2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/Confluence51648.2021.9377090

Early Prediction of Sepsis using Machine Learning

Anuraag Shankar*, Mufaddal Diwan†, Snigdha Singh‡,
Husain Nahrpurawala§ and Tanusri Bhowmick¶
Department of Computer Engineering, Pune Institute of Computer Technology
Pune, India
Email: *anuraag.k.shankar@gmail.com, †mufaddal.d.786@gmail.com, ‡snigdha920@gmail.com,
§hussainburhanuddin1@gmail.com, ¶tanusribhowmick0@gmail.com

***Abstract*—Sepsis is a fatal disease caused by infection. It has a significantly high mortality rate, particularly for patients in the ICU. The early and accurate detection of Sepsis is crucial as delayed treatment causes a sharp increase in the mortality rate. The proposed research aims to develop a classifier that accurately predicts Sepsis up to six hours before the clinical diagnosis of the disease. This is achieved using the patient's EMR, vital signs and demographics. The research shows several imputation techniques and proposes a new filling algorithm known as Mixed Filling. The main features contributing to the classifier's predictions have been described, thereby making the model more interpretable for medical personnel. Six models namely Random Forest, Logistic Regression, Light Gradient Boosting Machine, eXtreme Gradient Boosting, Neural Network and Long Short-Term Memory have been investigated for the classification of patients. The evaluation metrics that have been obtained are unprecedented and can be extremely useful for the timely and accurate prediction of Sepsis.**

***Index Terms*—Classification, Imputation, PhysioNet, Prediction, Sepsis**

I. Introduction

Sepsis is a fatal disease caused by an infection in the body. The Sepsis-3 definition describes the disease as a "life-threatening organ dysfunction caused by a dysregulated host response to infection" [1]. Managing Sepsis is crucial and challenging as it is a cause of high mortality, especially in the intensive care unit (ICU). It is responsible for placing an enormous burden on health care systems. There are approximately 48.9 million cases and 11 million deaths due to Sepsis worldwide. This is about one-fifth of the global death count [2]. It is seen that there is a significant rise in the mortality rate of the patient if there is a delay in receiving the antibiotics. Delayed treatment is associated with an approximate increase of 4-8% in mortality rate per hour [3]. The heterogeneous nature of possible infections and diversity of host responses makes it difficult for medical experts to recognize and diagnose the disease.

According to the Sepsis-3 definition [1], the organ dysfunction is quantified using the Sequential Organ Failure Assessment (SOFA) score. The score consists of six parameters, one each for the respiratory system, nervous system, cardiovascular system, liver, coagulation and kidneys. Each parameter is assigned a value from 0 to 4, zero representing the lowest chance of organ failure. An increase of at least 2 points in the SOFA score corresponds to an in-hospital mortality of greater than 10%.

Earlier researches use the MIMIC-III dataset [4] for the early prediction of Sepsis. The authors in [5] use this dataset and propose a method "Insight" for the classification of Sepsis patients. Recent studies use the PhysioNet Sepsis Challenge [6] dataset, which has been labelled to predict Sepsis in patients up to six hours before the SOFA score prediction. In [7], the authors use an ensemble machine learning model consisting of forest-based models to detect Sepsis in patients. Another research [8], shows the comparison of various LSTM (Long Short-Term Memory) based models which try to capture the interactions in the time series dataset [6]. The studies have been explained in detail in Section II.

This paper compares and analyzes six different models for various imputation algorithms. A new imputation algorithm "Mixed Filling" has been proposed to further enhance the performance of the models. The models include classical machine learning algorithms and user-defined deep learning models for the early prediction of Sepsis. The following methods of filling missing values in the dataset [6] have been assessed for every model:

1) Front Filling
2) Back Filling
3) Linear Filling
4) Mixed Filling

Several metrics have been calculated for the models. A detailed comparison of all models for each imputation technique has been shown in Section IV.

II. Literature Review

Two datasets have been widely used for the classification of Sepsis patients - MIMIC-III [4] and PhysioNet Sepsis Challenge [6]. The MIMIC-III is a time-series dataset that contains data related to patients who stayed in the ICU at the BIDMC (Beth Israel Deaconess Medical Center), Boston. The dataset consists of several features including the vital signs, demographics, etc. and is labelled manually to detect Sepsis. The authors in [5] propose a new scoring system known as the InSight score. The classifier shown in the paper uses 8 easily accessible patient variables from the MIMIC-III dataset [4]. The dataset was reduced to patients above the age of 15, with a few additional constraints. Hourly readings of a patient were recorded and the patient was labelled as Sepsis positive

if there was an increase in the SOFA score by at least 2 points in a time frame of 72 hours - 24 hours before to 48 hours after antibiotic administration or blood culture draw. The classifier obtains an ROC AUC score of 0.88 for predicting Sepsis at its onset, 0.74 four hours before onset and 0.73 four hours before onset on a sparse version of the dataset.

The second dataset was first provided during the PhysioNet Computing in Cardiology Challenge 2019 [6]. The public version of the dataset contains the data of 40,336 patients from two hospitals, with a total of 41 features per patient. This dataset is also a time series dataset with each row corresponding to one hour of the patient's data. The labelling of the dataset has been done in accordance with the Sepsis-3 definition [1], with the time frame of monitoring each patient being 36 hours. Additionally, if a patient is detected as Sepsis positive, data up to 6 hours prior to the onset has been labelled as positive for the early prediction of the disease. Submissions to the challenge were made on a hidden test set. Submissions were judged on a utility score which rewarded models for correct predictions and penalised them for incorrect ones.

In [7], the authors limit the training to 30 features from the dataset. Patients who were labelled as positive from the first hour or had less than 6 positive labels were excluded. Feature importances were calculated after performing 5-fold cross-validation of a variety of models. Ultimately, three models - Random Forest, LightGBM and XGBoost - were used as an ensemble model by assigning weights to their output probabilities. An ROC AUC score of 0.79 was achieved by the ensemble model.

A deep learning approach is seen in [8], where LSTMs have been used to predict the onset of Sepsis. The inspiration for this method comes from the temporal structure of the data. The sequential data was used to generate embeddings. These embeddings were aggregated using attentional multihead aggregation where 1, 8, 16 heads had been used. The highest ROC AUC score of 0.84 was achieved with the 16-heads-HEA-LSTM model.

In another attempt, the authors in [9] augment the dataset by including new features such as a PartialSOFA score, shock index (ratio of heart rate to systolic blood pressure), etc. Augmented features also included measurements of parameters over a look-back window. The window was treated as a hyper-parameter and its size was tuned accordingly. An alternative method of labelling was used based on the utility score defined in the challenge [6]. The Light Gradient Boosting model was used and a utility score of 0.36 was obtained, which was the highest among all submissions in the challenge.

III. Methodology

A. Data Imputation and Visualization

Data for this research was used from the PhysioNet Challenge [6]. Two columns - Unit 1 and Unit 2 - were dropped reducing the total features to 39. A variety of data imputation techniques were explored to enhance the performance of the models. Data was available in the form of 40,336 PSV (Pipe Separated Values) files, each file corresponding to one patient. Imputation was performed on each PSV separately. The overall mean of each feature was computed before the filling process. The following imputation techniques were explored:

1) Front Fill: Each null/missing value is filled with the last recorded data point value. The values before the first non-null data point 'A' are given the same value as 'A'.
2) Back Fill: Each null value is filled with the value of the next available non-null data point. The null values occurring after the last recorded non-null data point 'A' are given the same value as 'A'.
3) Linear Fill: The missing values between two non-null data points are filled in a linear fashion. For example, if there are 3 null data points X, Y, Z between two data points A and B having values 1 and 5, X, Y, Z are given values 2, 3, 4 respectively. The values before the first and after the last recorded non-null data points are filled using the front fill convention.
4) Mixed Fill: To perform this method of imputation, the correlation matrices of the datasets obtained from the previous mentioned techniques are used. Algorithm 1 shows the mixed filling algorithm.

For features containing all null values, the overall feature mean was used to fill all values for the patient.

Algorithm 1 Mixed Fill

correlation(m,f) returns correlation of feature f with the $SepsisLabel$ in filling method m

for f in feature list **do**
  $fill\_method \leftarrow NULL$
  $max\_correlation \leftarrow 0$
  for m in filling method **do**
    if $|correlation(m, f)| > |max\_correlation|$ **then**
      $fill\_method \leftarrow m$
      $max\_correlation \leftarrow correlation(m, f)$
    end if
  end for
  Fill feature f using filling method $fill\_method$
end for

Once the imputation was performed, a combined dataset was formed by concatenating the data of all the patients. This resulted in one large dataset which was then used to perform cross-validation for every model. In this way, each row could be treated as an independent sample to train and test the models. This was done for all models except the LSTM network as it required individual patient data to capture relations within the time-series dataset. Due to the huge bias towards negative samples in the dataset, undersampling was performed on the test set to get a clearer idea of the performance of the models. The negative samples were undersampled randomly to get the ratio of negative to positive samples in the test set down to 2:1. Training sets were undersampled in some cases to improve the false negative count.

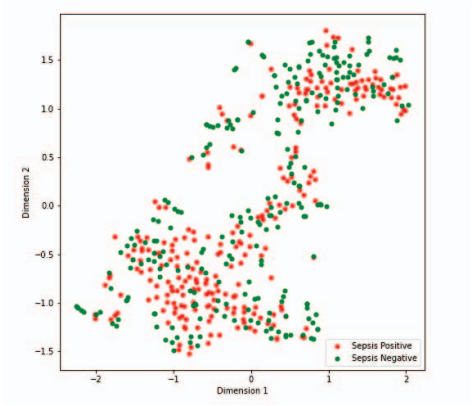


Fig. 1: Plot of mixed fill dataset with reduced dimensions using t-SNE with perplexity=50.

Since the dataset consisted of 38 features excluding the Sepsis label, the dimensionality had to be reduced to visualise it. The t-SNE algorithm [10] was used to reduce the dimensionality of the dataset to two dimensions. Fig. 1 shows the plot of randomly chosen data points with their dimensionality reduced on the mixed fill dataset. It is observed that there is no form of clustering in the dataset and there is a significant overlap between the two classes.

B. Models Used

The following six models were used for comparison:

- Random Forest [11]
- Logistic Regression [12]
- Light Gradient Boosting Machine (LGBM) [13]
- eXtreme Gradient Boosting (XGBoost) [14]
- User-defined Neural Network
- User-defined Long Short-Term Memory (LSTM) network [15]

The models were built using the Scikit-learn [16], LightGBM [13], XGBoost [14] and Keras [17] libraries in Python. Each model was trained using all the imputation methods separately. Section IV shows a detailed comparison of all models using every filling method.

IV. Results and Discussion

Cross-validation was performed on all the models for each imputation technique. K-fold cross-validation was chosen as the method of cross-validation. This would ensure that the models were performing well on the complete dataset and not specific segments of it. The models were tested based on five metrics - Accuracy, Precision, Recall, F1 Score and ROC AUC Score.

Tables 1, 2, 3, 4 show the metrics of the models after performing cross-validation on the front, back, linear and mixed fill datasets. A generic trend can be observed from all the tables. The LGBM, Random Forest and XGBoost models perform exceptionally well on all the datasets. The neural network model performs well but does not obtain a high precision score. The LSTM network produces average results and the Logistic Regression model fails and produces below par results. It is seen that the LGBM classifier shows the most superior overall performance for every task except the back filling dataset where the random forest model performs better. Even for the back filling task, the LGBM classifier obtains a higher recall than the random forest resulting in a lower false negative count. The random forest model obtains a high accuracy, recall and ROC AUC score but also produces higher false positives resulting in a lower precision and F1 score. The XGBoost model shows the best recall in every task indicating that it predicts positive samples remarkably. But the model shows an average precision indicating that it has a higher false positive count. The neural network model shows a high ROC AUC score and shows above par performance for the other metrics. The LSTM network performs better than the logistic regression model but does not perform well when compared to the other models. The logistic regression model obtains poor results consistently for each dataset.

To understand the reasoning of the model behind making the predictions, feature importances of the LGBM model were calculated. Two types of feature importance calculation techniques were used. Firstly, global feature importance of the model was calculated. This would provide an idea of which feature contributed the most during the complete training process. Secondly, local feature importance of the model was calculated to find out which feature contributed the most while making individual predictions.

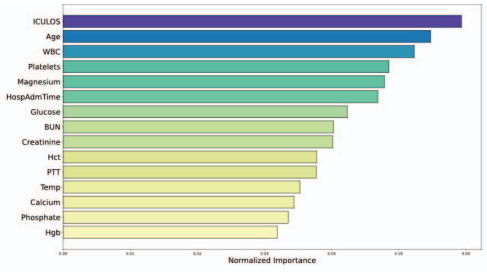


Fig. 2: Relative global importance of the top 15 features.

TABLE 1: Average metrics after cross-validation on the front fill dataset.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
LGBM Classifier	0.9843	0.9709	0.9824	0.9766	0.9973
Random Forest Classifier	0.9751	0.9495	0.9773	0.9632	0.9957
XGBoost	0.9277	0.8264	0.9912	0.9013	0.9754
Neural Network	0.8696	0.7944	0.8212	0.8076	0.9317
LSTM	0.8051	0.7215	0.6767	0.6983	0.8240
Logistic Regression	0.7097	0.5566	0.6346	0.5930	0.7471

TABLE 2: Average metrics after cross-validation on the back fill dataset.

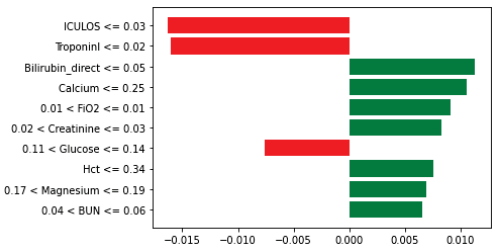
Model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
Random Forest Classifier	0.9795	0.9526	0.9876	0.9698	0.9967
LGBM Classifier	0.9794	0.9521	0.9880	0.9697	0.9964
XGBoost	0.9271	0.8248	0.9919	0.9007	0.9755
Neural Network	0.8903	0.8157	0.8666	0.8404	0.9477
LSTM	0.8121	0.7335	0.6852	0.7085	0.8402
Logistic Regression	0.7184	0.5690	0.6406	0.6026	0.7587

TABLE 3: Average metrics after cross-validation on the linear fill dataset.

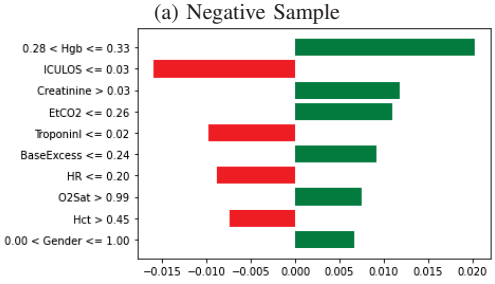
Model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
LGBM Classifier	0.9883	0.9784	0.9866	0.9824	0.9984
Random Forest Classifier	0.9789	0.9544	0.9838	0.9689	0.9969
XGBoost	0.9309	0.8305	0.9958	0.9057	0.9807
Neural Network	0.8855	0.8015	0.8728	0.8356	0.9453
LSTM	0.8126	0.7417	0.6719	0.7051	0.8274
Logistic Regression	0.7157	0.5648	0.6408	0.6004	0.7541

TABLE 4: Average metrics after cross-validation on the mixed fill dataset.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
LGBM Classifier	0.9888	0.9767	0.9902	0.9834	0.9987
Random Forest Classifier	0.9792	0.9527	0.9865	0.9693	0.9969
XGBoost	0.9335	0.8374	0.9934	0.9088	0.9795
Neural Network	0.8872	0.8085	0.8670	0.8367	0.9463
LSTM	0.8064	0.7290	0.6671	0.6967	0.8190
Logistic Regression	0.7168	0.5662	0.6429	0.6021	0.7577



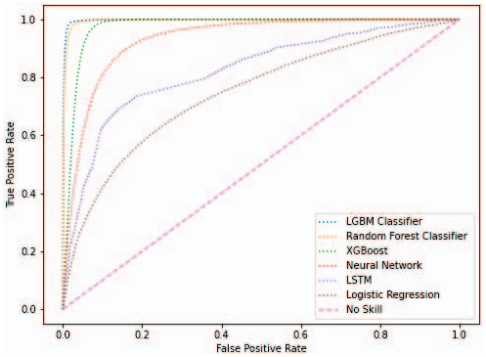
(a) Negative Sample



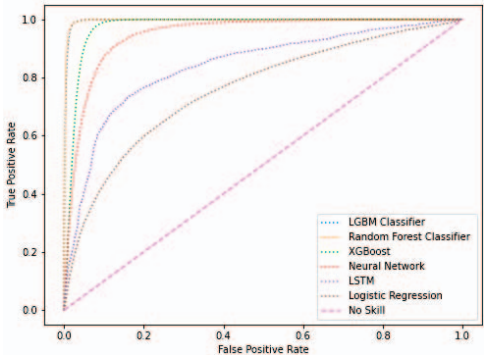
(b) Positive Sample

Fig. 3: Local feature importance generated using LIME.

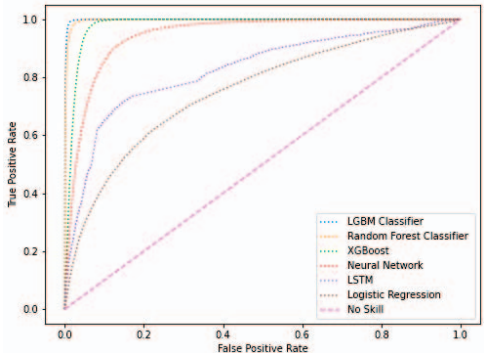
Fig. 2 shows the relative global importance of the top 15 features of the LGBM model when trained on the mixed fill dataset. It is observed that the three most influential features are the ICULOS (ICU Length of Stay), Age and WBC (White Blood Cell) count of the patient. Local feature importance for the model was calculated with the help of LIME [18]. Fig. 3 shows the relative importance of features generated using LIME for a Sepsis negative and Sepsis positive sample. For the negative sample, it can be seen that the ICULOS and Troponin I levels are among the most crucial factors while making the prediction. For the positive sample, the Hgb (Hemoglobin) and ICULOS contribute the most towards the result. It is noticed that these two factors affect the prediction inversely. Even though the ICULOS tends to drive the prediction towards the negative class, the higher magnitude of Hgb and other features results in the patient being predicted as Sepsis positive. Another key point that can be observed is that the global and local feature importances of the model tally for many of the most important features.



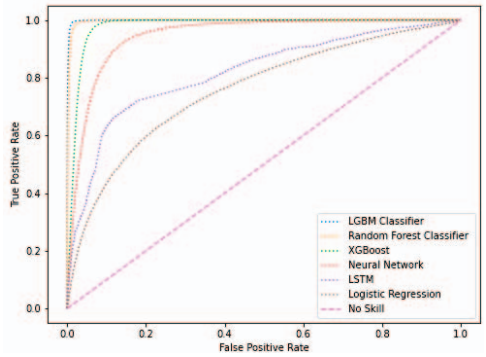
(a) Front Fill



(b) Back Fill

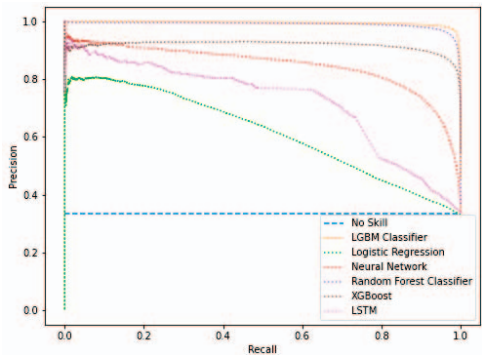


(c) Linear Fill

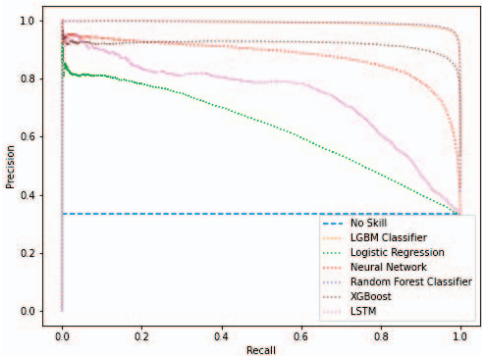


(d) Mixed Fill

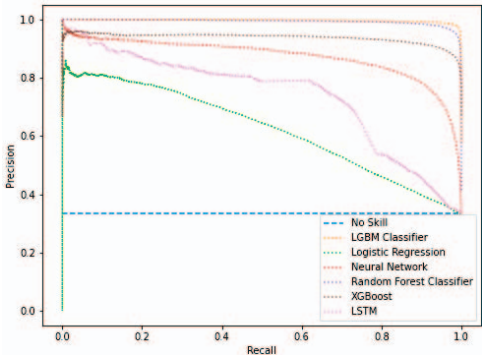
Fig. 4: ROC curves for all models in every imputation method.



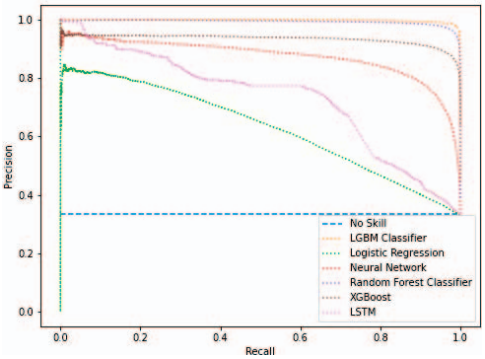
(a) Front Fill



(b) Back Fill



(c) Linear Fill



(d) Mixed Fill

Fig. 5: Precision-Recall curves for all models in every imputation method.

Figures 4 and 5 describe the ROC and Precision-Recall curves of the models for each filling method. Among the various models and imputation techniques tested, it is seen that the LGBM model produces exceptional results when trained on the mixed fill dataset. Since the mixed filling algorithm incorporates the best aspects of all the other filling algorithms, the metrics obtained on it are higher. Other models which produce high ROC AUC scores lack in either getting a high precision or a high recall resulting in a higher count of false positives and false negatives.

V. Conclusion

In this paper, four methods of imputation were analyzed for the early detection of Sepsis in patients. Each method was tested using six models. By comparing the different imputation techniques, it can be concluded that the mixed filling algorithm produces the best results as it selects and integrates the features from the most suitable filling algorithm. Of all the models tested, the LGBM classifier yields the best metrics. This model could play a pivotal role in the early and accurate detection of Sepsis in patients thereby ensuring proper treatment.

Acknowledgment

The authors would like to thank Mr. Gehan Chopade and Dr. Kavita Sultanpure for their help and support in making this work possible.

References

[1] Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J. D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J. L., Angus, D. C. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA, 315(8), 801–810. https://doi.org/10.1001/jama.2016.0287.

[2] Sepsis: https://www.who.int/news-room/fact-sheets/detail/sepsis.

[3] Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, Gurka D, Kumar A, Cheang M. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Crit Care Med. 2006 Jun;34(6):1589-96. doi: 10.1097/01.CCM.0000217961.75225.E9. PMID: 16625125.

[4] Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). MIMIC-III, a freely accessible critical care database. DOI: 10.1038/sdata.2016.35.

[5] Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., Shimabukuro, D., Chettipally, U., Feldman, M. D., Barton, C., Wales, D. J., Das, R. (2016). Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. JMIR medical informatics, 4(3), e28. https://doi.org/10.2196/medinform.5909.

[6] 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.

[7] M. Fu, J. Yuan, M. Lu, P. Hong and M. Zeng, "An Ensemble Machine Learning Model For the Early Detection of Sepsis From Clinical Data," 2019 Computing in Cardiology (CinC), Singapore, Singapore, 2019 (pp. 1-4), doi: 10.23919/CinC49843.2019.9005710.

[8] L. Liu, H. Wu, Z. Wang, Z. Liu and M. Zhang, "Early Prediction of Sepsis From Clinical Data via Heterogeneous Event Aggregation," 2019 Computing in Cardiology (CinC), Singapore, Singapore, 2019 (pp. 1-4), doi: 10.23919/CinC49843.2019.9005879.

[9] Morrill JH, Kormilitzin A, Nevado-Holgado AJ, Swaminathan S, Howison SD, Lyons TJ. Utilization of the Signature Method to Identify the Early Onset of Sepsis From Multivariate Physiological Time Series in Critical Care Monitoring. Crit Care Med. 2020 Oct;48(10):e976-e981. doi: 10.1097/CCM.0000000000004510. PMID: 32897664.

[10] van der Maaten, L., Hinton, G. (2008). Visualizing Data using t-SNE . Journal of Machine Learning Research, 9, 2579–2605.

[11] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324.

[12] Logistic Regression Scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[13] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146-3154).

[14] Chen, T., Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. arXiv:1603.02754v3.

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 15, 1997), 1735–1780. DOI:https://doi.org/10.1162/neco.1997.9.8.1735.

[16] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python. 12(85):2825-2830, 2011.

[17] Chollet, F., others, 2015. Keras: https://github.com/fchollet/keras.

[18] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938.