JP DASOLT
Going Beyond Numbers

Automated Medical Diagnosis: Early Prediction of Sepsis using Electronic Health Records (EHR)

Nour El Imane S.

Thursday 16th May, 2024

Abstract

This research paper presents the development and implementation of a predictive early sepsis detection system leveraging machine learning and deep learning methodologies. Through rigorous experimentation, the Random Forest model emerged as the optimal choice, exhibiting superior precision (96.9%) and F1 score (34.2%) compared to alternative models. Addressing challenges such as processing large datasets and handling data imbalance was integral to the study.

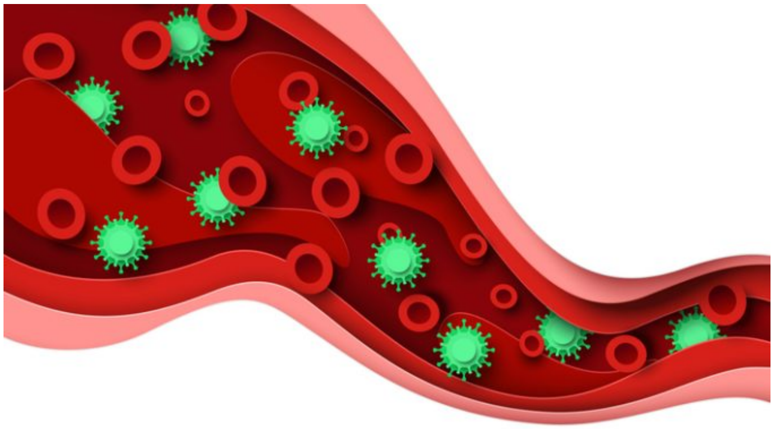
The clinical implications of this predictive system are profound, offering the potential for early sepsis detection, thereby facilitating timely interventions and potentially saving lives while alleviating strain on healthcare resources. Integration of the machine learning model into clinical workflows promises real-time monitoring and decision support, thus enhancing the quality of patient care.

1. Introduction

Sepsis is a life-threatening condition triggered by the body's response to infection [1]. Normally, when the body detects an infection, it releases chemicals into the bloodstream to fight it off. However, in sepsis, this response goes haywire, causing widespread inflammation and organ damage. Common causes of sepsis include bacterial, viral, or fungal infections like pneumonia or urinary tract infections [2].

The gravity of sepsis management became even more apparent with the release of the 2015 report from the National Confidential Enquiry into Patient Outcome and Death [3]. This report unearthed a troubling reality: in over a third of cases (36%), there were considerable delays in identifying sepsis. This revelation underscores the urgent need for improved diagnostic strategies and interventions.

This paper delves into the emergence of machine learning algorithms in medical diagnostics, a development that has garnered substantial attention in recent years. These algorithms offer promising prospects for enhancing the accuracy and efficiency of disease detection, potentially revolutionizing sepsis management and patient outcomes.



Given the elusive nature of sepsis symptoms, early detection and intervention are paramount for saving lives. Symptoms such as rapid heartbeat, fluctuating body temperature, and accelerated breathing often manifest ambiguously, complicating diagnosis. Accurate and timely identification of sepsis is pivotal for initiating targeted treatments, thereby significantly improving patient outcomes and survival rates[4].

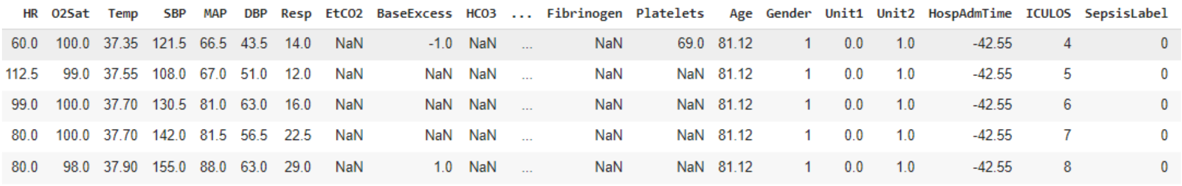
2. Objective

The objective of this project is to leverage physiological data for the early detection of sepsis, up to six hours in advance. By incorporating patient information such as vital signs, laboratory values, and demographics as inputs, the model will generate predictions to classify patients as either non-septic or septic prior to the onset of clinical symptoms.

This study aims to evaluate the efficacy of various supervised machine learning models and deep learning techniques in predicting sepsis. Through comprehensive analysis of their accuracy and reliability, we aim to gain insights into their performance in sepsis prediction. By assessing these models, our study seeks to contribute to ongoing research on the application of machine learning and deep learning in medical diagnosis, with the ultimate goal of improving the early detection of sepsis.

3. Dataset Description

For this study, we utilized clinical data of ICU patients from two distinct hospital systems made available by PhysioNet [4]. Data was collected from 40,336 patients across both hospitals. Each data file contains a table with measurements recorded over time. Upon aggregating ICU-hour-stay entries from all patients, the total dataset comprises 1544510 lines of data.



HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess	HCO3	...	Fibrinogen	Platelets	Age	Gender	Unit1	Unit2	HospAdmTime	ICULOS	SepsisLabel
60.0	100.0	37.35	121.5	66.5	43.5	14.0	NaN	-1.0	NaN	...	NaN	69.0	81.12	1	0.0	1.0	-42.55	4	0
112.5	99.0	37.55	108.0	67.0	51.0	12.0	NaN	NaN	NaN	...	NaN	NaN	81.12	1	0.0	1.0	-42.55	5	0
99.0	100.0	37.70	130.5	81.0	63.0	16.0	NaN	NaN	NaN	...	NaN	NaN	81.12	1	0.0	1.0	-42.55	6	0
80.0	100.0	37.70	142.0	81.5	56.5	22.5	NaN	NaN	NaN	...	NaN	NaN	81.12	1	0.0	1.0	-42.55	7	0
80.0	98.0	37.90	155.0	88.0	63.0	29.0	NaN	1.0	NaN	...	NaN	NaN	81.12	1	0.0	1.0	-42.55	8	0

The dataset encompasses 40 features, consisting of 8 vital signs, 26 laboratory values, and 6 demographic variables. Lastly, the outcome variable represents the presence or absence of Sepsis 1.

Table 1: Variable Descriptions

Variable	**Description**
Vital Signs	
HR	Heart rate (beats per minute)
O2Sat	Pulse oximetry (%)
Temp	Temperature (Deg C)
SBP	Systolic BP (mm Hg)
MAP	Mean arterial pressure (mm Hg)
DBP	Diastolic BP (mm Hg)
Resp	Respiration rate (breaths per minute)
EtCO2	End tidal carbon dioxide (mm Hg)
Laboratory Tests	
BaseExcess	Measure of excess bicarbonate (mmol/L)
HCO3	Bicarbonate (mmol/L)
FiO2	Fraction of inspired oxygen (%)
pH	N/A
PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
SaO2	Oxygen saturation from arterial blood (%)
AST	Aspartate transaminase (IU/L)
BUN	Blood urea nitrogen (mg/dL)
Alkalinephos	Alkaline phosphatase (IU/L)
Calcium	(mg/dL)
Chloride	(mmol/L)
Creatinine	(mg/dL)
Bilirubin_direct	Bilirubin direct (mg/dL)
Glucose	Serum glucose (mg/dL)
Lactate	Lactic acid (mg/dL)
Magnesium	(mmol/dL)
Phosphate	(mg/dL)
Potassium	(mmol/L)
Bilirubin_total	Total bilirubin (mg/dL)
TroponinI	Troponin I (ng/mL)
Hct	Hematocrit (%)
Hgb	Hemoglobin (g/dL)
PTT	Partial thromboplastin time (seconds)
WBC	Leukocyte count (count$\times 10^3/\mu$L)
Fibrinogen	(mg/dL)
Platelets	(count$\times 10^3/\mu$L)
Patient Information	
Age	Years (100 for patients 90 or above)
Gender	Female (0) or Male (1)
Unit1	Administrative identifier for ICU unit (MICU)
Unit2	Administrative identifier for ICU unit (SICU)
HospAdmTime	Hours between hospital admit and ICU admit
ICULOS	ICU length-of-stay (hours since ICU admit)
SepsisLabel	0 (Non-sepsis) and 1 (Sepsis)

4. Preprocessing and Analysis

The dataset consists of various vital, laboratory captured and demographic attributes hourly. The laboratory captured attributes are based on advice of medical practitioner and are centric towards actual patients. As a result, the dataset is extremely unbalanced and has multiple instances of missing attributes.

1.4. Cleaning Data

The initial steps in any data analysis endeavor often involve preprocessing, a critical phase to ensure data quality and reliability. Upon acquiring the dataset, the first task was to cleanse it of any unnecessary information, such as identifier columns like ’Unit1’ and ’Unit2’.

However, the primary challenge lay in handling missing values. The dataset exhibited a significant issue with null values, with a substantial portion of columns containing over 90% missing data, as depicted in Figure 1

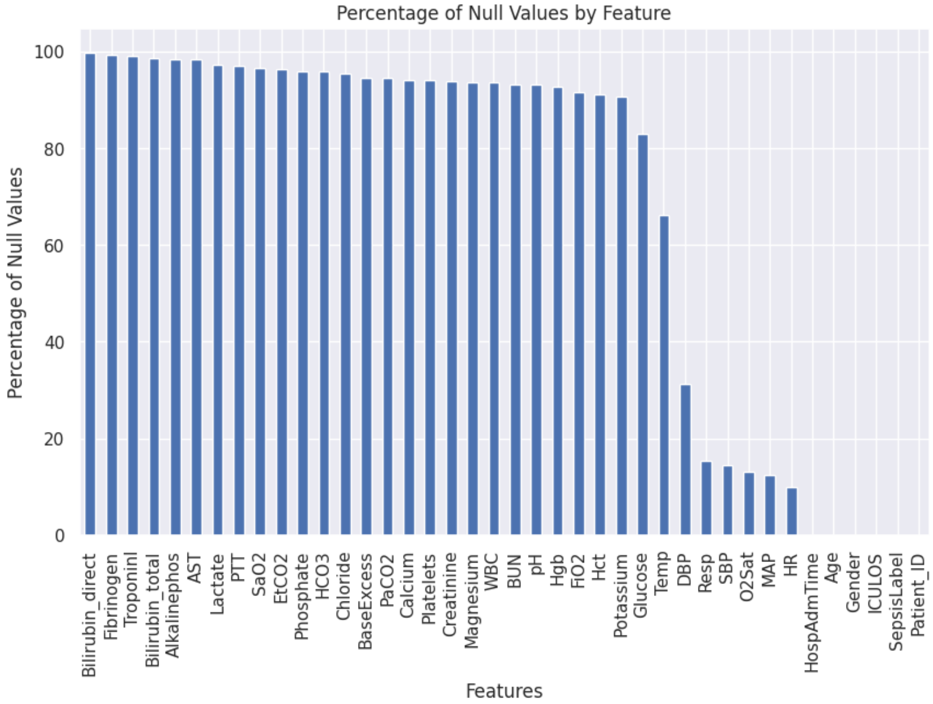
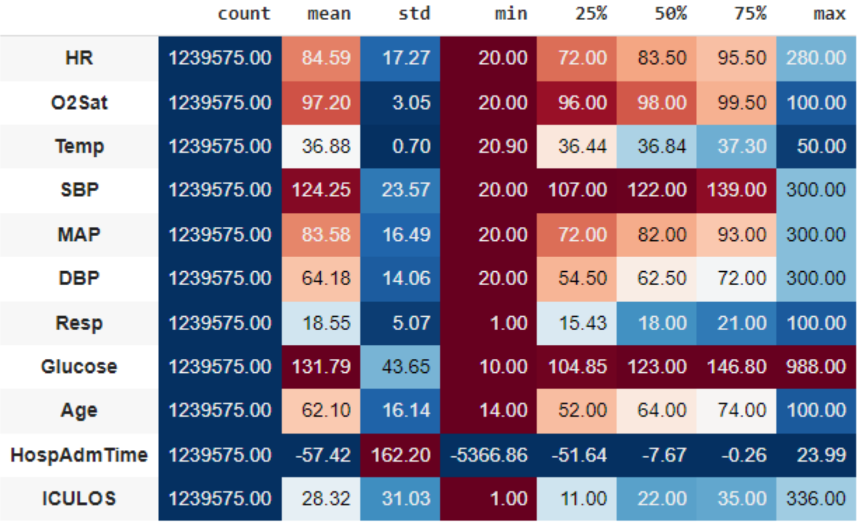


Figure 1: Percentage of Null Values by Feature

To tackle this issue effectively, I initially calculated the percentage of null values in each column. Subsequently, features with 90% or more missing data were eliminated from further analysis.

Following this initial cleanup, I proceeded to group the data by patient ID, a crucial step in organizing the dataset to facilitate subsequent analyses. Once grouped, I employed interpolation techniques followed by backward and forward filling to impute missing values effectively. Any remaining missing values were then dropped from the dataset to maintain data integrity.

To further refine the dataset, duplicate rows were removed, resulting in the final dataset statistics described in Figure 2. This refined dataset comprises 1,239,575 samples and 13 features, including the critical label indicating the presence of sepsis.



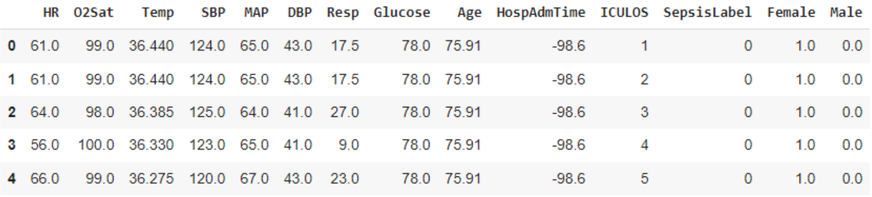
	count	mean	std	min	25%	50%	75%	max
HR	1239575.00	84.59	17.27	20.00	72.00	83.50	95.50	280.00
O2Sat	1239575.00	97.20	3.05	20.00	96.00	98.00	99.50	100.00
Temp	1239575.00	36.88	0.70	20.90	36.44	36.84	37.30	50.00
SBP	1239575.00	124.25	23.57	20.00	107.00	122.00	139.00	300.00
MAP	1239575.00	83.58	16.49	20.00	72.00	82.00	93.00	300.00
DBP	1239575.00	64.18	14.06	20.00	54.50	62.50	72.00	300.00
Resp	1239575.00	18.55	5.07	1.00	15.43	18.00	21.00	100.00
Glucose	1239575.00	131.79	43.65	10.00	104.85	123.00	146.80	988.00
Age	1239575.00	62.10	16.14	14.00	52.00	64.00	74.00	100.00
HospAdmTime	1239575.00	-57.42	162.20	-5366.86	-51.64	-7.67	-0.26	23.99
ICULOS	1239575.00	28.32	31.03	1.00	11.00	22.00	35.00	336.00

Figure 2: Describe the Dataset

2.4. Enabling Modeling

After initial preprocessing, the dataset underwent further modifications to prepare it for modeling:

- **Encoding Categorical Features:** The target variable was already encoded, and for the categorical feature 'Gender,' a one-hot encoding scheme was employed, expanding it into binary columns. Below is a representation of the dataset after encoding:



	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	Glucose	Age	HospAdmTime	ICULOS	SepsisLabel	Female	Male
0	61.0	99.0	36.440	124.0	65.0	43.0	17.5	78.0	75.91	-98.6	1	0	1.0	0.0
1	61.0	99.0	36.440	124.0	65.0	43.0	17.5	78.0	75.91	-98.6	2	0	1.0	0.0
2	64.0	98.0	36.385	125.0	64.0	41.0	27.0	78.0	75.91	-98.6	3	0	1.0	0.0
3	56.0	100.0	36.330	123.0	65.0	41.0	9.0	78.0	75.91	-98.6	4	0	1.0	0.0
4	66.0	99.0	36.275	120.0	67.0	43.0	23.0	78.0	75.91	-98.6	5	0	1.0	0.0

- **Scaling Numeric Features:** To ensure that features with different scales do not unduly influence the model, numeric features were scaled using the StandardScaler from the scikit-learn library. This transformation standardizes the distribution of each feature by removing the mean and scaling to unit variance.

- **Splitting the Data:** The dataset was split into training and testing sets with a ratio of 80% for training and 20% for testing. This ensures that the model is trained on a sufficiently large portion of the data while retaining a separate set for evaluation.

3.4. Analysis

As illustrated in Figure 3, the dataset exhibits a significant imbalance, with only 1.8% of the data representing sepsis patients. This is attributed to the rarity of sepsis cases in the population under study, which poses challenges for model training and generalization.

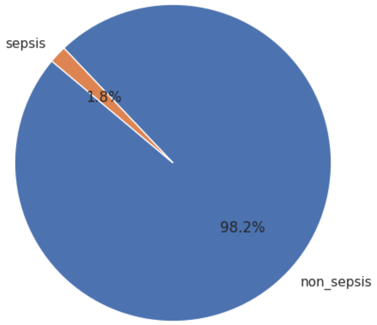


Figure 3: Distribution of Sepsis Label

The Figure 4 displays that males outnumber females in the dataset, with 699,437 instances compared to 540,138 instances, respectively, suggesting a slightly higher occurrence of sepsis among males.

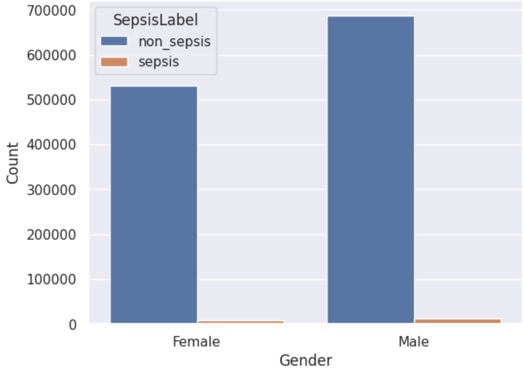


Figure 4: Distribution of Sepsis Label by Gender

Patients with sepsis tend to have a significantly longer Intensive Care Unit Length of Stay (ICULOS), often exceeding 60 hours, compared to non-sepsis patients, who typically exhibit around 27 hours. This prolonged ICULOS in sepsis cases indicates a more complex treatment and monitoring course (Figure 5).

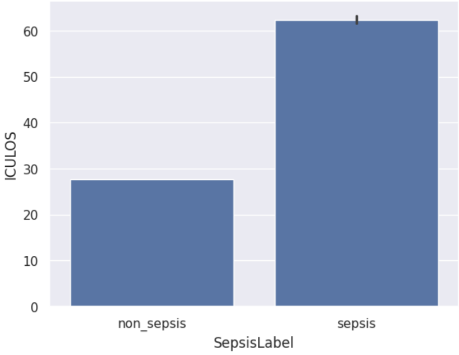


Figure 5: Sepsis Label VS ICULOS

In the correlation heatmap depicted in Figure 6, several insights emerge:

- **Heart Rate (HR):** Higher HR with higher temperature and respiration, lower with age due to metabolic demands and cardiac function changes
- **Temperature:** Higher temperature raises HR due to increased metabolic demand, slightly lowers blood pressure due to vasodilation

- **Respiration Rate (Resp):** Higher resp rates accompany higher HR and longer ICU stays due to increased metabolic demand and severity of illness

- **ICU Length of Stay (ICULOS):** Longer stays relate to higher resp rates and presence of sepsis, indicating severity of condition

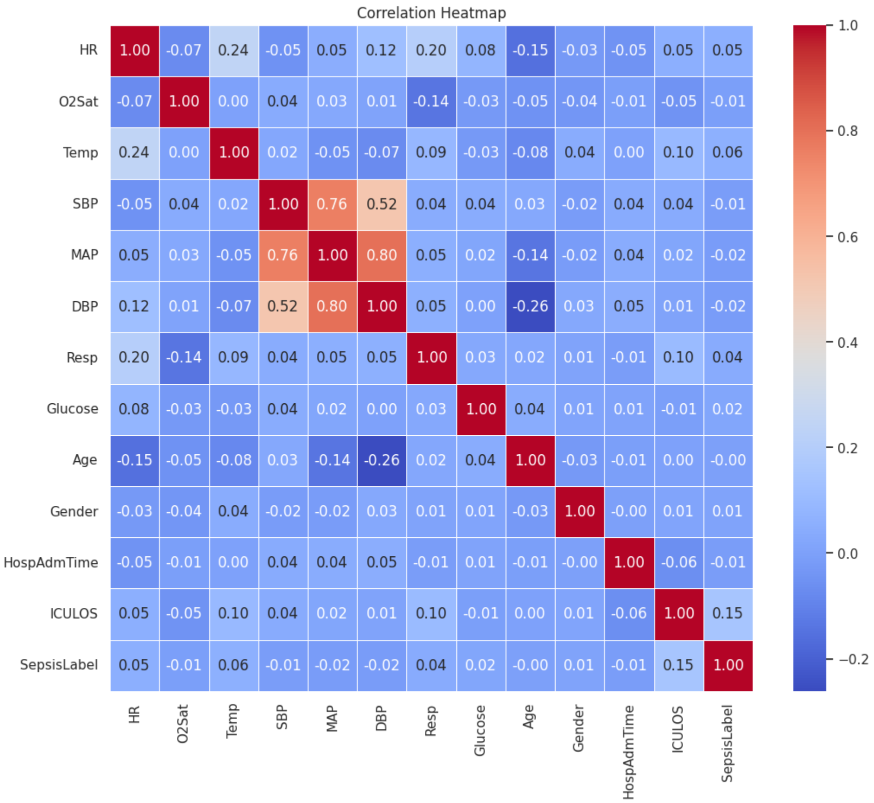


Figure 6: Correlation Heatmap

5. Methodology

The primary objective of this study is to predict sepsis and non-sepsis patients, which is framed as a *binary classification problem.* To address this, I employed a selection of machine learning (ML) and deep learning (DL) models. Specifically, **Random Forest (RF)** and **XGBoost** were chosen from ML, while **Multilayer Perceptron (MLP)** and **Recurrent Neural Networks (RNN)** with Long Short-Term Memory (LSTM) architecture were selected from DL .

The choice of these models was informed by a literature review in the field of illness diagnoses and sepsis diagnosis, where these algorithms have demonstrated efficacy.

Given the size of the dataset, I opted to utilize *Google Colab* with GPU support for both modeling and evaluation. This choice significantly expedited the computational process. Additionally, I utilized popular libraries such as *scikit-learn*, *TensorFlow*, and *Keras* for model implementation and evaluation.

In healthcare, with such imbalanced data, I will focus on metrics such as **Precision**, **F1 Score**, **Recall**, and **Confusion Matrice** because they are more effective in this case.

To mitigate the effects of class imbalance prevalent in datasets, I employed **class weights** during model training to balance the data distribution.

Hyperparameter tuning was conducted manually to optimize model performance. This process, while effective, incurred varying runtimes ranging from 10 minutes to over 2 hours, depending on the complexity of the model. Notably, RNN models required the most extensive computational resources, followed by MLP, XGBoost, and Random Forest.

6. Results & Discussion

Our four models yield varying results, with noticeable differences between them. Figure 7 depicts a comparison of model performance. We observe that all models exhibit perfect accuracy, indicating their ability to correctly predict the non-sepsis class. This is because of the larger number of non-sepsis samples compared to sepsis samples. Additionally, across all models, other metrics appear weak except for precision in the random forest model.

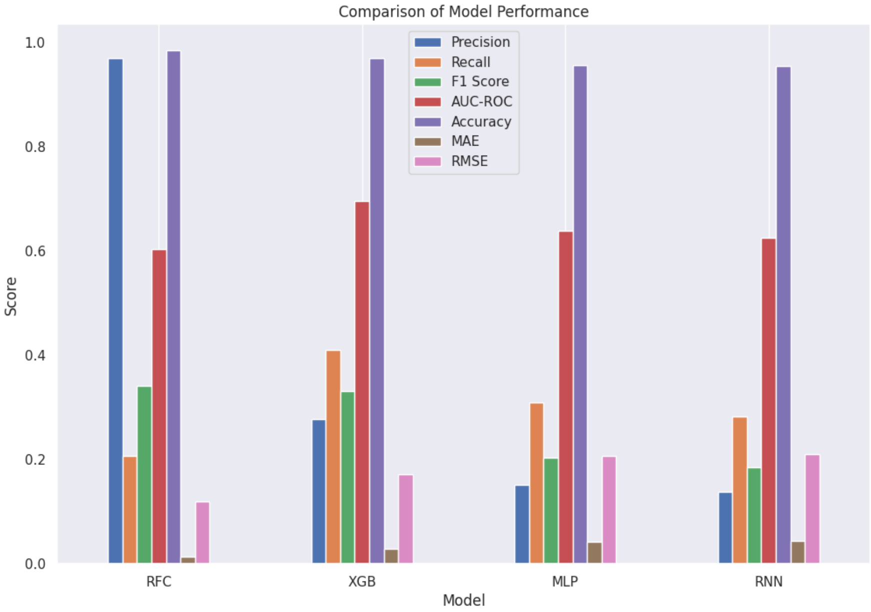
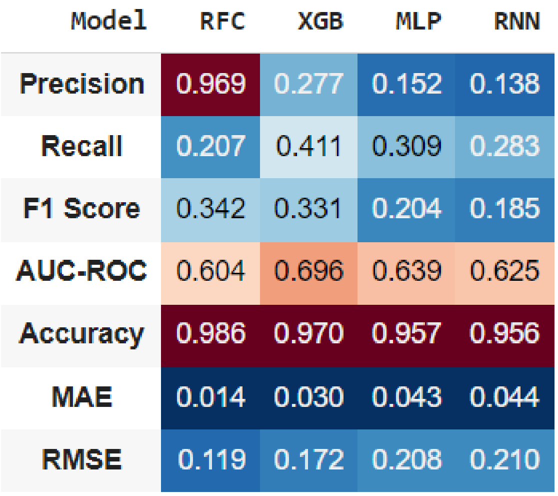


Figure 7: Comparison of Model Performance

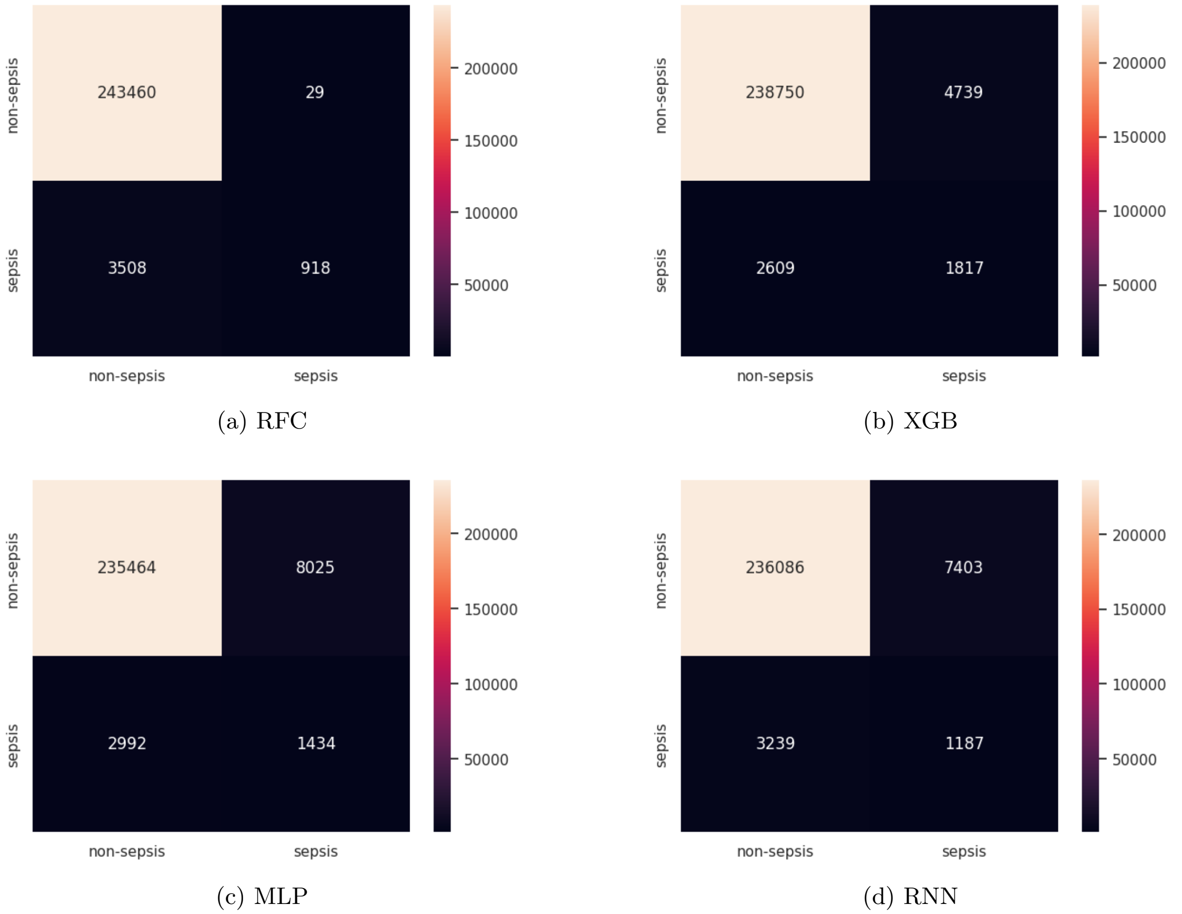
To better discern the differences in model results, Figure 8 presents a table displaying the metrics' values for each model. Overall, The Random Forest model and XGBoost demonstrate good performance compared to the others.



Model	RFC	XGB	MLP	RNN
Precision	0.969	0.277	0.152	0.138
Recall	0.207	0.411	0.309	0.283
F1 Score	0.342	0.331	0.204	0.185
AUC-ROC	0.604	0.696	0.639	0.625
Accuracy	0.986	0.970	0.957	0.956
MAE	0.014	0.030	0.043	0.044
RMSE	0.119	0.172	0.208	0.210

Figure 8: Comparison Table of Model Performance

Another crucial metric to consider is the confusion matrix (Figure 9), particularly focusing on false negatives—samples incorrectly predicted as non-sepsis when they are sepsis-positive. This misclassification can lead to delayed treatment, disease progression, and severe consequences for patients.



(a) RFC

(b) XGB

(c) MLP

(d) RNN

Figure 9: Confusion Matrix

In conclusion, the Random Forest model emerges as the optimal choice in this scenario due to its higher precision (96.9%) and F1 score (34.2%) compared to the other models.

7. Deployment

The first step in deployment is training a random forest on the entire dataset and saving it in a .pkl file. Following that, I created a simple interface using the Gradio package to ensure that the app functions correctly (you can access my full notebook project here[1]). Then, I began developing a **Flask** app using *Python*, *HTML*, and *CSS*, with **Anaconda** and **Visual Studio Code**.

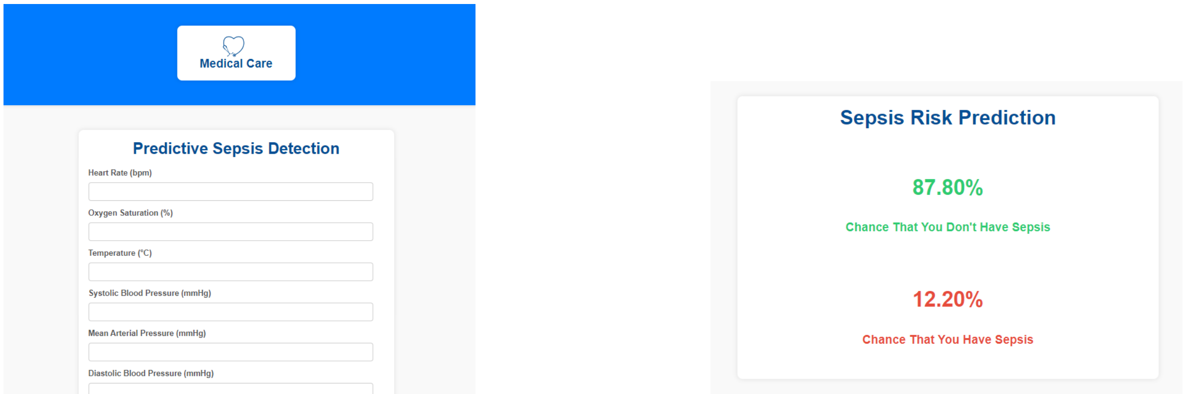
Here is what the app looks like:

Then I deploy the web app using **Docker** on **Azure Cloud**. Using Docker for deployment is helpful to avoid versioning problems and ensure consistent environments across different platforms.

This is the website[2], you can try it out. I hope you like it!

[1] https://github.com/nourelimanes/Predictive-Sepsis-Detection-Project

[2] https://medicalcare.azurewebsites.net



Medical Care

Predictive Sepsis Detection

Heart Rate (bpm)

Oxygen Saturation (%)

Temperature (°C)

Systolic Blood Pressure (mmHg)

Mean Arterial Pressure (mmHg)

Diastolic Blood Pressure (mmHg)

Sepsis Risk Prediction

87.80%

Chance That You Don't Have Sepsis

12.20%

Chance That You Have Sepsis

Figure 10: Web page interface

8. Conclusion

In conclusion, this research paper has provided a thorough exploration of the development and application of a predictive early sepsis detection system utilizing machine learning and deep learning techniques. Through extensive experimentation and analysis, we have showcased the effectiveness and potential clinical relevance of our predictive model.

Our experimental findings have yielded promising results, with our predictive model demonstrating commendable precision and F1 score specificity in identifying septic cases, despite the data's imbalance. Throughout the project journey, challenges such as handling large datasets and lengthy model implementation, particularly with hyperparameter tuning, were encountered.

It is crucial to recognize the limitations of our study. While our predictive model shows promise, further validation through large-scale clinical trials is essential to evaluate its performance in real-world settings. Continuous refinement and optimization of the model will be imperative to improve its accuracy and reliability over time. Additionally, enhancing the model's performance can be achieved by increasing the number of samples from patients with sepsis to achieve data balance.

Our research marks a significant advancement in the development of a practical and efficient tool for early sepsis detection. By harnessing the capabilities of machine learning, we have laid the groundwork for enhancing patient outcomes and healthcare delivery in the domain of sepsis management.

References

[1] M Singer, CS Deutschman, CW Seymour, and et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, 2016. https://pubmed.ncbi.nlm.nih.gov/26903338/.

[2] Shashwat Nayak and Priya Singh. Early prediction of sepsis from clinical data. Year. {Priya.Singh2, Shashwat.Nayak}@utdallas.edu.

[3] Saba Sarwar and Mahmood I. Shafi. National confidential enquiry into patient outcome and death. *Obstetrics, Gynaecology & Reproductive Medicine*, 17(9):278–279, 2007.

[4] Matthew Reyna, Chris Josef, Jeter, and et al. Early prediction of sepsis from clinical data: The physionet/computing in cardiology challenge 2019. *PhysioNet*, 2019. Published: Aug. 5, 2019. Version: 1.0.0.