Random Forest	Logistic Regression	Gradient Boosting	LSTM
Balanced Accuracy	0.747	0.720	0.586	**0.766**
Precision	0.282	0.258	**0.603**	0.291
Recall	**0.758**	0.724	0.189	0747
F1 Score	0.411	0.381	0.288	**0.419**
AUROC	0.828	0.776	0.838	**0.842**

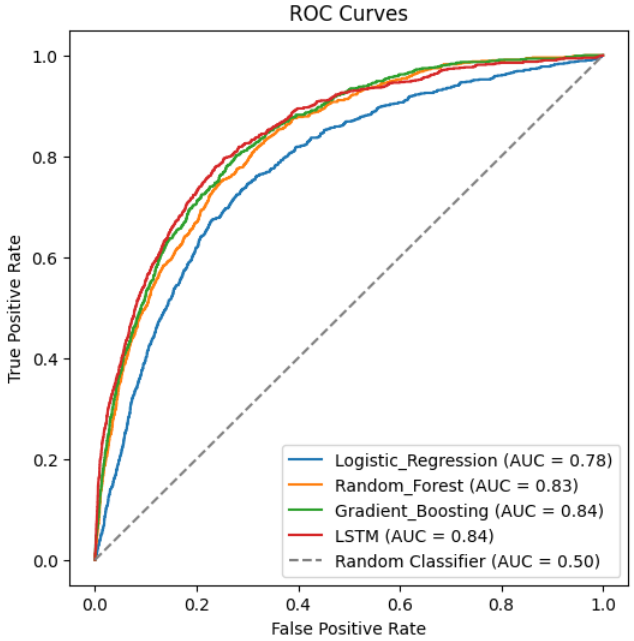


FIG. 20: ROC curves (Traditional ML vs Deep Learning)

B. Model transparency and interpretability

Given the high stakes of healthcare decisions, if a ML model gives a false prediction, understanding who or what is responsible is vital for all stakeholders involved in patient care [14]. We are attempting the mitigate the problem of clarity and establish trust in the model through the prioritisation of interpretable results, and through close collaboration with RPH domain experts.

Based on our results, the traditional ML models stand out in terms of interpretability. These models provide tangible insights into local and global feature importance, providing a clear understanding of the variables significantly influencing predictions. In this project, we have attempted to improve interpretability of machine learning predictions by applying SHAP [17] to quantify the contribution of each feature to model predictions in a consistent manner, further enhancing transparency and accountability of results. Given the black-box nature of LSTM models, we apply TIMESHAP (Time Series Shapley Values) as an explanatory model to calculate the attributes of feature, timestep, and cell levels, providing a comprehensive understanding of the model's behavior [5]. TimeSHAP not only offers global and local interpretations of the LSTM model's predictions but also enhances our understanding of how the LSTM model processes time series data. However, we have encountered challenges related to the availability of official guidance for the functions within this module, leading to difficulties when implementing TimeSHAP.

TABLE XVI: Top Global Features in Traditional ML and LSTM

Rank	Random Forest	Logistic Regression	Gradient Boosting	LSTM
1	**Lactate**	**Neutophils**	**Lactate**	**Lymphocytes**
2	Creatinine	**Lymphocytes**	**Neutophils**	**Neutophils**
3	**Lymphocytes**	RDW	**White Blood Cells**	RDW
4	SOFA	Red Blood Cells	**Lymphocytes**	Urea Nitrogen
5	**Neutophils**	**Lactate**	RDW	Platelet Count
6	RDW	INR(PT)	Creatinine	**White Blood Cells**
7	**White Blood Cells**	PT	**Bilirubin, Total**	Creatinine
8	**Bilirubin, Total**	**Bilirubin, Total**	SOFA	**Bilirubin, Total**
9	Urea Nitrogen	**White Blood Cells**	Hemoglobin	SOFA
10	Hemoglobin	Bicarbonate	PT	

C. Strengths and Limitations

In this project, the strength of our feature selection methodology is underscored by the balanced integration of machine learning techniques with practical, readily available clinical measurements at RPH. This synergy ensures our model's applicability and robustness, relying not solely on clinical expertise—which could introduce

subjectivity and overlook significant predictors [9]—but also on objective, data-driven criteria. Despite this, our methods are limited by the high granularity of patient data and data structure requirements. In our methods, we only consider hourly intervals. While we consider missing values at different time points, our methods don't consider an analysis of patients subgroups at these time points. Certain cohorts might show symptoms or indicators earlier or later than others.

To ensure interpretability and trust in the model's output for effective medical decision making, we have attempted to approach this project first-and-foremost with a clear understanding and rigorous consideration of: the healthcare domain; RPH resource constraints; and how the model will be used by RPH clinicians.

From the outset, some of the target variables (sepsis diagnosed patients) used for training data points may carry the wrong labels. The results are influenced by diagnostic criteria used for identifying sepsis patients (ICD-9 codes). Given 50-60% of sepsis cases are missed in clinical coding, we likely have missing labels for sepsis patients, which were not accounted for in our tested methods. Additionally, as vital signs are not monitored across all wards, they were not included in this study. Vital signs, specifically mean arterial pressure (MAP) are however used to calculate the SOFA score. Though the SHAP results indicate its significance in model predictions, its validity is questionable given its altered nature.

V. CONCLUSION

This project delves into the intricacies of timing logistics and the data trajectory, investigating their impact on accuracy within diverse hospital settings with varied data collection methodologies. Furthermore, we probe the algorithms' resilience to data perturbations, pinpointing specific cohorts and data configurations that exhibit suboptimal performance. Our methods focus on identifying features that occur prior to the onset of sepsis.

In the clinical environment where immediate interpretability is essential, traditional models might be favoured, despite lower overall accuracy. As such, in light of the available data and resources, we contend that the Random Forest model stands out as the most suitable choice. It balances high interpretability and explainability, which is vital for healthcare experts' comprehension, while still maintaining a high performance in comparison to other tested models. Furthermore, on top of the RF model, we advocate for the incorporation of survival analysis models based on their ability to provide insights into the likelihood of sepsis onset occurring for a patient over a given time frame. These temporal insights are invaluable for healthcare experts, enabling them to anticipate patient needs and intervene proactively. Moreover, survival analysis accounts for censored data, ensuring that all available information is used effectively. In the context of sepsis prediction, it offers a more comprehensive view of patient risk, complementing the predictions made by the other models.

Our project is a very promising start for making accurate predictions of sepsis. The methods explored may be used to help improve current sepsis identification protocols, reduce the burden on RPH healthcare staff, support clinicians in making timely, informed decisions, and ultimately to help decrease the mortality rate of patients.

VI. FUTURE IMPROVEMENTS

Beyond this project, we believe better model performance could be obtained if we could (i) consider more advanced data processing techniques, such as Synthetic Minority Oversampling Technique (SMOTE) for handling imbalanced class; (ii) perform more detailed features selection based on missing data and model performance; (iii) interrogate more thoroughly on the strength of the associations between comorbidities of sepsis with sepsis using causal inference models; (iv) train and compare more ML algorithms, including XGBoosting; (v) put more efforts in enhancing the interpretability of our LSTM model, such as finding more representative explanatory models; (vi) leverage NLP techniques for the unstructured clinical note data, particularly those available after patient triage, including employing transformers (BERT) to provide context-aware embeddings to identify features or patterns indicative of sepsis, such as symptoms, observations, and relevant past medical history, including conditions and sources of initial infection; and (vii) conduct a thorough error analysis using UMAP to better identify specific patient cohorts that may have been missed by the models, given 50-60% of sepsis cases are missed in clinical coding.

Moreover, we could improve our model in providing real-time sepsis risk prediction. However, we also aware that integrating the proposed models with the HIVE EHR systems for real-time predictions comes with serious ethical and logistical considerations.

ACKNOWLEDGMENTS

This work was supported by resources provided by The Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.

Eun-Jung Holden Shiv Akarsh Meka, Health in a Virtual Environment (HIVE), Royal Perth Hospital Jonathon Burcham, Clinical Nurse Manager - Emergency Research, Department of Health (WA Health)

Appendix A: ERD of dataset

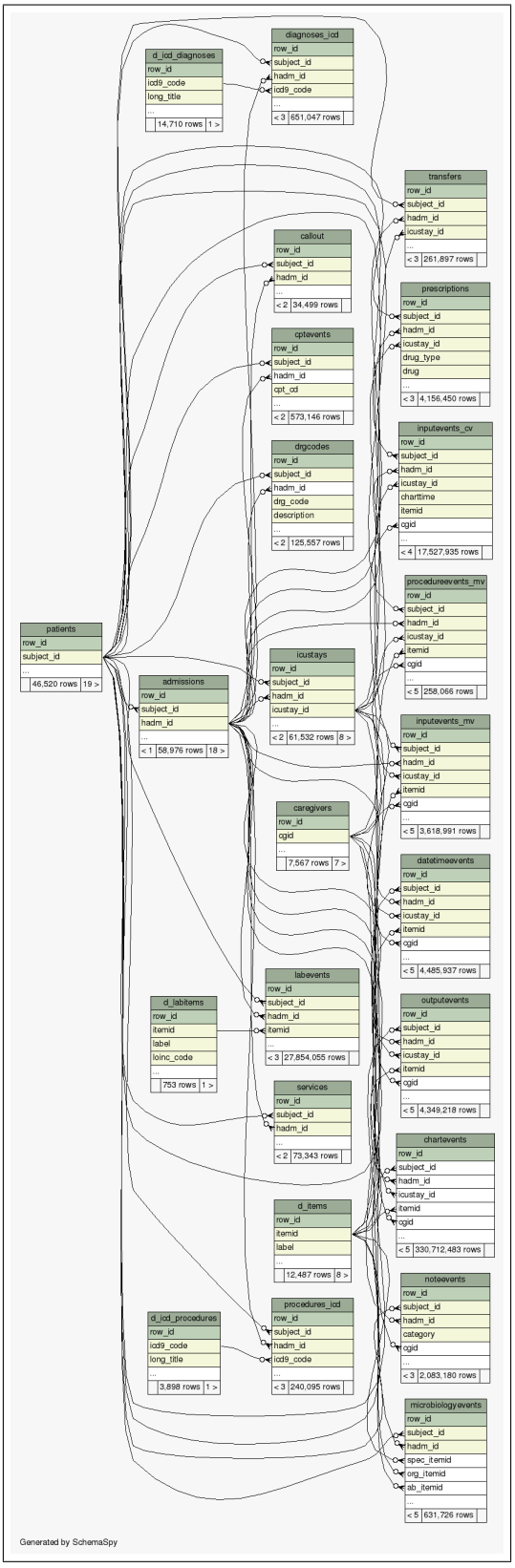


FIG. 21: Diagram sourced from [1]

[1] SchemaSpy - mimic.mimiciii - All Relationships.

[2] ICD - ICD-10-CM - International Classification of Diseases,(ICD-10-CM/PCS Transition), June 2023.

[3] Roy Adams, Katharine E. Henry, Anirudh Sridharan, Hossein Soleimani, Andong Zhan, Nishi Rawat, Lauren Johnson, David N. Hager, Sara E. Cosgrove, Andrew Markowski, Eili Y. Klein, Edward S. Chen, Mustapha O. Saheed, Maureen Henley, Sheila Miranda, Katrina Houston, Robert C. Linton, Anushree R. Ahluwalia, Albert W. Wu, and Suchi Saria. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nature Medicine*, 28(7):1455–1460, July 2022.

[4] Martin K Angele, Sebastian Pratschke, William J Hubbard, and Irshad H Chaudry. Gender differences in sepsis: Cardiovascular and immunological aspects. *Virulence*, 5(1):12–19, January 2014.

[5] João Bento, Pedro Saleiro, André Ferreira Cruz, Mário A. T. Figueiredo, and Pedro Bizarro. Timeshap: Explaining recurrent models through sequence perturbations. *CoRR*, abs/2012.00073, 2020.

[6] Ryan J. Delahanty, JoAnn Alvarez, Lisa M. Flynn, Robert L. Sherwin, and Spencer S. Jones. Development and Evaluation of a Machine Learning Model for the Early Identification of Patients at Risk for Sepsis. *Annals of Emergency Medicine*, 73(4):334–344, April 2019.

[7] Brandon DeShon, Benjamin Dummitt, Joshua Allen, and Byron Yount. Prediction of sepsis onset in hospital admissions using survival analysis. *Journal of Clinical Monitoring and Computing*, 36(6):1611–1619, December 2022.

[8] Muhammad Faisal, Andy Scally, Donald Richardson, Kevin Beatson, Robin Howes, Kevin Speed, and Mohammed A. Mohammed. Development and External Validation of an Automated Computer-Aided Risk Score for Predicting Sepsis in Emergency Medical Admissions Using the Patient's First Electronically Recorded Vital Signs and Blood Test Results*. *Critical Care Medicine*, 46(4):612–618, April 2018.

[9] Ernest V. Garcia, J. Larry Klein, and Andrew T. Taylor. Clinical decision support systems in myocardial perfusion imaging. *Journal of Nuclear Cardiology*, 21(3):427–439, June 2014.

[10] Richert E Goyette, Nigel S Key, and E Wesley Ely. Hematologic Changes in Sepsis and Their Therapeutic Implications. *Seminars in Respiratory and Critical Care Medicine*, 25(06):645–659, December 2004.

[11] Katharine E. Henry, David N. Hager, Peter J. Pronovost, and Suchi Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*, 7(299), August 2015.

[12] Nianzong Hou, Mingzhe Li, Lu He, Bing Xie, Lin Wang, Rumin Zhang, Yong Yu, Xiaodong Sun, Zhengsheng Pan, and Kai Wang. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: A machine learning approach using XGboost. *Journal of Translational Medicine*, 18(1):462, December 2020.

[13] Hye Jin Kam and Ha Young Kim. Learning representations for the early detection of sepsis with deep neural networks. *Computers in Biology and Medicine*, 89:248–255, October 2017.

[14] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, December 2019.

[15] Pei-Chen Lin, Kuo-Tai Chen, Huan-Chieh Chen, Md Mohaimenul Islam, and Ming-Chin Lin. Machine learning model to identify sepsis patients in the emergency department: Algorithm development and validation. *Journal of Personalized Medicine*, 11(11):1055, 2021.

[16] Vincent X. Liu, Vikram Fielding-Singh, John D. Greene, Jennifer M. Baker, Theodore J. Iwashyna, Jay Bhattacharya, and Gabriel J. Escobar. The Timing of Early Antibiotics and Hospital Mortality in Sepsis. *American Journal of Respiratory and Critical Care Medicine*, 196(7):856–863, October 2017.

[17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[18] Pedro Sanchez, Jeremy P. Voisey, Tian Xia, Hannah I. Watson, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, August 2022.

[19] Deepak Kumar Sharma, Parul Lakhotia, Paras Sain, and Shikha Brahmachari. Early prediction and monitoring of sepsis using sequential LONG SHORT TERM MEMORY model. *Expert Systems*, 39(3), March 2022.

[20] Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom Van Der Poll, Jean-Louis Vincent, and Derek C. Angus. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):801, February 2016.

[21] Ethan A. T. Strickler, Joshua Thomas, Johnson P. Thomas, Bruce Benjamin, and Rittika Shamsuddin. Exploring a global interpretation mechanism for deep learning networks when predicting sepsis. *Scientific Reports*, 13(1):3067, February 2023.

[22] Yingjie Su, Cuirong Guo, Shifang Zhou, Changluo Li, and Ning Ding. Early predicting 30-day mortality in sepsis in MIMIC-III by an artificial neural networks model. *European Journal of Medical Research*, 27(1):294, December 2022.

[23] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure: On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine (see contributors to the project in the appendix). *Intensive Care Medicine*, 22(7):707–710, July 1996.

[24] Dong Wang, Jinbo Li, Yali Sun, Xianfei Ding, Xiaojuan Zhang, Shaohua Liu, Bing Han, Haixu Wang, Xiaoguang Duan, and Tongwen Sun. A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients. *Frontiers in Public Health*, 9:754348, October 2021.

[25] with the HITEC Investigators, Jessica S. Ancker, Alison Edwards, Sarah Nosal, Diane Hauser, Elizabeth Mauer, and Rainu Kaushal. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision

support system. *BMC Medical Informatics and Decision Making*, 17(1):36, December 2017.

[26] Teh Xuan Ying and Asma Abu-Samah. Early prediction of sepsis for icu patients using gradient boosted tree. In *2022 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pages 78–83. IEEE, 2022.