**Comparative study on early prediction of Sepsis**

**Abstract:**

Sepsis is a major concern in a health care industry with high mortality rate. Therefore, it necessary for us to predict it early for intervening with treatment which would decrease the mortality rate. An appropriate model with a good prediction is necessary as the false predictions would increase the consumption of limited hospital resources. The model with the best prediction is to be selected by conducting various experiments through machine learning models. Naïve Bayes, Ada Boost, Random Forest and Resnet machine learning models are used for the classification of Sepsis data. Imbalanced sepsis data set and balanced dataset using SMOTE are used for the analysis. The Labels are shifted forward to predict the onset of sepsis early. The investigation is then run on the dataset with shifted labels. The best classifier for each of the shifted dataset is to be selected with the performance evaluation metrics such as Accuracy, F1- Score, sensitivity, and specificity. Random forest classifier with over sampling technique outperforms the other classifiers with the high accuracy score of 95%, low sensitivity value of 44%, high specificity value 99% and a high F1 score of 70%.

1. **Introduction**

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection [1]. The mortality rate with the incidence of Sepsis is not reduced and remains high [2]. Treatment for sepsis is major economic concern of the health care industry [3]. The most sepsis patients are treated in Intensive Care Unit. The cost associated with the treatment of sepsis depends on several factors such as age, Intensive unit considered. The patient incident with sepsis who has not been treat with antibiotics six hours from hospital admit has 9% high mortality rate than patients treated with sepsis [4]. Therefore, it is necessary for us intervene early and start with the antibiotic treatment. The onset of sepsis is calculated using the Systematic Inflammation Response syndrome (SIRS) criteria and Sequential Organ Failure assessment (SOFA) score. While it is necessary for the early prediction of sepsis, false alarms in the predictions would consume the limited hospital utilities. Delayed treatment could inflate the mortality rate therefore a model which has a high positive rate and low negative rate. Accurate system for the early detection of sepsis remains unresolved. Also, the huge imbalance in the dataset in the health is a major concern as it affects the accuracy making prediction biased towards the majority class. The accuracy of conventional model classifiers is low on the imbalanced dataset [5].

The paper is aimed at early detection of sepsis. A comparative study with four different models such as Naïve Bayes, Random Forest, Ada Boost and Resnet are implemented on four different time shifted dataset. The classifiers are also trained and classified on both the balanced and imbalanced dataset. The appropriate classifier is to be selected based on the performance metrics such as Accuracy, F1score, sensitivity and specificity.

Section 2 illustrates the related work in this study and Section 3 illustrates the data collection, data preprocessing and methodologies implemented. Experimental results are presented in Section 4 and conclusion in section 5.

1. **Related Work**

To optimize the available treatment and to reduce the high mortality caused by sepsis it is necessary for us to detect it early. It should be detected as early as possible. Various models have been proposed by researchers for the detection of Sepsis on the different dataset available. Hidden Markov Model, deep learning models such as LSTM, Naïve Bayes, Artificial Neural Network, Random Forest, Logistic Regression are some of the popular model used in the detection of sepsis.

An approach to predict sepsis based on the hypothesis that instable patients prone to Sepsis is presented in [6]. Different machine learning such as logistic regression, Support vector machine and artificial neural networks were used to predict sepsis based on SIRS criteria. Area Under Curve (AUC) was used to evaluate the models. A hybrid approach that combines convolutional neural network and Long shot term memory network was proposed in [7] for timely detection of sepsis. The AUROC show the model performs well the early detection of sepsis. [8] shows the application of machine learning models for the prediction of sepsis on a pediatric dataset. AUROC along with sensitivity and specificity were used an evaluation metric.

The Convolutional neural network and recurrent neural network were applied to physiological to detect sepsis six to twelve hours early is showed in [9]. The performance of the model is evaluated with AUROC. The framework to continuously monitor the mortality risk with the help of Hidden Markov Model is proposed in [10]. The Sepsis data was used a case study for the analysis of proposed framework. Various machine learning models such as decision trees, Random forest, Naïve Byes, Support Vector Machine (SVM) were analyzed to compare the performance. Accuracy and AUROC were the evaluation metric used for the comparison of models. [11] showed the necessity to balance the imbalance dataset in the healthcare data. The comparison of the performance metric showed the improvement in F1 score and sensitivity values.

From the above literature it is evident that popular machine learning model along with deep learning models are used in the early detection of sepsis. The imbalance in the dataset is handled by very few studies. The importance is also given to class imbalance in this article. Synthetic Minority oversampling technique (SMOTE) is applied to balance the dataset through oversampling. The standard classifier such Random Forest, Naïve Bayes, Ada Boost along with Resnet deep learning method is used to predict the onset of Sepsis. The new contribution for this paper includes the shift in the label to predict one hour, six hour and twelve hours ahead of the occurrence of sepsis. The comparative study is proposed on the four different classifier, four different label shifts and two different datasets. On all datasets Random Forest classifier with the over sampling technique performed better than other classifiers with the good accuracy, F1 score, specificity, and a low sensitivity.

1. **Methodology**

The binary classifiers to classify the sepsis data such as Naïve Bayes, Random Forest, Ada Boost and Resnet model are implemented with Python, Keras and TensorFlow libraries. The pipeline for the model is data collection, data preprocessing, training, and testing the data, evaluation of metrics. The imbalance in the data set is balanced using the SMOTE technique. Labels are shifted for the early detection of sepsis. The models are implemented on both the balanced and imbalanced datasets on all four shifts in the sepsis label. Totally 24 models using 2 datasets, 4 label shifts and 4 classifiers are implemented.

* 1. **Dataset Description**

The physiological data used for early prediction of Sepsis is extracted from the PhysioNet Computing in Cardiology Challenge 2019[12-18]. Data is collected from three different Electronic Medical Record (EMR) systems from three different hospitals across U.S. These information were accumulated over the past decade with approval from the precise institutional review boards. The deidentified patient record was not labeled for the onset of sepsis. The dataset comprises of 40336 patients’ records grouped into two different training sets. Every patient data is represented in a pipe separated value (PSV) file with header denoting the variables considered for the analysis. Each row in a PSV file represents one hour of patient data. Each patient data has forty different time subordinate components comprising of eight vital signs, twenty-five laboratory observations and six demographic information variables. The vital signs include Systolic blood pressure (SBP), Diastolic blood pressure (DBP), Mean arterial pressure (MAP), Respiratory rate etc. Calcium, Chloride, Creatinine, Bilirubin\_direct, potassium and Magnesium are the few laboratory values observed. Age, gender, type of ICU, Hospital admit time and ICULOS are the demographic information considered. Hospital Admit time indicates the hours between the admit time and patient being moved to ICU. ICULOS variables denotes the length of patient stay in ICU**.** The values in the row represent the corresponding vital sign or laboratory values observed at that point of time. Every row is marked for sepsis label as 0 or 1 based on the suspicion variable and SOFA (Sequential Organ Failure Assessment) score. Label 1 denotes the patients affected by Sepsis and 0 for non-sepsis patients. The variables whose measurement was not recorded at the observed time was specified by Nan (Not a number) values and these variables are to be handled during the preprocessing. The variables observed differ in their unit of measure and their value differs in the range. Longitudinal data representing the length of stay would vary for each patient. This causes the imbalance in the dataset where most observation stay in non-sepsis class and a small minority class representing sepsis.

* 1. **Data preprocessing**

Raw data cannot be given as an input to the model and therefore it must be processed and transformed. Preprocessing is the vital step in machine learning because it influences the accuracy of the results. Data cleaning is especially important in case of health care data as there are lots of missing data since vital signs considered are recorded on an hourly basis, but laboratory values are recorded daily. Each patient file is appended with column indicating a sequential patient id. The separate patient files are then grouped into a single comma separated value file. Huge missing values in the dataset would create a bias in the prediction. Therefore, the following steps are to be followed to preprocess the data

* Missing Values Ratio dimensionality reduction method is used to filter out columns with more than 95% of the missing value
* Patients are then grouped by the patient id
* The missing value of the laboratory variables are then imputed with values carry forwarded from the previous observation of the patient as they are not observed at all point of time unlike the vital signs.
* Data set is standardized using a standard scaler since the variables vary in the unit measure and variables have high range difference. The mean of the dataset would be zero and the standard deviation should remain the same after rescaling the data.
* The dataset is then shifted by the one value greater than minimum value in the dataset. The remaining missing values are imputed with the constant value of 0 to handle the remaining missing values in the dataset.

Hence, the resultant dataset obtained cleaned and transformed to fit to the model. It is necessary to follow the preprocessing in the same order as any change in them would result in most missing values imputed to zero affecting the accuracy of the model.

The entire dataset could not be used to train the data. Hence it is divided into training and test dataset. The training data is imbalanced with number of observations belonging to each class not same. The imbalance ratio of sepsis to non-sepsis label is 9:1. This imbalance in the data have a huge impact on learning process of the model since the model would tend to classify most of the data to its majority normal class. The accuracy of the model remains high for the imbalanced data which would mislead to the conclusion that the model performed better but it is not. In such cases F1 score, Sensitivity and specificity are could be used to evaluate the performance of the models. The relationship between f1 score and specificity and sensitivity are moving.

* 1. **Models**

The binary classification problem of predicting sepsis at time t is resolved by using machine learning classification models like Naïve Bayes, Ada Boost, Random Forest and Resnet algorithm. Input to these models is a feature vector containing the vital signs, laboratory observations and demographic information. The output of the model is the sepsis label indicating the occurrence of sepsis at that time. Entire available data cannot be given as input to the classifier to train the model since the predicted results must be evaluated. Evaluation of the predicted result require some data that was not used to train the data. Hence, the preprocessed dataset for each experiment is further divided into train set containing 70% of records in the dataset and test data containing 30% of remaining records in the dataset. The classifier is trained with the training data set and the predicted results are evaluated with test data. Accuracy, F1-score, sensitivity, and specificity are calculated for each experiment conducted as an evaluation of performance metric of the classifier. These experiments were aimed at selecting an algorithm with a balanced sensitivity and specificity values as a performance metric for the model predicting sepsis.

The random sample of 2000 patients are initially used to train and test the classifiers. The classifiers are then applied to data set comprising of 10000 patients. The patient files for experimental setup are chosen randomly from the entire dataset to eliminate the bias in the setup. All classifiers are then applied to the three different datasets. There is a disproportion in the ratio of the positive classes (Non-Sepsis) to negative class (Sepsis) in the training datasets. Synthetic Minority Oversampling Technique (SMOTE) was used to proportionate them. Classifiers were also applied to the balanced data set. The labels were then time shifted to predict one hour, six hour and 12 hours ahead of the actual occurrence of Sepsis. The models were also executed on the shifted labels.

Naïve Bayes is a probabilistic machine learning model based on Bayes theorem. Naïve Bayes model assumes that all the features are independent and have equal effect in the prediction of sepsis. This model performs well for binary classification tasks especially on the medical data for its transparency in predicting the outcomes and justifications for the prediction. It is commonly used as a benchmark to compare with other machine learning models.

Random forest is a popular decision tree-based prediction model for both classification and regression. To avoid overfitting the data the random forest models take the average of the results obtained from its individual estimators. This makes random forest to outperform the decision tree. The random forest model performs better even where there is a huge missing value in the dataset. It is best suited for our application since there is a moderate number of missing values in the dataset.

One other commonly used machine learning algorithm is the Adaptive boosting (Ada Boost) algorithm. The Ada boost classifier combines multiple week learner into a single strong learner. At each iteration, the weights of incorrect classification are reassigned or recalculated until either maximum number of iterations is reached or there no more error in the classification algorithm. Since the incorrect classifier has a huge impact in our prediction, Ada boost algorithm would be a good choice. Also, the model resist overfitting the data.

Resnet is a deep learning model based on neural network. There are many variants for the Resnet model such as Restnet-50, Resnet -101, Resnet -152 etc. Our model is experimented with Resnet50 architecture. The skip function in Resnet is used to transfer the result from a previous layer to the next layer. Deep learning model works well when the data set is large, and it is much suited for complex problems.

There is a problem when the predictor class is skewed toward the negative class (Sepsis). The classifier would predict all the results towards the positive Non sepsis class. Hence it is necessary to balance the dataset. To balance the dataset, we could either under sample the majority normal class or oversample the minority sepsis class. The under sampling the majority class would result in the loss of information. So, the moderate imbalance in the dataset is handle using Synthetic Minority Oversampling Technique (SMOTE). The minority class (Sepsis) is oversampled by replicating them such that no new information or variation adding to the data. Both classes are balanced after the application of SMOTE. The following figure present the skewness in the dataset for various time shifts in the dataset.

**Figure1: Class Distribution for various time shifts**

The goal of the system is to detect the sepsis before the actual occurrence as much as possible. The label is shifted forward to achieve the early prediction. The labels are shifted one row to predict sepsis one hour ahead of the actual prediction. Similarly, it is shifted six rows and twelve rows forward to predict six hours, twelve hours ahead of the actual sepsis occurrence. The following figure depicts the label shift with respect to time. The graph is plotted for time along the X axes and sepsis label plot along the Y axes.

**Figure 2 Label shift graph with respect to time**

We train the different machine learning models and a model with better prediction is to be chosen. Since there is an imbalance in the data accuracy must not be the only metric evaluating the performance of the model as it would be high and mislead the selection of the model. Hence sensitivity, specificity and f1 score must be considered along with the accuracy to assess the different models implemented.

1. **Model Performance**

Initially the classifiers are executed for the imbalanced datasets of different sizes. Then the balanced dataset after using smote is trained by all the classifiers. Labels are then shifted for one hour, six hour and 12 hours. Finally, each shifted data is then trained with classifier both before and after application of SMOTE. All 48 (4 \* 4\* 2 = 32) models computed by 4 classifier, 4 label shift and 2 different data sizes were classified to predict sepsis occurrence at that time. The following tables depicts the comparative performance of the models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Label Shift** | **Model** | **SMOTE** | **Accuracy** | **Specificity** | **Sensitivity** | **F1 score** |
| **No Shift** | Naïve Bayes | No | 80.44% | 0.99 | 0.94 | 0.07 |
| Random Forest | No | 99.15% | 0.99 | 0.09 | 0.69 |
| Ada Boost | No | 98.24% | 0.99 | 0.01 | 0.10 |
| Resnet | No | 97.31% | 1.0 | 0.0 | 0.09 |
| Naïve Bayes | Yes | 32.19% | 0.95 | 0.99 | 0.04 |
| Random Forest | Yes | 99.19% | 0.98 | 0.44 | 0.76 |
| Ada Boost | Yes | 85.32% | 0.99 | 0.93 | 0.09 |
| Resnet | Yes | 98.00% | 0.99 | 0.17 | 0.29 |
| **1 Hour** | Naïve Bayes | No | 80.57% | 0.99 | 0.94 | 0.07 |
| Random Forest | No | 99.13% | 0.98 | 0.10 | 0.71 |
| Ada Boost | No | 98.11% | 0.99 | 0.01 | 0.01 |
| Resnet | No | 98.12% | 1.0 | 0.0 | 0.99 |
| Naïve Bayes | Yes | 40.43% | 0.96 | 0.99 | 0.07 |
| Random Forest | Yes | 99.18% | 0.98 | 0.47 | 0.77 |
| Ada Boost | Yes | 84.35% | 0.99 | 0.93 | 0.09 |
| Resnet | Yes | 22.22% | 0.99 | 0.93 | 0.03 |
| **6 Hour** | Naïve Bayes | No | 75.34% | 0.98 | 0.94 | 0.10 |
| Random Forest | No | 99.10% | 0.98 | 0.11 | 0.80 |
| Ada Boost | No | 97.35% | 0.99 | 0.01 | 0.01 |
| Resnet | No | 95.00% | 0.99 | 0.01 | 0.01 |
| Naïve Bayes | Yes | 34.51% | 0.93 | 0.99 | 0.06 |
| Random Forest | Yes | 99.14% | 0.97 | 0.47 | 0.83 |
| Ada Boost | Yes | 83.63% | 0.98 | 0.91 | 0.12 |
| Resnet | Yes | 97.11% | 0.99 | 0.01 | 0.03 |
| **12 Hour** | Naïve Bayes | No | 61.75% | 0.96 | 0.97 | 0.10 |
| Random Forest | No | 99.29% | 0.97 | 0.12 | 0.87 |
| Ada Boost | No | 96.86% | 0.99 | 0.01 | 0.02 |
| Resnet | No | 94.35% | 0.99 | 0.01 | 0.03 |
| Naïve Bayes | Yes | 61.75% | 0.96 | 0.97 | 0.10 |
| Random Forest | Yes | 99.38% | 0.97 | 0.44 | 0.90 |
| Ada Boost | Yes | 96.86% | 0.99 | 0.01 | 0.02 |
| Resnet | Yes | 96.87% | 0.99 | 0.01 | 0.01 |

From the above result it is evident that all classifiers perform poorer with low F1 score, low specificity, and high sensitivity for all the imbalance datasets except for random forest model. Random Forest model performs better than other classifiers for both the balanced and imbalanced dataset. The model also produces best results for the all the shifted labels and is its well suited for our analysis. The model has a high specificity (true positive rate) and low sensitivity (true negative rate) along with high F1score. Feature importance obtained from the results of random forest is used to rank the features given as input to the model. Hospital admit time ranks first in all the analysis conducted and plays a predominant role in our prediction. Temperate, Age, calcium, WBC, and BUN are other highly ranked features. When the hospital admit time is dropped from the analysis and the models are re-executed. There is a slight decrease in the accuracy and F1score for the recomputed models. Random forest model accurately predicts sepsis exact occurrence, one hour, six hour and twelve hours ahead of the actual occurrence of sepsis. We therefore select random forest classifier along with oversampling technique as a classifier for early detection of Sepsis

1. **Conclusion**

The dataset containing physiological data on sepsis is used to compare the performance of different machine learning classifiers such as Naïve Bayes, Random Forest, Ada Boost and Resnet are compared this paper. This paper also intensively investigates the performance for prediction of sepsis on time shifted datasets for the early detection of sepsis. The paper also investigates the problem of the imbalanced dataset using the SMOTE technique.

Among the different classifiers investigated Random Forest performance better than the other models both on balanced and imbalanced data. Random Forest also outperforms other models in early prediction of sepsis for all the time shifted datasets. Random forest with over sampling technique has better accuracy with high specificity predicting the true positive and a low sensitivity depicting the false negative rates. Also, the models have high F1 score with balances sensitivity and specificity rate. Hospital admit time plays a crucial role in the detection of Sepsis.

**References:**

1. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). J Am Med Inform Assoc 2016; 315: 801–810.
2. Ferrer R, Artigas A, Levy MM, Blanco J, Gonzalez-Diaz G, Garnacho-Montero J, Ibanez J, Palencia E, Quintana M, De la Torre-Prados MV, et al. Improvement in process of care and outcome after a multicenter severe sepsis educational program in Spain. JAMA. 2008;299(19):2294–303.
3. Chalupka AN, Talmor D. The economics of sepsis. Crit Care Clin. 2012;28(1):57–76. vi
4. The Vincent X. Liu, Vikram Fielding-Singh, John D. Greene, Jennifer M. Baker, Theodore J. Iwashyna, Jay Bhattacharya, Gabriel J. Escobar, “The Timing of Early Antibiotics and Hospital Mortality in Sepsis” https://doi.org/10.1164/rccm.201609-1848OC
5. G. M. Weiss, “Mining with rarity: a unifying framework,” ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 7–19, 2004.
6. Karjalainen, Pasi A Bloch, Eli Rotem, Tammy Cohen, Jonathan Singer, Pierre, Aperstein, Yehudit “Machine Learning Models for Analysis of Vital Signs Dynamics: A Case for Sepsis Onset Prediction”, Journal of Healthcare Engineering <https://doi.org/10.1155/2019/5930379>
7. Simon Meyer Lauritsen, Mads Ellersgaard Kalør, Emil Lund Kongsgaard, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, Bo Thiesson, “Early detection of sepsis utilizing deep learning on electronic health record event sequences”. <https://doi.org/10.1016/j.artmed.2020.101820>.
8. Le S, Hoffman J, Barton C, Fitzgerald JC, Allen A, Pellegrini E, Calvert J, Das R. Pediatric Severe Sepsis Prediction Using Machine Learning. Front Pediatr. 2019 Oct 11;7:413. doi: 10.3389/fped.2019.00413. PMID: 31681711; PMCID: PMC6798083.
9. Li Xin, Ng G, Schlindwein Fernando, A Convolutional Neural Network for Early Detection of Sepsis using Hourly Physiological Data from Patients in Intensive Care Unit

<https://www.researchgate.net/publication/335842117_A_Convolutional_Neural_Network_for_Early_Detection_of_Sepsis_using_Hourly_Physiological_Data_from_Patients_in_Intensive_Care_Unit>

1. Gupta A, Liu T, Crick C. Utilizing time series data embedded in electronic health records to develop continuous mortality risk prediction models using hidden Markov models: A sepsis case study. Statistical Methods in Medical Research. 2020;29(11):3409-3423. doi:10.1177/0962280220929045
2. Yang Zhao, Zoie Shui-Yee Wong, Kwok Leung Tsui “A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events’ Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection”
3. MLA Reyna, Matthew, et al. "Early Prediction of Sepsis from Clinical Data -- the PhysioNet Computing in Cardiology Challenge 2019" (version 1.0.0). PhysioNet (2019), https://doi.org/10.13026/v64v-d857.
4. APA Reyna, M., Josef, C., Jeter, R., Shashikumar, S., Moody, B., Westover, M. B., Sharma, A., Nemati, S., & Clifford, G. (2019). Early Prediction of Sepsis from Clinical Data -- the PhysioNet Computing in Cardiology Challenge 2019 (version 1.0.0). PhysioNet. https://doi.org/10.13026/v64v-d857.
5. Chicago Reyna, Matthew, Josef, Chris, Jeter, Russell, Shashikumar, Supreeth, Moody, Benjamin, Westover, M. Brandon, Sharma, Ashish, Nemati, Shamim, and Gari Clifford. "Early Prediction of Sepsis from Clinical Data -- the PhysioNet Computing in Cardiology Challenge 2019" (version 1.0.0). PhysioNet (2019). https://doi.org/10.13026/v64v-d857.
6. Harvard Reyna, M., Josef, C., Jeter, R., Shashikumar, S., Moody, B., Westover, M. B., Sharma, A., Nemati, S., and Clifford, G. (2019) 'Early Prediction of Sepsis from Clinical Data -- the PhysioNet Computing in Cardiology Challenge 2019' (version 1.0.0), PhysioNet. Available at: https://doi.org/10.13026/v64v-d857.
7. Vancouver Reyna M, Josef C, Jeter R, Shashikumar S, Moody B, Westover M B, Sharma A, Nemati S, Clifford G. Early Prediction of Sepsis from Clinical Data -- the PhysioNet Computing in Cardiology Challenge 2019 (version 1.0.0). PhysioNet. 2019. Available from: <https://doi.org/10.13026/v64v-d857>.
8. Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, Clifford GD, Sharma A. Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge. Critical Care Medicine 48 2: 210-217 (2019). <https://doi.org/10.1097/CCM.0000000000004145>
9. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.