Prediction of risk of Sepsis using Machine Learning on Big Data

Sifat Naseem
MS Health Informatics
fsifatna@mtu.edu

Mabel Shekinah Rose Komanduru
MS Health Informatics
mkomandu@mtu.edu

***Abstract*—Sepsis is an inflammatory and deadly disease that occurs when the body reacts to an infection. According to the Centre for Disease Control (CDC), in America, about 1.7 million adults develop sepsis and 1 in 3 people who die in a hospital had sepsis during that hospitalization. Early recognition of sepsis in patients enables prompt intervention, proactive measures, and the avoidance of serious problems. The main goal of our project was to predict the risk of sepsis in hospital-administered patients earlier than the clinical prediction using various machine learning models on big data collected from three different hospitals. Over 1 million ICU patient records are collected, including 40 different features ranging from vital signs, laboratory values, and demographics. We tried 4 different machine learning models that include Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF), and concluded that different models performed differently in terms of predicting the outcome. Overall, RF performed the best in sepsis risk prediction with an accuracy rate of 87'%', as it samples a small portion of characteristics and data points to create several decision trees which then based their prediction on the majority view of all the trees.**

I. Introduction

Sepsis being a fatal disorder that develops when the body's immune reaction to an infection runs out of control and begins to harm its own tissues and organs. It may result from several bacterial, viral, or fungi infections, such as bloodstream infections, pneumonia, and urinary tract infections. Inflammatory chemicals are released by the immune system when it becomes activated in response to an infection. This response amplifies in sepsis and can cause extensive tissue damage and inflammation. Organ failure, shock, and even death may result from this. Treatment must start right away for sepsis, which is a medical emergency. The chances of recovery are improved if sepsis is identified and treated quickly. Sepsis can quickly worsen and lead to septic shock and severe sepsis if ignored, both of which pose a serious risk to life. Sepsis is a serious illness that can be hazardous even with therapy and leaves a permanent physical impact.

According to the WHO, 4.2 million newborns and children worldwide are impacted by sepsis, which is projected to affect 30 million people worldwide and cause 6 million deaths annually. The most expensive medical condition for U.S. hospitals is sepsis, which accounts for 24 billion dollars (13'%' of U.S. healthcare spending) annually (Paoli et al., 2018). Most of these patients did not have a diagnosis of sepsis when they were admitted.

Sepsis must be diagnosed and treated as soon as possible since doing so can improve a patient's response to antibiotics, lower treatment costs, and improve patient outcomes. To do this, researchers have investigated several early sepsis detection techniques. In order to predict sepsis, one method is to employ machine learning algorithms that learn from supervised or unsupervised data input. Studies have demonstrated that when it comes to identifying the likelihood of sepsis in emergency patients, machine learning models like SVM and Random Forest outperform conventional screening technologies. This shows how machine learning could help in the early detection and treatment of sepsis.

Logistic Regression works well with binary classification problems and is computationally efficient, works well with large datasets. A probabilistic model called Naive Bayes determines the likelihood that a new observation will belong to each class and then chooses the class with the highest likelihood. The advantage of SVM is its ability to handle high-dimensional data, which is common in clinical datasets that contain numerous features. Random Forest can handle missing values, outliers, and nonlinear relationships between features, which can make it a good choice for analyzing complex clinical datasets. So, all of the above properties of these model makes them a better fit for predicting any clinical condition like Sepsis.

II. Data

The data we used for developing this project is collected from the Early Prediction of Sepsis from Clinical Data: The Physio Net/Computing in Cardiology Challenge 2019 [1]. It consists of over 60,000-1 million ICU patient records with up to 40 clinical variables for each hour of a patient's ICU stay.

The data is collected from three separate hospital systems in different parts of the United States each having its own Electronic Medical Record (EMR) system i.e., Beth Israel Deaconess Medical Center (hospital A), Emory University Hospital (hospital B), and unnamed hospital system (hospital C) [1]. The data collected contained eight vital sign variables, 26 laboratory variables, and six demographic variables which gives a total of 40 variables including 15 million data points over 2.5 million hourly time windows.

The data is de-identified and the target label is determined using Sepsis-3 clinical criteria for sepsis onset. There are

three-time points to determine the onset time t(sepsis) of sepsis for each septic patient:

- t(suspicion): The clinical suspicion of infection is identified at this step. The earlier timing of IV antibiotics and blood cultures within a specific time interval is used to determine it. If IV antibiotics were administered first, the cultures had to be obtained within 24 hours. If cultures were taken initially, IV antibiotics had to be ordered within 72 hours. In either situation, IV antibiotics must have been provided for a minimum of 72 hours.
- T(SOFA): The occurrence of organ failure as indicated by a two-point increase in the Sequential Organ Failure Assessment (SOFA) score within 24 hours.
- T(sepsis): The onset of sepsis is designated as the earlier of t(suspicion) and t(SOFA) as long as t(SOFA) happened within 24 hours of t(suspicion) and no more than 12 hours after t(suspicion). [1]

For sepsis patients, the target label is 1 if t $>$ t(sepsis) – 6 and 0 if t $<$ t(sepsis) – 6. For non-sepsis patients, the target label is 0. [1]

III. Data Preprocessing

After data extraction, we performed various data preprocessing steps so that data can be analyzed properly for better accuracy and reliability which is very important in terms of clinical data.

A. Missing Values, Duplicate data:

Our dataset was having some missing values which could have hampered the outcome. So, missing values were replaced by zeros in order to prevent the distortion of data and false assumptions. Data were checked for duplicate values and unique values were identified and analyzed.

B. Correlation Analysis:

A correlation matrix was plotted for every feature to show the correlation coefficients between variables in a dataset. It is an important tool in data analysis and is used to identify patterns and relationships between variables. It is plotted using the 'seaborn' package. It was found that there were some features that were highly correlated. Some of them are:

- MAP, diastolic, and systolic blood pressure were highly correlated. so diastolic, and systolic blood pressure were dropped and MAP was kept.
- Lactate, PTT, and phosphate have a high correlation as all are indicators of kidney health. We dropped PTT and phosphate.
- Fibrinogen and platelets have a high correlation as both define the clotting ability, so platelet was dropped.
- Bilirubin (total) and bilirubin (direct) have a high correlation as well, so the latter was dropped.
- Hct and Hgb were highly correlated, so Hct was dropped and Hgb was preserved.

Similarly, there were some other correlated features as well as shown in Fig.2 below. This image helps us to better understand the correlation as it is plotted with the help of the

Row	Measurement	Description
1	HR	Heart rate (beats/min)
2	O_2Sat	Pulse oximetry (%)
3	Temp	Temperature (°C)
4	SBP	Systolic BP (mm Hg)
5	MAP	Mean arterial pressure (mm Hg)
6	DBP	Diastolic BP (mm Hg)
7	Resp	Respiration rate (breaths/min)
8	$Etco_2$	End tidal carbon dioxide (mm Hg)
9	BaseExcess	Excess bicarbonate (mmol/L)
10	Hco_3	Bicarbonate (mmol/L)
11	Fio_2	Fraction of inspired oxygen (%)
12	pH	pH
13	$Paco_2$	Partial pressure of carbon dioxide from arterial blood (mm Hg)
14	Sao_2	Oxygen saturation from arterial blood (%)
15	AST	Aspartate transaminase (IU/L)
16	BUN	Blood urea nitrogen (mg/dL)
17	Alkalinephos	Alkaline phosphatase (IU/L)
18	Calcium	Calcium (mg/dL)
19	Chloride	Chloride (mmol/L)
20	Creatinine	Creatinine (mg/dL)
21	Bilirubin direct	Direct bilirubin (mg/dL)
22	Glucose	Serum glucose (mg/dL)
23	Lactate	Lactic acid (mg/dL)
24	Magnesium	Magnesium (mmol/dL)
25	Phosphate	Phosphate (mg/dL)
26	Potassium	Potassium (mmol/L)
27	Bilirubin total	Total bilirubin (mg/dL)
28	TroponinI	Troponin I (ng/mL)
29	Hct	Hematocrit (%)
30	Hgb	Hemoglobin (g/dL)
31	PTT	Partial thromboplastin time (s)
32	WBC	Leukocyte count (count/L)
33	Fibrinogen	Fibrinogen concentration (mg/dL)
34	Platelets	Platelet count (count/mL)
35	Age	Age (yr)
36	Gender	Female (0) or male (1)
37	Unit 1	Administrative identifier for ICU unit (medical ICU); false (0) or true (1)
38	Unit 2	Administrative identifier for ICU unit (surgical ICU); false (0) or true (1)
39	HospAdmTime	Time between hospital and ICU admission (hours since ICU admission)
40	ICULOS	ICU length of stay (hours since ICU admission)

Fig. 1. Features in the data file. [1]

heatmap feature which is a popular tool in correlation analysis for displaying the strength and direction of the association between two pairs of variables in a given dataset. The correlation coefficient between each pair of variables is depicted on the heatmap using colors.

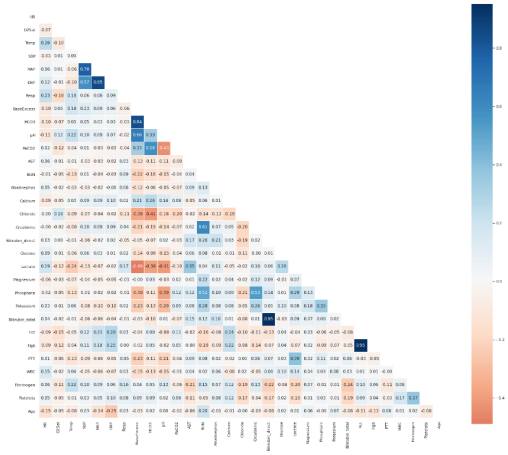

Fig. 2. Correlation Analysis.

C. Outlier Detection:

Data were checked for outliers because outliers have a substantial impact on statistical analysis, potentially leading to incorrect or misleading results. Particularly it was checked for features like heart rate and Temperature because they are two critical vital signs that are commonly used to monitor the health status of patients.

D. Data Aggregation:

On observing the data further, it was found that there were multiple records belonging to a single patient because the data is collected hourly in the ICUs. So in order to reduce the complexity of the data set, data were aggregated based on the patient ID. This helps to reduce the dimension of data from 1 million to 40,000. After performing aggregation, features have been reduced from 41 to 26.

E. Under-sampling:

There was an issue with data imbalance with respect to the target label which was handled by using the under-sampling method which is a technique where instances from the majority class are removed so that the number of instances in both classes becomes more balanced. For example, in our dataset, we have 40,000 instances, where 32,000 instances belong to class A and 7000 instances belong to class B. In this case, the dataset is highly imbalanced, and the model may struggle to predict the minority class accurately. By under-sampling the majority class, we have reduced the number of instances in class A to 15000, making the dataset more balanced. The model can then learn to distinguish between the two classes more effectively and make more accurate predictions.

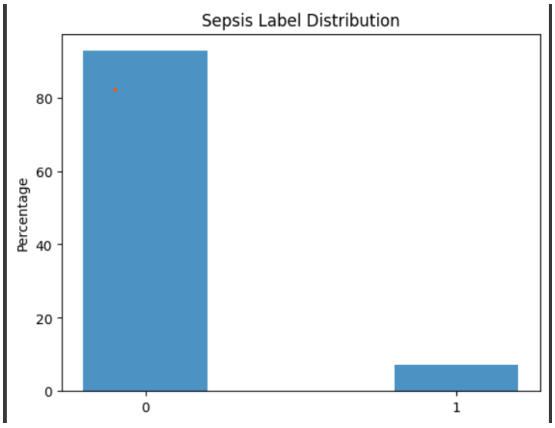

Fig. 3. Data Imbalance.

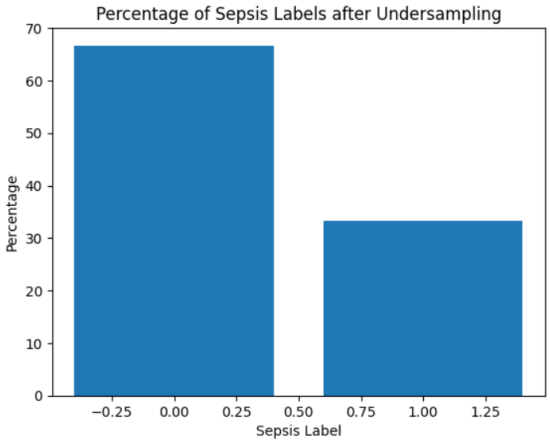

Fig. 4. Under-sampling.

IV. Methodology

As described above, four different supervised classifiers are implemented depending on our dataset. We used the 'sklearn' python package for various functions like splitting the dataset, importing the classifier libraries, hyperparameter tuning, and for accuracy/classification reports. For all the models, the dataset is divided into three parts: training, validation, and testing, with split ratios of 60-20-20. Hyperparameter tuning is done using grid search on the training data using 5-fold cross-validation for logistic regression and naïve Bayes classifiers only.

For logistic regression, hyperparameter tuning for 'the regularization parameter' ('C') is done using the values [0.01, 0.1, 1, 10, 100]. The regularization parameter C governs the degree of regularization. A lower C value results in greater regularization, whereas a higher C value results in weaker regularization and a more complex model. We can achieve a reasonable balance between overfitting and underfitting the data by adjusting the value of C. The model is then trained using the optimal hyperparameters on the training set, and its performance on the validation and testing sets is evaluated using accuracy score and classification report metrics which

gives us values of precision, recall, and f1 score.

For naïve bayes, variable smoothening is the hyperparameter that is tuned using [1e-10, 1e-9, 1e-8] values.

For SVM, the best hyperparameters are defined as 'kernel=rbf', which is abbreviated as "radial basis function." As a result, the classes will be separated by the SVM classifier using a non-linear decision boundary. The trade-off between increasing the margin and decreasing the classification error is controlled by setting the C parameter to 5. A more modest C worth will take into consideration a balance between maximizing the margin and minimizing the classification error, while a bigger C worth will prompt a more modest margin.

For random forest, we used n-estimators as 100 which specifies the number of trees in the model, with a maximum depth of 10 which helps to reduce the overfitting of the model by reducing the complexity of the model.

V. Results

Models	Val Acc	Test Acc	Precision	Recall	f1 score
LR	70	73	0.73	0.74	0.71
NB	75	76	0.76	0.76	0.74
SVM	92	67	0.65	0.68	0.55
RF	92	87	0.87	0.87	0.87

Fig. 5. Performance Analysis.

The output of the four classifiers includes a classification report which provides us with a table containing various evaluation metrics i.e., precision, recall, f1 score of sepsis, and non-sepsis prediction. The report also gives the macro-average, which is the average performance of the model over both classes, and the weighted average which considers the number of instances of each class.

For logistic regression, the accuracy was 0.736. With regards to precision, the model achieved 0.75 precision for class 0, which means that out of all samples predicted to belong to class 0, 75'%' belong to class 0. The model achieved 0.69 precision for class 1, which means that out of all samples predicted to belong to class 1, 69'%' belong to class 1. For recall, the model achieved 0.91 recall for class 0, which means that out of all the samples that belong to class 0, 91'%' were correctly predicted by the model. The model achieved 0.38 recall for class 1, which means that out of all the samples that belong to class 1, only 38'%' were correctly predicted by the model. The model achieved an f1-score of 0.82 for class 0 and 0.49 for class 1. Overall, the model achieved better performance for class 0 compared to class 1.

For naïve Bayes, the accuracy is 0.763. For the non-sepsis class, the precision, recall, and F1-score were 0.77, 0.93, and 0.84, while for the sepsis class, they were 0.75, 0.43, and 0.55, respectively. The weighted average f1-score was 0.74. This means that the overall performance of the model is not very high and could potentially be improved.

In SVM, the testing accuracy of the model is 0.665. The precision for class 0 (non-sepsis) is 0.67, which indicates that 67'%' of the samples the model predicted to belong to class 0 do. If the recall is one hundred percent, then the model correctly predicted all the samples in class zero. The harmonic mean of precision and recall, the f1-score, is 0.80, indicating a balance between precision and recall. The precision for class 1 is 0.00, indicating that of all the samples that the model predicted would belong to class 1, only 0'%' actually do. The f1-score is 0.00.

The random forest model has a satisfactory accuracy score of 0.86. However, it is evident that the model predicts instances of non-sepsis with a higher accuracy (precision of 0.85 and recall of 0.96) than it does instances of sepsis. Label 0(non-sepsis) also has a higher F1 score than label 1(sepsis). The model predicts label 0 slightly better than label 1, as shown by the macro-average of precision, recall, and F1-score. Finally, a bar plot is plotted for validation and test accuracies for a better visual understanding of the classifier performances. The plots showed that the random forest classifier performed better than all other models for both testing and validation sets with SVM being the least-performing model overall.

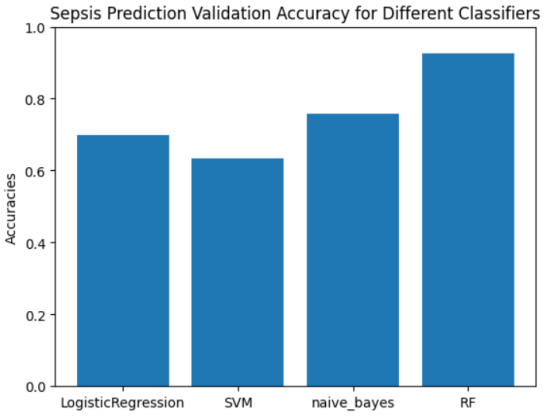

Fig. 6. Validation Accuracy.

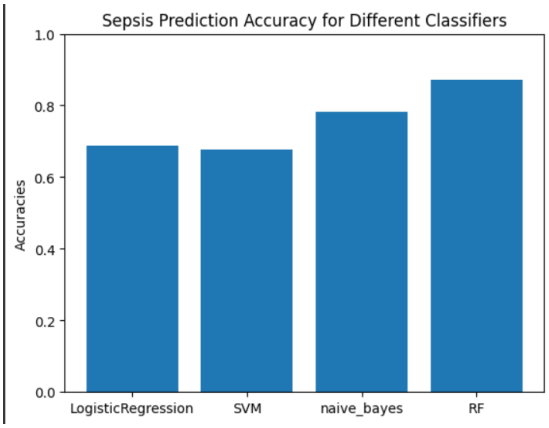

Fig. 7. Test Accuracy.

VI. Conclusion

The risk of sepsis was predicted in this project utilizing patient data from different hospitals. Various data pre-processing steps were followed to analyze the data and filter it for better prediction and suitability. Different machine Learning Models with different hyperparameters were used like Random Forest, SVM, Naïve Bayes, and logistic regression for prediction and it is concluded that Random Forest being the best-performing classifier in this project helps in predicting the risk of sepsis much more accurately with an accuracy of 87'% followed by Naïve Bayes. We believe RF's ability to handle imbalanced data, like ours, makes it the best classifier for our dataset as it samples a subset of features and data points to build multiple decision trees, thus making predictions based on the majority vote of all the trees. The performance of each model was compared based on various scores like F1 score, precision, and Recall, and the best performing mode was chosen.

In order to identify patients who are at a high risk of developing sepsis and to take preventive steps to do so, healthcare providers can benefit from these findings. Data imbalance was the biggest challenge in this project which was somehow handled by under-sampling but not completely. While it has been demonstrated that machine learning models may accurately predict a patient's risk of sepsis, deep learning models may have even more potential to boost the precision and accuracy of these predictions provided sufficient computation resources should be available.

There is therefore a lot of opportunity for deep learning models to be used in sepsis risk prediction studies in the future. We might be able to create even more potent tools for forecasting the danger of sepsis and enhancing patient outcomes by integrating the benefits of both machine learning and deep learning models.

VII. Contributions

- Sifat Naseem- Data loading, Data preprocessing.
- Mabel Komanduru- Machine learning classifiers, Performance evaluation.

References

[1] Reyna, Matthew A. PhD1; Josef, Christopher S. MD1; Jeter, Russell PhD1; Shashikumar, Supreeth P. B.Tech2,3; Westover, M. Brandon MD, PhD4; Nemati, Shamim PhD1,3; Clifford, Gari D. DPhil1,2; Sharma, Ashish PhD1. Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019. Critical Care Medicine 48(2):p 210-217, February 2020. — DOI: 10.1097/CCM.0000000000004145

[2] Barghi B, Azadeh-Fard N. Predicting risk of sepsis, comparison between machine learning methods: a case study of a Virginia hospital. Eur J Med Res. 2022 Oct 28;27(1):213. doi: 10.1186/s40001-022-00843-4. PMID: 36307887; PMCID: PMC9617383.