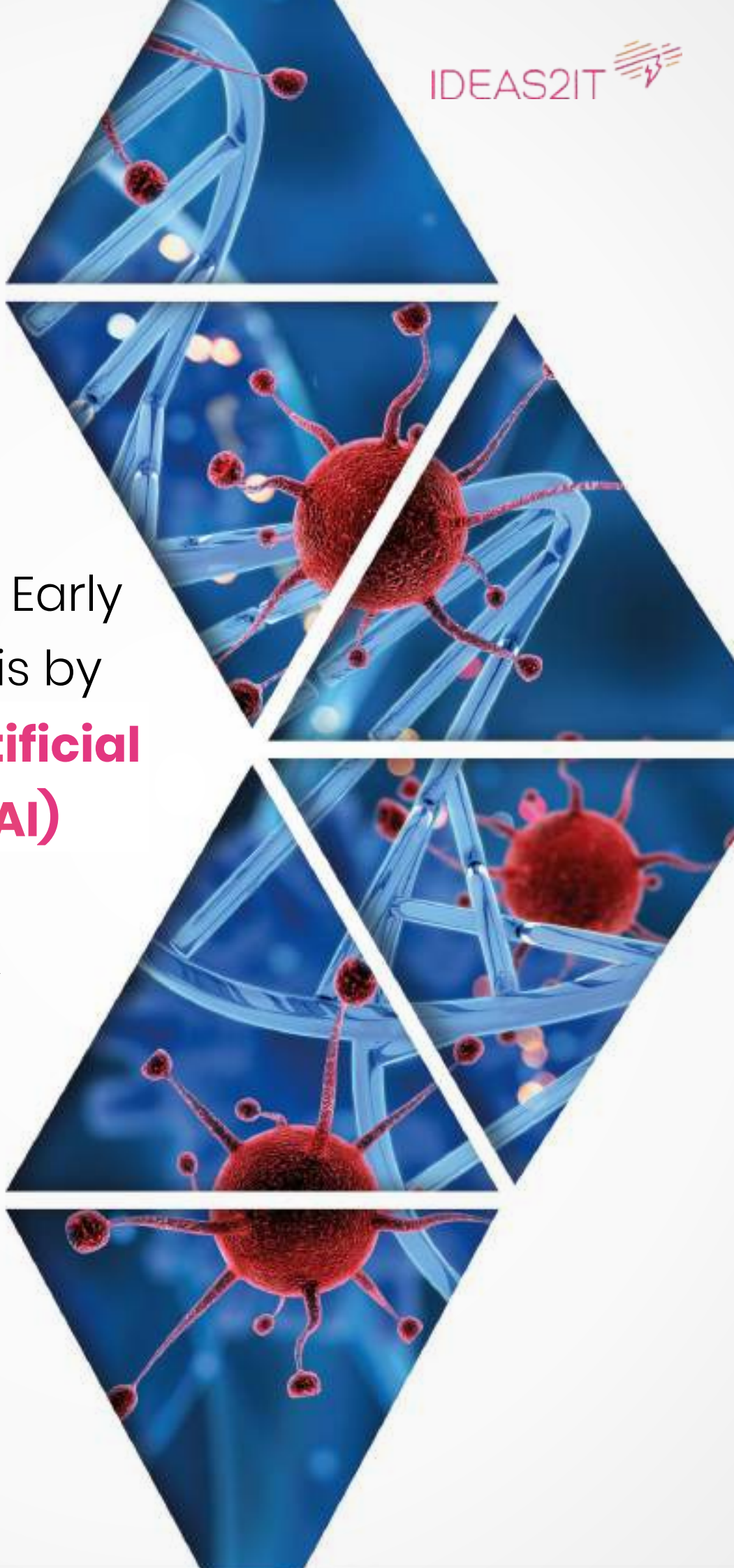
IDEAS2IT



Predicting the Early Onset of Sepsis by leveraging **Artificial Intelligence (AI)**

by
Karthik Chandramohan &
Karthik Sundararaman

Overview

Healthcare statistics reveal that Sepsis affects about 1.7 million people across the United States every year and contributes to over 250,000 deaths annually [1]. The mortality rate for patients diagnosed and hospitalized with severe Sepsis or septic shock is 25%; that's over eight times higher than the mortality rate for any other diagnosis [2]. Mortality due to Sepsis makes it one of the leading causes of preventable deaths in US hospitals.

Additionally, Sepsis accounts for more than 50% of hospital deaths, and mortality increases dramatically with greater disease severity: 10–20% for Sepsis, 20–40% for severe Sepsis, and 40–80% for septic shock [3]. Patients with Sepsis have an average Length of Stay (LOS) that is 75% longer than those hospitalized for other conditions. Patients who survive Sepsis hospitalizations are more likely to have long-term cognitive and physical function impairments [4]. Owing to the above mentioned reasons, Sepsis poses a huge economic burden on the US healthcare system, accounting for over $20 Billion in annual costs [5].

The perception that Sepsis could be handled with better care has forced hospitals to revisit their strategies and reduce the burden on the US' healthcare system. This has led to a lot of initiatives targeted at improving Sepsis detection and mitigation process. Diagnosing Sepsis is rather difficult because it shares its symptoms with several other disorders and there are no reliable biomarkers before its onset.

But AI models could help in early detection and intervention, which are the key to minimizing mortality. By leveraging patients' historical data, routine vital signs and metabolic levels from Electronic Medical Records (EMR), Machine Learning (ML) models could highlight patients who are prone to developing Sepsis. Early and accurate Sepsis onset predictions would definitely give healthcare providers to leverage more aggressive and targeted treatments.

Objectives of the Case Study

- Build a machine learning model to Identify patients with higher risk of acquiring Sepsis
- Develop a mechanism by which early Sepsis onset is identified for patients with high risk of acquiring Sepsis (objective 1) based on clinical symptoms

Data Source

Data used for the study was harvested from a competition hosted by Physionet and it contained datasets from ICUs of two hospitals. There are 40 time-dependent variables (inclusive of patients demography, vital parameters and lab tests) such as Age, Length of Stay in ICU (ICULOS), Heart Rate (HR), Pulse Oximetry (O2Sat), Temperature (Temp), Time of Admission to the Hospital (HospAdmTime), Gender, Sepsis Label, Respiratory Rate (Resp), Blood Pressure (DBP and SBP), Mean Arterial Pressure (MAP), etc. Each row of the dataset was aggregated data recorded for a period of an hour. The dependent variable, Sepsis Label, indicates the onset of Sepsis according to the Sepsis-3 definition, where 1 indicates Sepsis and 0 indicates no Sepsis. Entries of NaN (not a number) indicate that there was no recorded measurement of a variable at the time interval. A sample snapshot of the dataset is shown in Figure 1.

HR	O2Sat	Temp	...	HospAdmTime	ICULOS	SepsisLabel
NaN	NaN	NaN	...	-50	1	0
86	98	NaN	...	-50	2	0
75	NaN	NaN	...	-50	3	1
99	100	35.5	...	-50	4	1

Figure 1: Sample Snapshot of the Data

It was observed that the mean age of patients and gender ratio (Male:Female) in one dataset was 63.01±16.33 years and 58.19:41.81 respectively. Whereas in the other dataset the mean age and gender ratio(Male:Female) was found to be 60.96±16.58 years and 53.66:46.34 respectively.

Feature Selection

Every variable in the raw dataset had missing values. Missing proportionality revealed that the majority of the variables were having missing values of more than 67% as shown in the Figure 2.

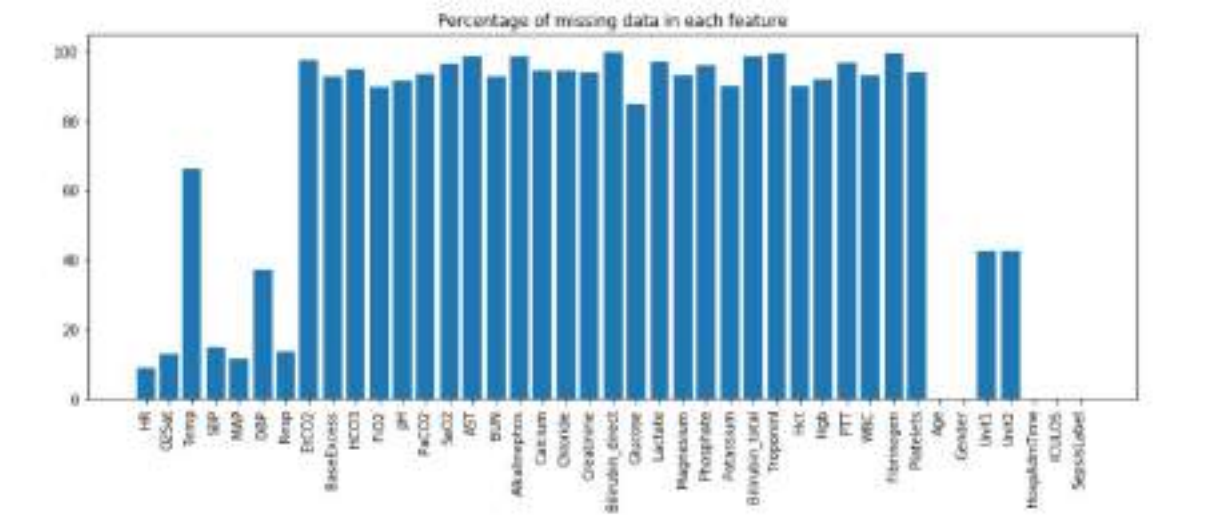


Figure 2: Missing Proportion Analysis for the Raw Dataset

Hence the current study excluded variables with more than 67% missing values. The remaining variables were grouped into continuous and categorical variables. While the missing values for continuous variables were imputed with mean, the missing values for categorical features were imputed with 999.

Subsequent to the imputation, a correlation study was performed on the remaining variables to understand how each variable was correlated to each other. Correlation Plot for the variables is shown in Figure 3.

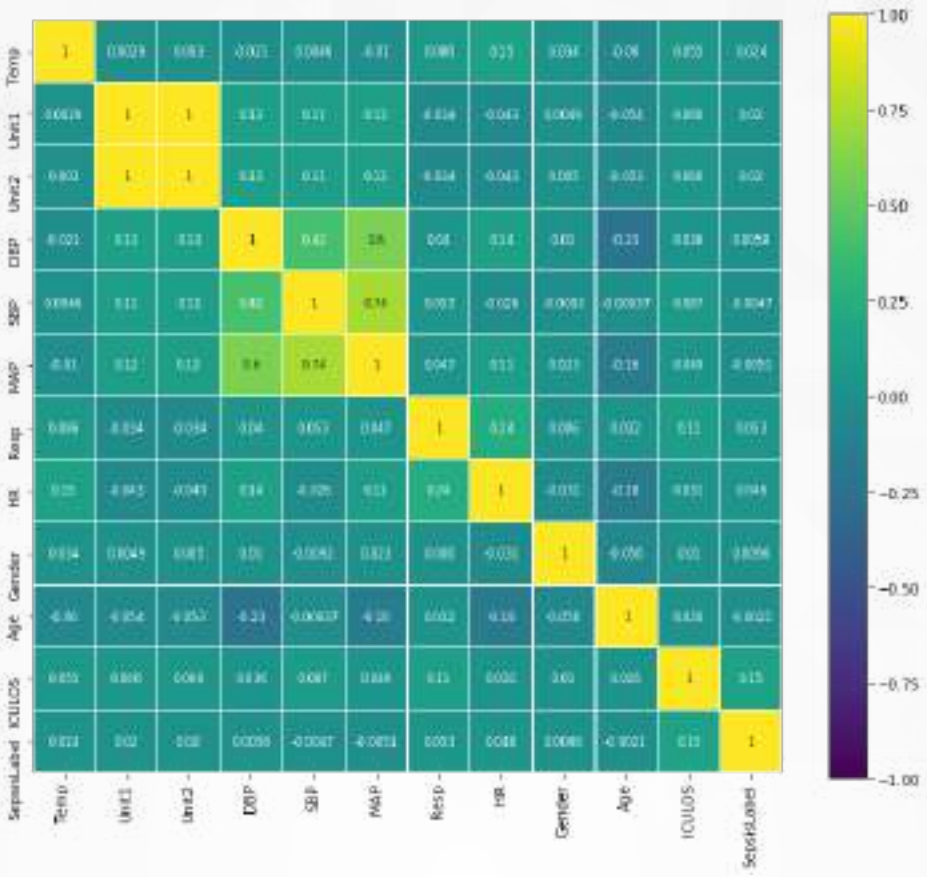


Figure 3: Correlation Plot for the Variables

It must be noted that the pressure terms (Systolic / Diastolic and Mean Arterial Pressures) would be correlated. However as these variables have different impacts on biological systems, we decided to have them as part of the study. The other highly correlated variables were Unit 1 and Unit 2, which are the MICU and SICU admission times. Therefore these variables were excluded from the study.

Data Aggregation

For each patient, measure of central tendency (mean), deviation, measure of extremes (min, max), measure of symmetry (skewness) and measure of data tails (kurtosis) were computed for all the variables except age, gender, ICULOS and SepsisLabel. So in all, 39 independent variables were used alongwith SepsisLabel for modeling.

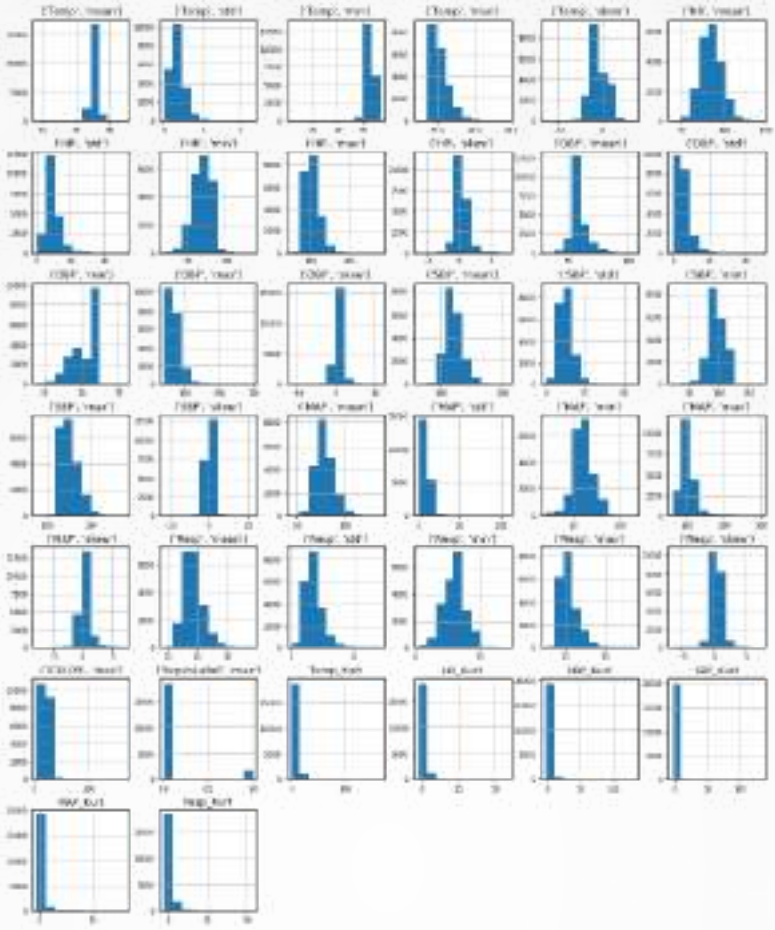


Figure 5: Distribution of the Aggregated Variables

The list of variables used for the model is also shown in the table below.

SNo	Features Used For Model – Derived Features						Feature Data Type
	Mean	Std	Min	Max	Skewness	Kurtosis	
1 - HR	✓	✓	✓	✓	✓	✓	Continuous
2 -Resp	✓	✓	✓	✓	✓	✓	Continuous
3-Temp	✓	✓	✓	✓	✓	✓	Continuous
4 -MAP	✓	✓	✓	✓	✓	✓	Continuous
5-DBP	✓	✓	✓	✓	✓	✓	Continuous
6-SPB	✓	✓	✓	✓	✓	✓	Continuous
7	ICULOS						Continuous
8	Gender						Categorical
9	Age						Continuous
10	Label						Categorical

Table 1:: List of Variables used as Input for the Model

Prevalence of Sepsis: Imbalanced Dataset

It was observed that the dataset is a highly imbalanced dataset as shown in Figure 4.

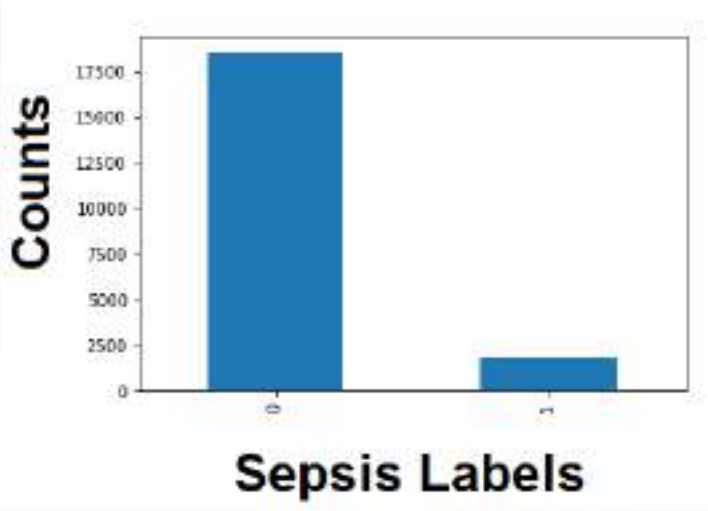


Figure 4: Prevalence of Sepsis in the Dataset

The current study addresses imbalance in the dataset using a multistage approach. During the first stage, the dataset was split into a training dataset and validation dataset (in proportions of 75:25). While the validation dataset was kept as is, the majority class of training dataset was undersampled (only 10% of the majority class was utilized) and the training dataset was converted into a balanced dataset. A powerful ensemble technique called Random Forest built with 100 trees was used to build the model on the balanced dataset. The validation dataset was used to test how well the model predicted. The other dataset was preprocessed similarly and used as a testing dataset for the model. The results of the model are shown below in Figures 5a and 5b.

	precision	recall	f1-score	support
0	0.98	0.87	0.92	4642
1	0.36	0.77	0.49	442
accuracy			0.86	5084
macro avg	0.67	0.82	0.71	5084
weighted avg	0.92	0.86	0.88	5084

Figure 5a: Validation Results of the Random Forest Model

	precision	recall	f1-score	support
0	0.98	0.85	0.91	18858
1	0.23	0.74	0.35	1142
accuracy			0.84	20000
macro avg	0.61	0.80	0.63	20000
weighted avg	0.94	0.84	0.88	20000

Figure 5b: Testing Results of Random Forest Model

Root Cause Analysis

Feature importance of the model was studied and the Figures 6a and 6b shows the importance in terms of variance exhibited by each of the variables used in the model.

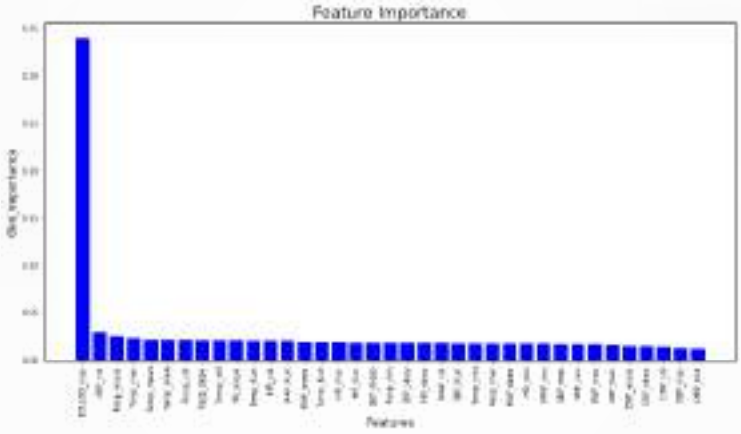


Figure 6a: Importance of the Variables Used for building the model

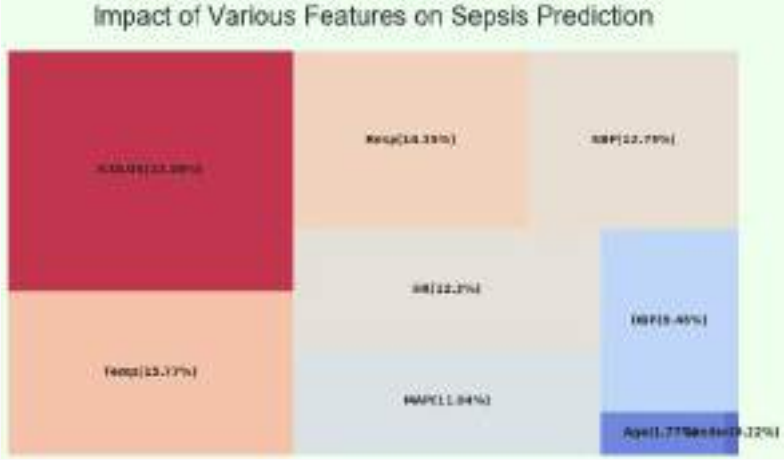


Figure 6b: Aggregated Importance of Features Used in the model

From the importance analysis we observed that the study concurred with studies in literature with regards to the Variable Length of Stay (ICULOS) and age. It has been observed that in general, Sepsis patients stay longer in a hospital than the controlled group and hence it is a good indicator for Sepsis prediction [6]. Apart from that, all other variables such as Temperature (indicator of infection), Pressure and Heart Rate (variation in cardiac functionality) have also been observed to have a significant impact on the prediction of Sepsis [7]. However this part of the study was about predicting whether the patient had Sepsis or not.

The risk of the patient is measured in terms of the probability of the model and higher the probability, higher the risk associated with acquiring Sepsis. The model uses equal probability for both the events Sepsis and non-Sepsis. However in reality, patients with a probability of 0.3 and above should be provided proper medical care as the risk could actually be on the higher side.

Early Onset Prediction of Sepsis

This part of the study deals with finding a timeline associated with the early prediction of Sepsis. Two-hours data and four-hours data prior to the onset Sepsis in true positives were identified from the model results and were extracted. The two-hours and four-hours datasets were preprocessed similarly and provided as input to the model. Risk of acquiring Sepsis was observed to be better in the two-hours data prior to the onset of Sepsis.

Limitations of the Study

There were missing values in a large number of variables, which could have altered the results. The data related to many lab results were missing due to lack of information. The data was aggregated at an hour's interval. Even though the norms pointed out to 3 and 6-hour bundles, the information recorded in real time would be in intervals of minutes.

Such data would have helped the model predict the onset time in minutes rather than an whole hour interval. Comorbidities, Medication and treatment information which are cardinal to hospitalization were found missing in the dataset used.

Benefits

- Predictions were used to identify patients who are at a higher risk of Sepsis. This could help physicians put them under a 3-hour or 6-hour surveillance bundle or even higher depending on the risk score [8].

- The readmission due to Sepsis could be avoided if proper care is provided at the initial hospitalization stage itself.

- An early prediction could prevent the onset of Sepsis or reduce its effect drastically, if critical care is timely provided to patients.

Conclusion

The model predicts the conditions of patients, under which they are more likely to acquire Sepsis. The model's performance could be improved by adding other parameters as mentioned above. The early onset prediction of Sepsis is a boon to healthcare providers and model results concur with the results of similar reported studies. The study observes that a two-hour prior data is a good indicator of onset of Sepsis.

References

- Rhee C, Jones TM, Hamad Y, et al. Prevalence, Underlying Causes, and Preventability of Sepsis-Associated Mortality in US Acute Care Hospitals. JAMA Netw Open. 2019;2(2):e18757
  https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2724768

- Dellinger, R. Phillip, et al. "Surviving Sepsis Campaign Guidelines Committee including the Pediatric Subgroup Surviving Sepsis campaign: international guidelines for management of severe Sepsis and septic shock: 2012." Crit Care Med 41.2 (2013): 580-637.
  https://www.ncbi.nlm.nih.gov/pubmed/23353941

- Martin GS. Sepsis, severe Sepsis and septic shock: Changes in incidence, pathogens and outcomes. Expert Rev Anti Infect Ther 2012; 10:701–706.
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6250243

- Iwashyna TJ, Ely EW, Smith DM, Langa KM. Long-term Cognitive Impairment and Functional Disability Among Survivors of Severe Sepsis. JAMA. 2010;304(16):1787–1794.
  https://www.ncbi.nlm.nih.gov/pubmed/20978258

- Torio CM, Andrews RM. National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2011: Statistical Brief #160. 2013 Aug. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2006 Feb-. Available from:
  https://www.ncbi.nlm.nih.gov/books/NBK169005/
  https://www.ncbi.nlm.nih.gov/pubmed/24199255

- Sakr, Y., Jaschinski, U., Wittebole, X., Szakmany, T., Lipman, J., Ñamendys-Silva, S. A., Martin-Loeches, I., Leone, M., Lupu, M. N., Vincent, J. L., & ICON Investigators (2018). Sepsis in Intensive Care Unit Patients: Worldwide Data From the Intensive Care over Nations Audit. Open forum infectious diseases, 5(12), ofy313.
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6289022/

- Jozwiak M, Monnet X, Teboul JL. Implementing Sepsis bundles. Ann. Transl. Med. 2016; 4: 332.
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5050197/

About Ideas2IT

Founded by an ex-Googler, Ideas2IT started its journey as a high-end product engineering partner for Silicon Valley startups. Ideas2IT has produced 150+ top-quality applications for 100+ clients such as Microsoft, Oracle and Opportun. Ideas2IT oers specialist capabilities in the domains of Data Science, IIoT, Blockchain, Cloud-based SaaS, Robotic Process Automation, Frontend, Backend & Fullstack Development and Intelligent Chatbots.

To know more, talk2us@ideas2it.com or visit www.ideas2it.com