# Early Sepsis Prediction of ICU patients using Clinical Data

*An ML-based Approach Leveraging PhysioNet Dataset*


**Team Members:**

Advika Lakshman
Himanshu Pendyala
Vishnu Vardhan
Vidhi sharma
Vignesh Venkatesan


CORPORATE GURUKUL

# Introduction

CORPORATE GURUKUL

## ***What is Sepsis?***

Sepsis is a critical condition arising from the body's extreme response to infection. It can lead to tissue damage, organ failure, and death if not treated promptly.

## ***Why is early detection vital?***

Each hour of delayed treatment increases mortality by **8%**.
Early signs are often subtle and missed by clinicians.
Timely intervention can save lives and reduce ICU burden.

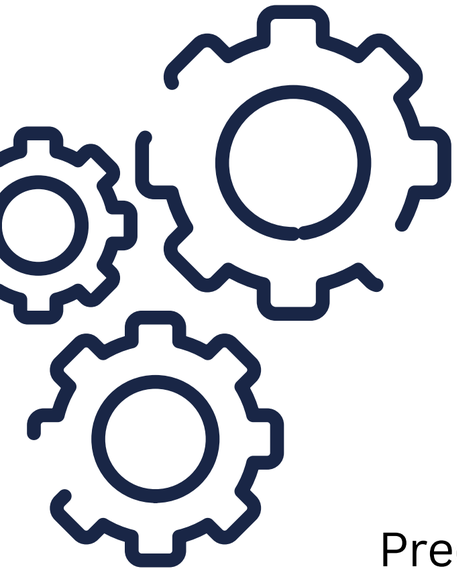
## ***Motivation for AI in Healthcare:***

With advancements in machine learning, predictive models can now analyse complex ICU data to alert clinicians before sepsis escalates.

# Problem Statement

**AIM:**

Build a machine learning model to predict sepsis onset in advance based on hourly clinical data of ICU patients.



**CHALLENGE:**

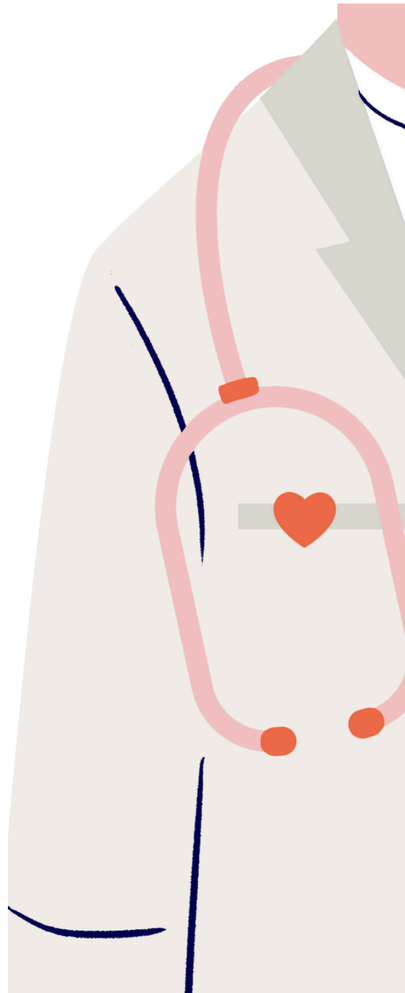
Large-scale, high-dimensional, irregular data

Significant missingness and imbalance in sepsis labels

Need for real-time, interpretable predictions in clinical settings

**SIGNIFICANCE:**

Predictive models can become crucial decision-support tools in Intensive Care Units.



CORPORATE GURUKUL

# Business Use Case

## Healthcare Transformation via AI

- Reduces ICU stays and associated healthcare costs
- Minimizes complications by ensuring early intervention
- Enables scalable solutions across hospitals globally

## Application

A real-time alert system integrated into hospital Electronic Health Records (EHRs) to notify clinicians about high-risk sepsis patients.

## Stakeholders Benefited

- ICU Doctors and Nurses
- Hospital Administrators
- Patients and Families
- Insurance Providers

CORPORATE GURUKUL

# Data Overview

***Source:***

***PhysioNet Dataset***

***A large public dataset of ICU patients from multiple*** **hospitals**

**Dataset Characteristics:**

1. 20,000+ patient files, one file per patient
2. Pipe-delimited (.psv) format with a fixed header
3. Each row = 1 hour of ICU stay
4. ~800,000 hourly entries in total
5. 40 time-dependent variables: vital signs, lab tests, and demographics
6. Target Variable: Binary label indicating whether sepsis was diagnosed during ICU stay
7. Missing Data: Represented as NaN due to irregular clinical measurements

CORPORATE GURUKUL

# Dataset

| | HR | O2Sat | Temp | SBP | MAP | DBP | Resp | EtCO2 | BaseExcess | HCO3 | ... | Fibrinogen | Platelets | Age | Gender | Unit1 | Unit2 | HospAdmTime | ICULOS | SepsisLabel | patient_id |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97.0 | 95.0 | NaN | 98.0 | 75.33 | NaN | 19.0 | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 2.0 | 0.0 | 1 |
| 1 | 89.0 | 99.0 | NaN | 122.0 | 86.00 | NaN | 22.0 | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 3.0 | 0.0 | 1 |
| 2 | 90.0 | 95.0 | NaN | NaN | NaN | NaN | 30.0 | NaN | 24.0 | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 4.0 | 0.0 | 1 |
| 3 | 103.0 | 88.5 | NaN | 122.0 | 91.33 | NaN | 24.5 | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 5.0 | 0.0 | 1 |
| 4 | 110.0 | 91.0 | NaN | NaN | NaN | NaN | 22.0 | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 6.0 | 0.0 | 1 |
| 5 | 108.0 | 92.0 | 36.11 | 123.0 | 77.00 | NaN | 29.0 | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 7.0 | 0.0 | 1 |
| 6 | 106.0 | 90.5 | NaN | 93.0 | 76.33 | NaN | 29.0 | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 8.0 | 0.0 | 1 |
| 7 | 104.0 | 95.0 | NaN | 133.0 | 88.33 | NaN | 26.0 | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 9.0 | 0.0 | 1 |
| 8 | 102.0 | 91.0 | NaN | 134.0 | 87.33 | NaN | 30.0 | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 10.0 | 0.0 | 1 |
| 9 | 104.0 | 92.0 | 37.17 | 138.0 | 86.67 | NaN | 19.0 | NaN | 23.0 | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 11.0 | 0.0 | 1 |
| 10 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 12.0 | 0.0 | 1 |
| 11 | 102.0 | 93.0 | NaN | 129.0 | 77.00 | NaN | 24.0 | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 13.0 | 0.0 | 1 |
| 12 | 108.0 | 90.0 | NaN | 122.0 | 96.67 | NaN | 27.0 | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 14.0 | 0.0 | 1 |
| 13 | 106.0 | 90.0 | NaN | NaN | NaN | NaN | 25.0 | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 15.0 | 0.0 | 1 |
| 14 | 109.0 | 91.0 | 36.56 | 132.0 | 96.67 | NaN | 24.0 | NaN | NaN | NaN | ... | NaN | NaN | 83.14 | 0.0 | NaN | NaN | -0.03 | 16.0 | 0.0 | 1 |

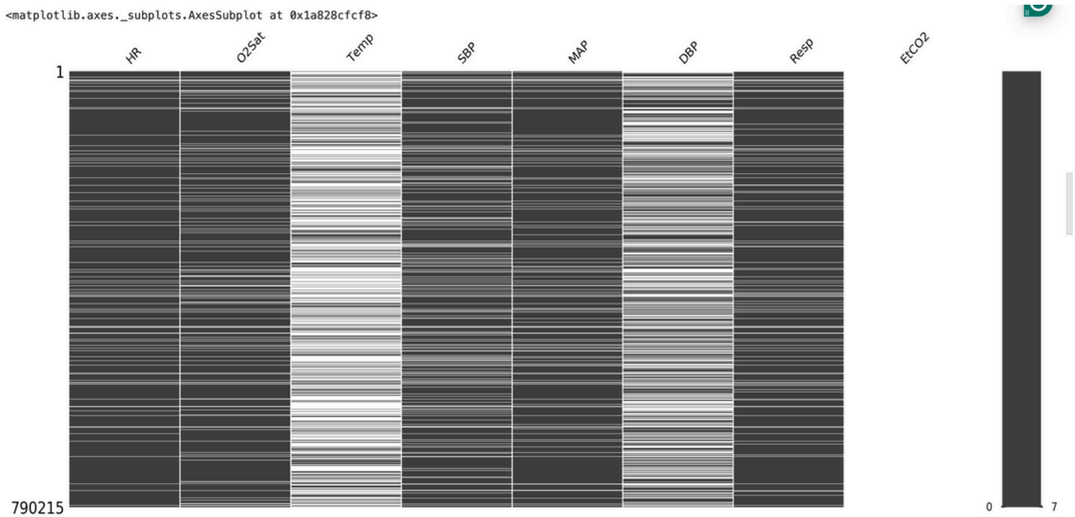
# Exploratory Data Analysis

- The dataset is heavily imbalanced: only around 2% of the data points are labeled as septic.
- Some vitals like HR, O2Sat, and Resp are consistently recorded, while labs like TroponinI, Bilirubin, and AST have high levels of missing data.
- Many features show non-normal distributions, often skewed due to the nature of ICU monitoring.
- Outliers are present, particularly in features like HR, Lactate, and WBC, reflecting acute patient states.

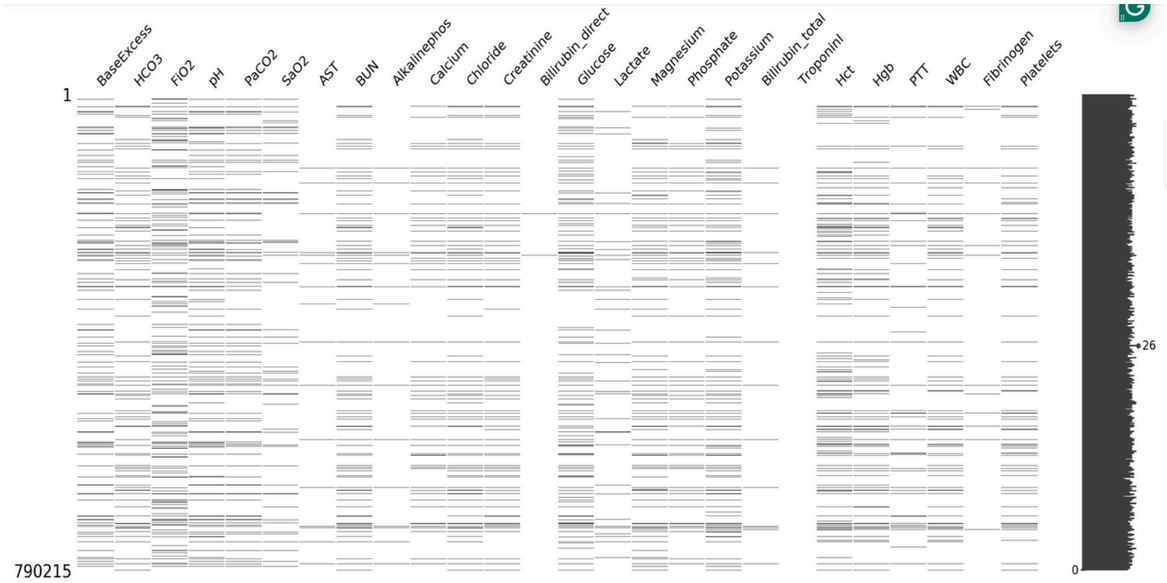
**Key insights:**

- Septic patients tend to have higher values of Lactate, WBC, and longer ICULOS.
- Non-septic patients exhibit more stable vital signs over time.

# Missingness Matrix



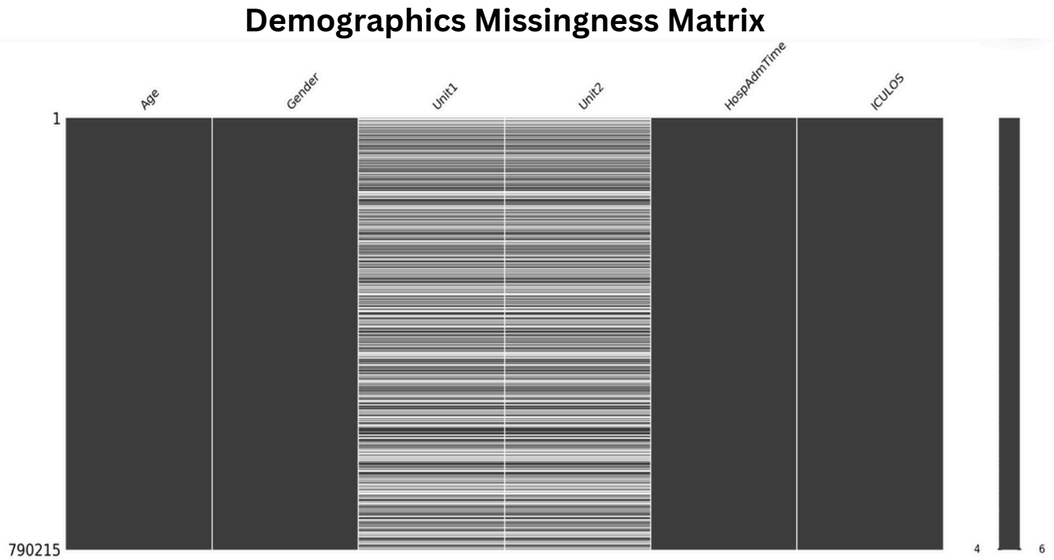

Vitals Missingness Matrix




Labs Missingness Matrix

CORPORATE GURUKUL

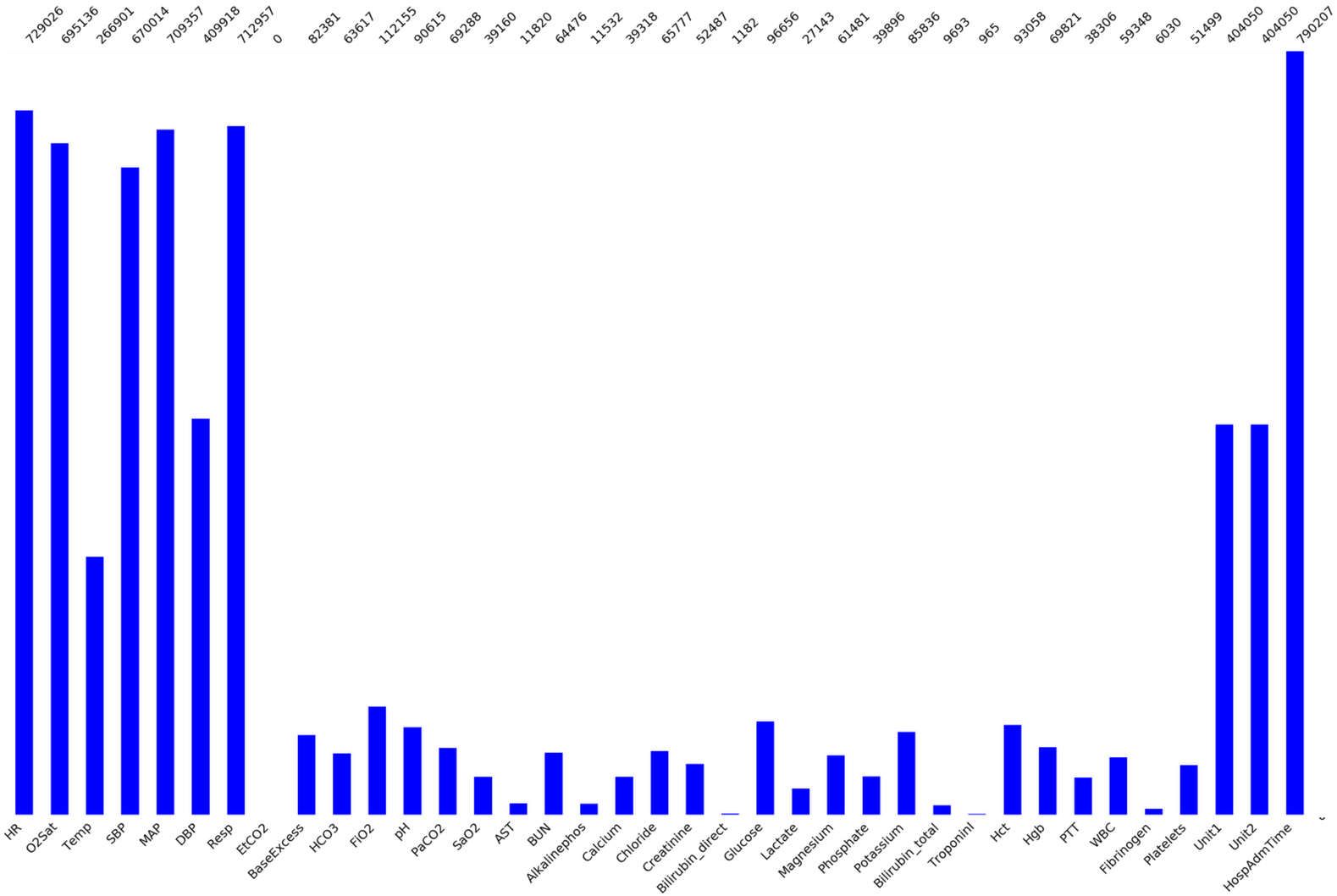
# Missingness Matrix




**Key Insights: Labs are sparsely sampled (high missing rate), vitals are moderately complete, demographics are nearly fully populated**

CORPORATE GURUKUL

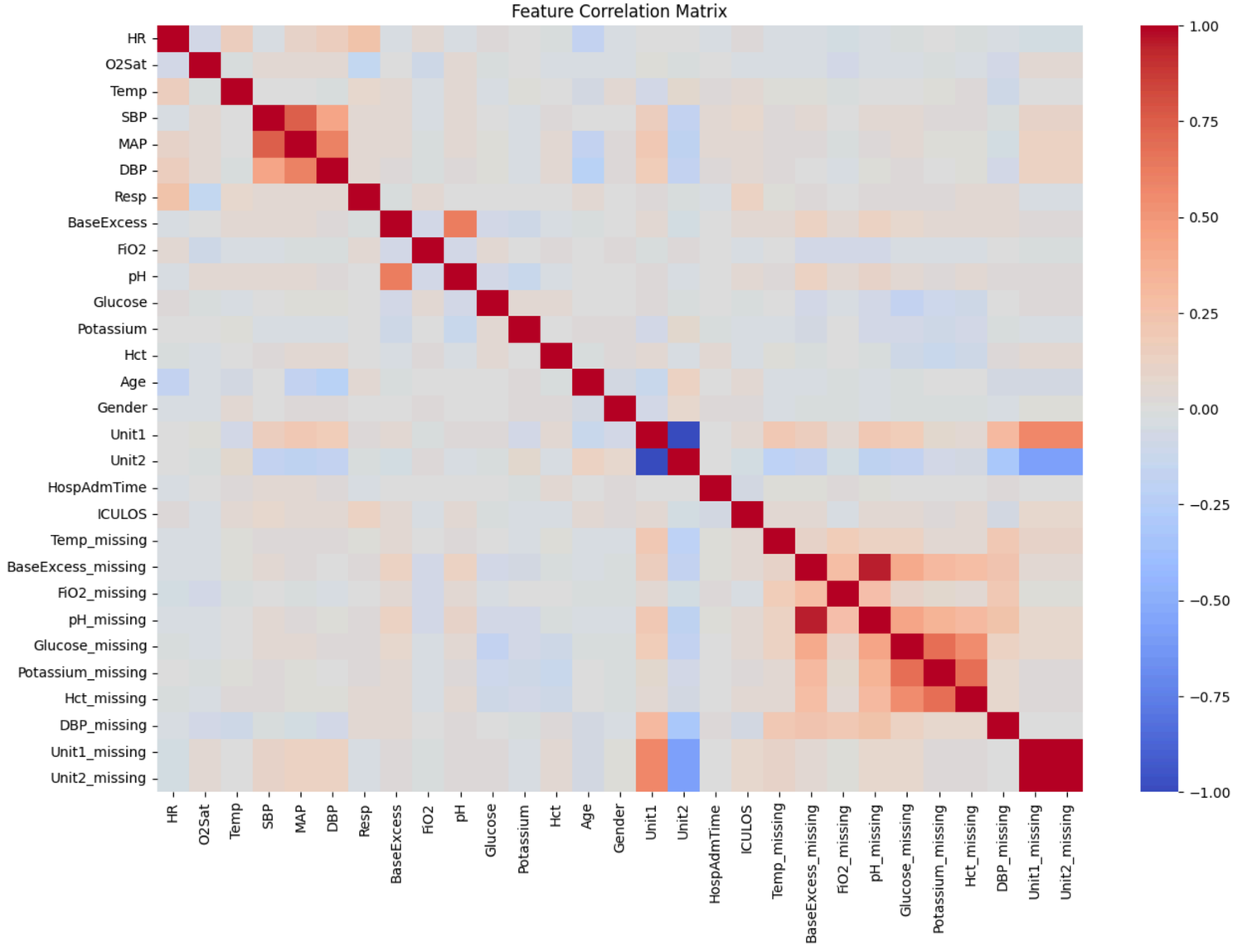
# Barplot




**Count of Non-missing Values**

CORPORATE GURUKUL

# Correlation Matrix




CORPORATE GURUKUL

# Data Preprocessing

- Several preprocessing steps were applied to prepare the data for modeling.
- Handling Missing Values.
- Features with excessive missingness (>90%) were dropped and binary masking was applied.
- Others were imputed using median(50-90%) or forward-fill techniques(<50%).
- Feature Scaling: transforms the values of independent variables (features) to a similar scale, ensuring all features contribute equally to the model.
- StandardScaler was used to normalize the data .
- Dealing with Imbalance: class weight balance.
- Train/Test Split: 80-20.

# Methodology

**The following models were used:**

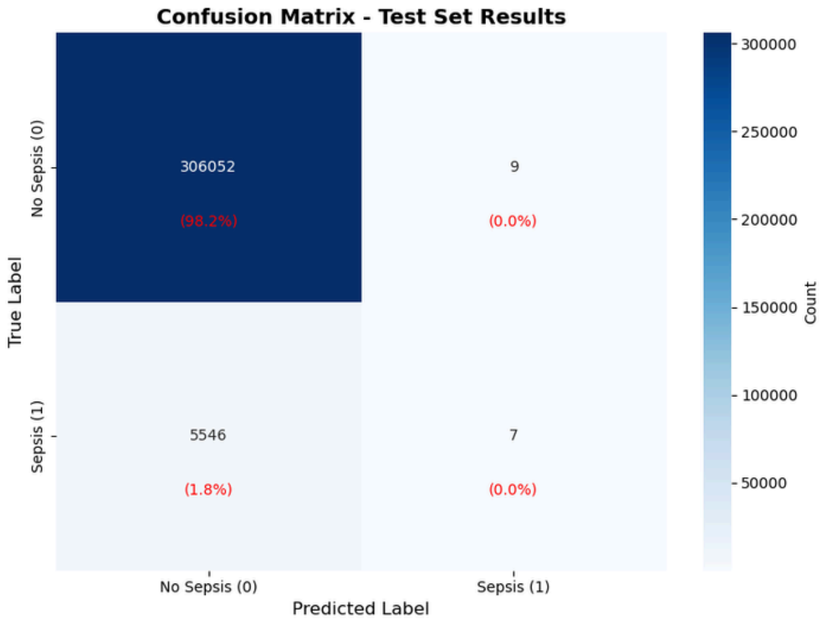
- Logistic Regression
- XGBoost
- XGBoost with custom interpolation methods
- XGBoost with recall optimisation

CORPORATE GURUKUL

# Logistic Regression

- **Data Pipeline**: Loaded cleaned CSVs, split out SepsisLabel as the target, and scaled all features to standardize ranges.
- **Model Setup & Training**: Initialized LogisticRegression (with a high max_iter), then fit on the scaled training data.
- **Evaluation**: Predicted on the test set using a 0.5 threshold, and reported accuracy, precision, recall, F1, plus a confusion matrix to assess positive-class detection

# Results and Analysis

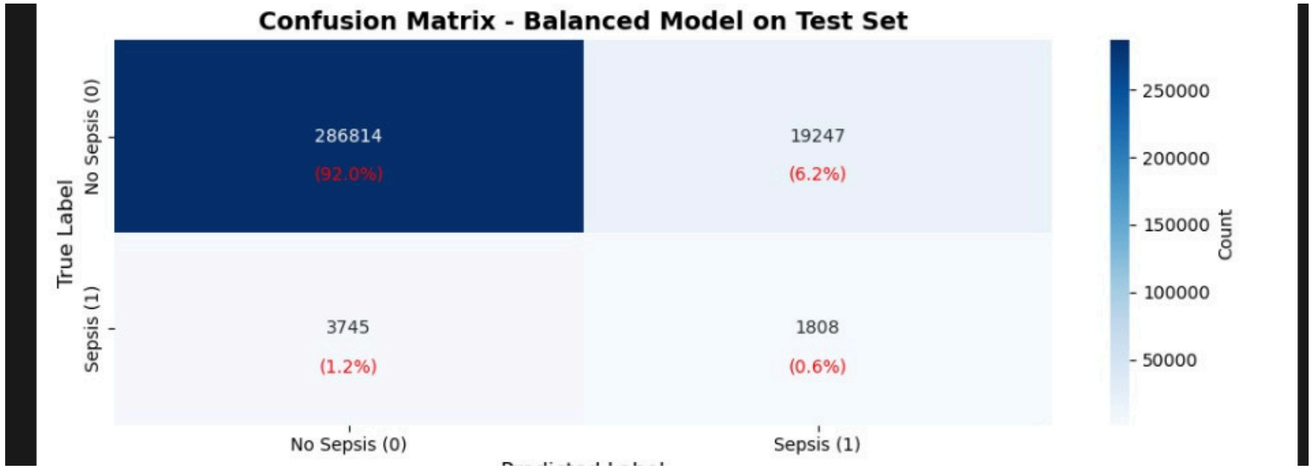



```
=== TRAINING vs TEST ACCURACY ===
Training Accuracy: 0.9819
Test Accuracy: 0.9822
Difference: -0.0003
✅ Good: similar training and test performance
```

```
=== TEST SET PERFORMANCE METRICS ===
Accuracy: 0.9822
Precision: 0.4375
Recall: 0.0013
F1 Score: 0.0025
Number of False Negatives: 5546

=== CONFUSION MATRIX BREAKDOWN ===
True Negatives (TN): 306052
False Positives (FP): 9
False Negatives (FN): 5546
True Positives (TP): 7
```

CORPORATE GURUKUL

# 70-30 balanced results



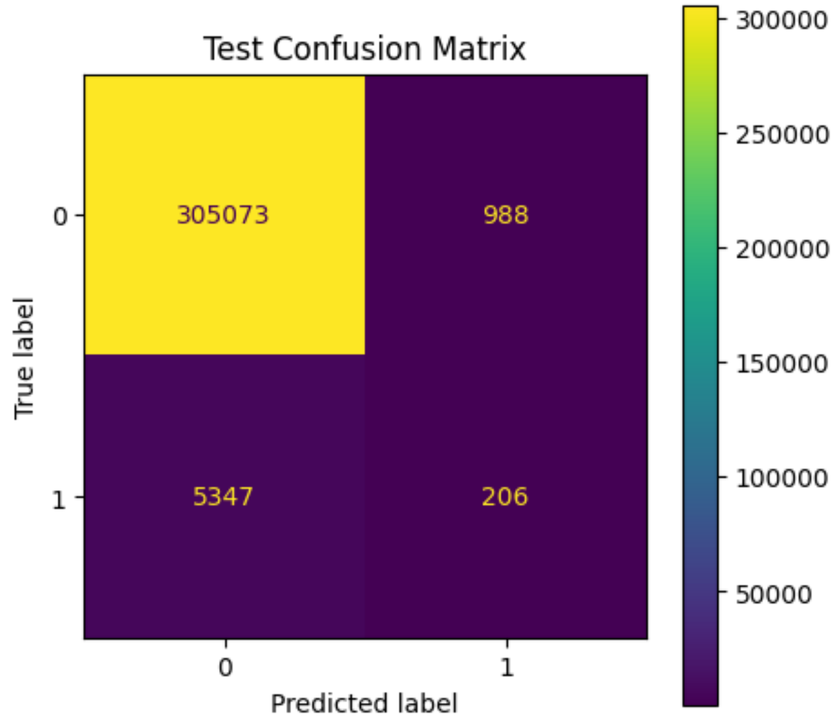

```
=== TEST SET PERFORMANCE METRICS
Accuracy: 0.9262
Precision: 0.0859
Recall: 0.3256
F1 Score: 0.1359
Number of False Negatives: 3745
```

```
=== TRAINING vs TEST ACCURACY (BALANCED MODEL) ===
Training Accuracy: 0.7572
Test Accuracy: 0.9262
Difference: -0.1691
⚠️  Unusual: test accuracy higher than training accuracy
```

# XGBoost

- **Fast, Accurate Predictions**: XGBoost’s boosted tree ensemble quickly captures complex patterns in ICU time-series data, helping detect sepsis onset earlier.
- **Robust to Missing Data**: It natively handles gaps in lab and vital sign measurements, so we can train directly without elaborate imputation.
- **Feature Importance Insights**: Built-in importance scores highlight which measurements (e.g., BUN, MAP, $O_2$Sat) drive sepsis risk, guiding clinical interpretation.
- **Regularization Controls Overfitting**: L1/L2 penalties and tree-shrinkage parameters keep the model generalizable to new patients, crucial for reliable deployment in the ICU

# Result and Analysis




```
=== TRAINING SET ===
Accuracy        : 0.9833
Precision       : 0.8696
Recall          : 0.0889
F1 Score        : 0.1612
False Negatives : 20376

=== TEST SET ===
Accuracy        : 0.9797
Precision       : 0.1725
Recall          : 0.0371
F1 Score        : 0.0611
False Negatives : 5347
```

CORPORATE GURUKUL

# XGBoost with custom interpolation

- **Two-Stage Imputation Strategy**

– Vital signs (HR, $O_2$Sat, Temp, SBP, MAP, DBP, Resp, Et$CO_2$) filled via **linear interpolation** to maintain smooth temporal trends

– All other lab features imputed with **MICE (IterativeImputer)** to capture multivariate correlations

- **Feature Assembly**

– Recombined interpolated vitals and MICE-imputed labs into a single, complete feature matrix for both train and test sets

- **Model Training**

– Trained the same XGBClassifier (log-loss objective, fixed seed) on the fully imputed data without dropping any rows

Evaluation Framework

– Generated train/test predictions and computed accuracy, precision, recall, F1, and false-negative counts

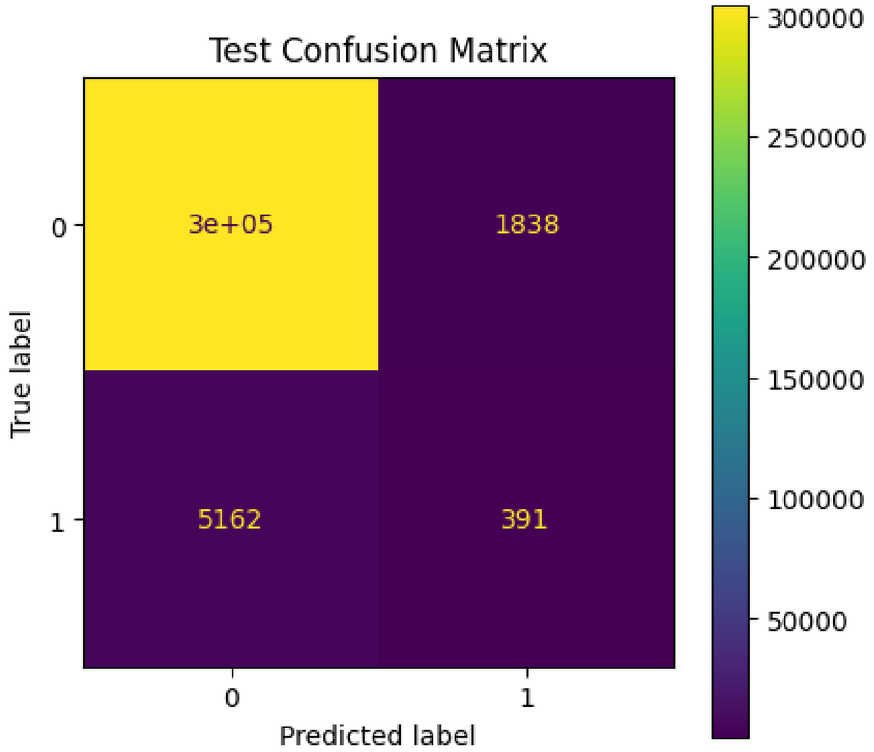
– Plotted confusion matrices to visually assess error trade-offs

- **Key Benefit**

– By addressing missingness explicitly, we supply XGBoost with richer, more reliable inputs—crucial for improving sepsis detection sensitivity in ICU time-series data.

CORPORATE GURUKUL

# Result and Analysis



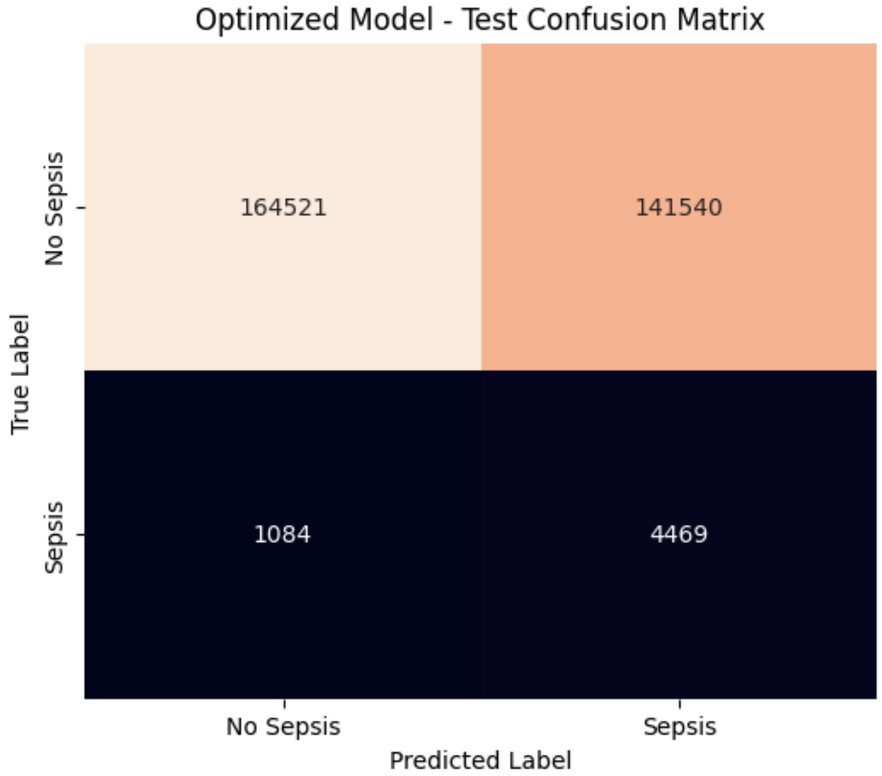

```
=== TRAINING SET ===
Accuracy        : 0.9838
Precision       : 0.8646
Recall          : 0.1211
F1 Score        : 0.2124
False Negatives : 19655

=== TEST SET ===
Accuracy        : 0.9775
Precision       : 0.1754
Recall          : 0.0704
F1 Score        : 0.1005
False Negatives : 5162
```

CORPORATE GURUKUL

# XGBoost with recall optimisation

- **Class-Balance Calibration**: Incorporating scale_pos_weight in the search space ensures the model explicitly up-weights rare sepsis cases, dramatically improving recall versus the default (which largely ignored positives).
- **Recall-Driven Tuning:** By using recall as the optimization metric in cross-validation, we bias the hyperparameter search toward configurations that catch more true sepsis events, rather than just maximizing overall accuracy.
- **Automated Hyperparameter Exploration**: Randomized search over a broad range of tree depths, learning rates, subsample ratios, and regularization terms finds a much better bias–variance trade-off than one-off, manually chosen defaults.
- **Cross-Validation Robustness:** Three-fold CV during tuning guards against overfitting; the final model generalizes better to unseen test data than the single-split default model.
- **Measurable Sensitivity Gain:** The optimized model yields higher test recall and F1 (with fewer false negatives) compared to the untuned version—directly translating to more reliable early sepsis alerts in the ICU.

# Result and Analysis




```
=== TEST SET METRICS ===
Accuracy         : 0.5423
Precision        : 0.0306
Recall           : 0.8048
F1 Score         : 0.0590
False Negatives  : 1084
Best recall (CV): 0.6865
```

CORPORATE GURUKUL

# Challenges

**Challenges in dataset :**

- Dataset is highly imbalanced with only 2% sepsis patients.
- lots of nan values
- dataset from only 3 different hospitals
- very low or very high f1 score due to unbalanced dataset
- time series format adds more complexity
- computational complexity because of large data and huge number of features

CORPORATE GURUKUL

# References


- Reyna, Matthew, et al. "Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019" (version 1.0.0). PhysioNet (2019). RRID:SCR_007345.
- The Signature-Based Model for Early Detection of Sepsis From Electronic Health Records in the Intensive Care Unit James Morrill1,3 , Andrey Kormilitzin1,2 , Alejo Nevado-Holgado2 , Sumanth Swaminathan3 , Sam Howison1 , Terry Lyons1
- Early Prediction of Sepsis Using Multi-Feature Fusion Based XGBoost Learning and Bayesian Optimization Meicheng Yang1 , Xingyao Wang1 , Hongxiang Gao1 , Yuwen Li1 , Xing Liu2 , Jianqing Li1* , Chengyu Liu1*
- Representation Learning for Early Sepsis Prediction Luan Tran1 , Manh Nguyen2 , Cyrus Shahabi1

# THANK YOU

CORPORATE GURUKUL