Gradient Boosting Machine for Early Prediction of Sepsis Using Clinical Data

Soufiane Chami[1], Kouhyar Tavakolian[1]
[1] School of Electrical Engineering & Computer Science
University of North Dakota, Grand Forks, ND, United States

Abstract

Sepsis is a severe medical condition caused by the body's extreme response to an infection leading to tissue damage, organ failure, and even death. The emergence of advanced technologies such as Artificial Intelligence and machine learning allowed faster exploration of advanced ways to recognize sepsis cases. In this paper, we present an efficient model, based on gradient boosting machines, to optimize sepsis risk monitoring and mitigation. Our model can also serve as a general framework to solve survival analysis problems in different domains. We also provide generic techniques for statistical feature extraction, which can drastically increase the aptitudes of detection without specific domain knowledge. The proposed model is also capable of handling a high rate of missing values of patient records. There are four main components of the model: features engineering, sepsis risk estimation, alert system optimization, and finally sepsis alarm predictions. Using the utility score metrics provided by PhysioNet/Computing in Cardiology Challenge 2019, the local validation score are 32.9% for utility score and 82.1% for Area Under the ROC Curve. Such results places us among the 10 places in the public leader-board.

1. Introduction

Sepsis is a severe medical condition caused by the body's extreme response to an infection leading to tissue damage, organ failure, and death. Every 3-4 seconds, at least one person dies because of sepsis worldwide. Sepsis affects about 1.6 million Americans yearly [1]. As a leading cause of death in US hospitals, sepsis costs the US about 24 billion USD, 6.2% of the total hospital costs in 2013 [2].

The early detection of sepsis has proven to be a key factor in increasing the efficiency of antibiotic treatment for septic patients. Related studies [3] demonstrated that early detection prevents 80% of death cases caused by sepsis. On the other hand, it was reported that the sepsis mortality significantly increases with the length of stay of the septic patient in the hospital. In other words, delayed recognition of sepsis exacerbates the risk of death of a septic patient by 7.6% every hour [4].

Furthermore, with the emergence of more advanced technologies, there is a significant amount of Electronic Health Records (EHRs) that became available. The EHRs are a systematic collection of data used as health indicators of a patient. The growing availability of the EHRs brought so many interests and opportunities to develop more advanced predictive models to early recognize septic patients. The electronic health records can be time-based features that change over time (time series) like heart rate or blood pressure. Also, they can be static features like demographic information such as age and gender.

The clinical definition of the systemic inflammatory response (SIRS) to infection, specifies four conditions in which only two of them are sufficient to trigger an alert of sepsis [5]. These conditions are listed as follows:

1. Temperature $> 38^o$C or $< 36^o$C
2. White blood cell count $> 12,000$ per ml or $< 4,000$ per ml, and $> 10\%$ for immature (band) forms
3. Heart rate of > 90 beats per minute
4. Respiratory rate of > 20 breaths per minute or partial pressure of CO2 of $<$32 mmHg

The rest of the paper is organized as follows: In section II, we outline the task definition of the PhysioNet challenge of this year. The clinical data provided is also discussed and presented. Section III presents the proposed methodology. In section IV, we present the obtained results. Finally, we discuss our conclusion and some future research directions.

2. Physionet Challenge and Data

Saving the life of septic patients comes with a combination of efforts from a different perspective of sepsis. While more advanced instruments allow closer and more accurate follow-up of the septic patient situation, they only can tell the current situation of a patient in a potential risk of sepsis [6].

2.1. Problem definition

The objective of this challenge is to build a machine learning model to predict sepsis 6 hours before the clinical prediction of sepsis.

Most of the time, when patients arrive at the confirmed sepsis shock stage, it's most probably too late to save their lives. Launching the antibiotic culture takes time and is less effective when a patient is in a serious stage of sepsis.

Medical instruments can tell what has already happened to the patient but alone they cannot tell a lot about the future development of the patient situation. This later information is more relevant for saving an infectious patient from death.

Each time we detect septic subjects one hour earlier, we get a more 4-8% chance to save their lives. With the ancient tools of data analysis in the last decade, scientists still face a lot of limitations to detect sepsis early enough [7].

With the emergence of machine learning techniques, significant progress has been made to improve the current detection tools. [1, 8, 9].

2.2. Related works

Early detection is not a conventional classification/regression problem. It's a hybrid type of problem that requires classification accuracy that looks to distinguish septic patients from others (is-septic), and regression to estimate how early a septic patient can be detected (time-to-event). The classification aspect is evident in our training process, however, the regression aspect is fulfilled in the process of definition of the optimal threshold. The threshold is the only parameter that can help the model to detect sepsis early enough within the defined time limit set by the end-user. During the training and evaluation process, the AUC is the principal component. Then, others expect should also be taken into consideration such as misdetection and false alarm rates. These factors allow a better interpretation of the model and an easier comparison between different approaches.

The top solution proposed during the Physio-Net Challenge was very similar in terms of features engineering techniques and the modeling mechanisms. In these sections, three approaches will be reviewed and presented. The three of them obtained a rank in the top 10 range at the Physio-Net Challenge 2019.

2.2.1. Sliding Windows and XGBoost Ensemble approach

The most performing approach [10] included two steps to analyze sepsis data: features selection, features engineering and then ensembling. The features engineering schema separated all the covariates in the data into two clusters: the first cluster of covariates with the minimum number of missing values with a threshold of 10%. For the features with low missing values:

- Aggregation using Sliding windows of 5- and 11-hours frames while applying different methods:
  - Min, max, mean, median, and variance.
  - Min, max, mean, median, and variance.
  - Quantile of 95%,99%,5%, and 1%
  - Other features to capture Long and Short-Term Dependencies like Shannon Entropy energy, mean of the first differences and the length of stay for a specific patient.

The total number of features is around 410 features. To reduce the number of covariates and reduce the bias and variance, the second layer of feature selection was applied for two main objectives: select the best performing features and five best hyperparameters. The described procedure above was applied to only 10% of the data. One interesting point about the feature selection in this approach is that it was performed using BoostARoota algorithms based on XGBoost. The procedure is that for each feature we create another shadow feature that is basically similar to noise and use only both of them on the training. Then, the model should report the importance of both features. If shadow feature is more relevant than the original features, this latter will be dropped.

Once the choice is made about which are the best performing parameters and best-performing covariates, the training start using the remaining 90% of the training data. In fact, in this approach, we construct five randomly disjoint sets and apply an under-sampling technique to balance class in each set separately. In the end, 5-XGBoost models are trained using each set with 5-fold-cross-validation. The final output is calculated using the geometric mean of the five outputs of the trained model. As a result, this model training structure has achieved an AUC score of 0.833 and an accuracy of 0.8440.

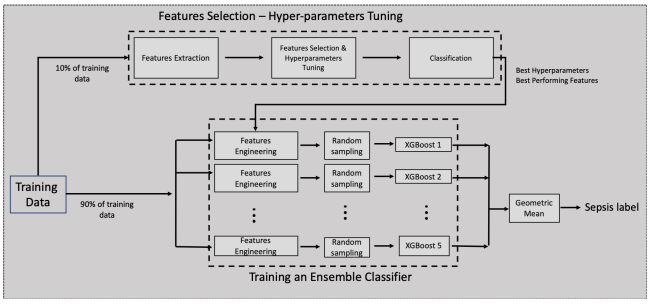

Figure 1. The training strategy of the proposed method to solve survival analysis problem.

2.2.2. Signature transformation

The approach of the second winning solution was a using similar technique to the first team in terms of its objective (i.e. capturing the long- and-short-term time dependencies) and also in the use of learning algorithms. However, it was more abstract and mathematically complex. This technique is called: signature-based features in time series. It's highly used in healthcare and bioinformatics to solve DNA sequences related problems. Ilya Chevyrev [2] is the inventor of this technique in the modeling and the machine learning domain. The second winning used the technique and applied a sliding approach to each column. Then they injected the resulted table to an Ensemble of XGBoost models. While it's very powerful to use signature-based features, it's still more complex methods and remains less flexible than the first winning solution. In this case, it's harder to interpret the features of the columns and their impact on the model performance. This approach obtained a score of 0.433 on the public leaderboard.

2.2.3. Time Phased Model

There is a particular characteristic of the dataset of the PhysioNet challenge. The sepsis events are hugely time-dependent, and the risk of sepsis depends a lot on the previous events in the long and short terms. That's why most of the top winners used time-based features to capture all these dependencies. While the last two winners have based their learning models on GBM, the third team led by Xiang Li from the University of Ping An technology at Beijing, proposed a combination of neural network and XGBoost in different stages depending on the length of stay of the patient. These stages are defined as follows: Early stage (1-9 hrs.), Middle Stage (10-49 hrs.), and Late Stage (50+ hrs.). This approach obtained a score of 0.415 with 10-fold cross-validation.

Approach	AUC score	Utility Score
Signature Transformation	86.8%	0.433
Sliding Windows	83.3%	0.422
Time Phase Stages	N.A	0.415

3. Proposed Approach

In this section, we present the data pre-processing stage, features extraction, and an overview of the proposed model.

3.1. Data Features

There are more than 40 health variables used to track the health situation of the patients in this challenge. The 40 columns are classified into three classes: vital signs, lab tests, and static variables.

3.1.1. Vitals signs

Most of these features concern the main screening signal that would reflect the continuous situation of patient health. There are 8 vital signs provided in the data and we can cluster them into three main categories: ECG signals, Pulse signals, and Temperature.

- **ECG signals:** The main columns related are the heart rate, systolic blood pressure, and diastolic blood pressure. The severity of sepsis is very correlated to the number of beats per minute. Many studies have shown that the more sepsis gets worse, the more we observe ECG abnormalities. This can be explained by the loss of excitability in cardiac tissue during the sepsis.
- **Pulse signals:** Respiration rate, O2Sat, and EtCO2 (End-tidal carbon dioxide)
- **Temperature:** Symptoms of sepsis include a fever above 101^{o}F (38^{o}C) or a temperature below 96.8^{o}F (36^{o}C) heart rate higher than 90 beats per minute. So, tracking the temperature is very important for the early detection task. The predictive model uses a certain number of the previous data points from the current time and tries to detect any significant shift on temperature (drop or rise).

3.1.2. Lab tests

The role of laboratory tests is very significant for the early detection of sepsis. Since the main definition of sepsis is the body's systemic inflammatory response to a bacterial infection [11]. The spread of bacteria in the blood (bacteremia) makes it a big indicator of sepsis infection. About 25 lab tests have been provided in the data for sepsis detection. These features include: White blood cells count, Blood urea nitrogen, Lactic acid, partial thromboplastin time, Leukocyte count, Platelets.

3.1.3. Demographics features

Some studies have shown that sepsis mortality has a little thing to do with demographics. Especially age and length of stay. In our current model, we still find this is not the case. More results will be shown in the next paragraphs [12]. The main demographic features in the data are gender, age, and length of stay. Surprisingly, we found out in our model, that age is a critical factor to predict sepsis risk.

3.2. Data Pre-processing & Features

Before the learning step, the clinical dataset we have for the challenge contains a lot of inconsistencies. For our model, we have performed many preprocessing techniques to ensure the consistency of the data and create new features. The processing pipeline can be described as follow:

(1) Handling missing values: Several lab tests features have up to 90% of missing values. The data resolution is hourly based, and it is difficult to perform lab experiments every hour for every patient. Which explains the high ratio of missing values in lab test columns. However, the values in these columns are very important and removing them would not be the best idea. In order to handle missing values:

1. We have performed interpolation which fills a missing value with the mean of the two consecutive non-missing values.
2. The remaining missing values are filled using backward and then forward values.

(2) Lag features: The purpose of these features is to capture the long- and short-time dependencies. We performed different sliding windows with mean of the last 1,2,..,9 hours.

(3) Data binning: To reduce some variance in the signal columns. We create new features that aggregate the signal value into ranges and intervals. We have based the data binning on the max and min of each signal. There is an option to define range limits using Random Forest. We expect to explore it in future work.

(4) Count Encoding of Categorical features: applied mainly on the demographic features such as age and gender.

3.3. Gradient Boosting Machine Ensemble

3.3.1. Model description

As explained earlier, our proposed model to predict sepsis is based on two layers: sepsis risk prediction (classification task), and optimal cut-off assessment (optimization layer).

The first layer will be based on Ensemble of 5 Gradient Boosting Models of Light-GBM. GBM algorithms are based on a fundamental idea, which is weak learners (decision trees) make strong predictors. In other words, the GBM model makes an iterative learning process that starts with one weak classifier, then, builds a second model that learns from the errors made by its predecessor model. The process is iterated multiple times to achieve the best possible performance. This is a common treat between all the GBM models such: XGBoost and Light-GBM.

In our proposed model, we have decided to pick Light-GBM as a candidate to predict sepsis. There are multiple benefits of considering Light-GBM as one of the best candidates to make effective sepsis predictions. First, its capability to handle missing values. During the training process of LGBM, the missing values are handled in a way that reduces the loss error as much as possible during the trees split process. Second, LGBM has also the capability to handle categorical features without the need for manual processing. Finally, in terms of computational expenses, the Light-GBM is much faster and more accurate than XGBoost [13]. Finally, Light-GBM supports parallel computing using GPU and OpenMPI. This gives a computational convince and advantage to make faster iterations and reduce the experimentation time.

Finally, it's worth noting that on the contrary of most proposed models during the PhysioNet challenge, we are the only team using Light-GBM to predict sepsis.

3.3.2. Training Process

The training data provided is seriously unbalanced. Among 40336 patients, only 2932 patients are septic. This is a ratio of 7.27, and most of the samples are negative with a ratio of 92.73%. One of the fundamental strategies we considered for sepsis detection is to consider stratified cross-validation for validation and stratified sampling for testing. In other words, the data is split into 2 sets: train data, test data. The train data will be split further during the cross-validation process with stratified 5-folds. Stratification is based on the Sepsis Labels and consists of keeping the same ratio of negative (non-septic) and positive samples (septic). The figure 2 can describe the process:

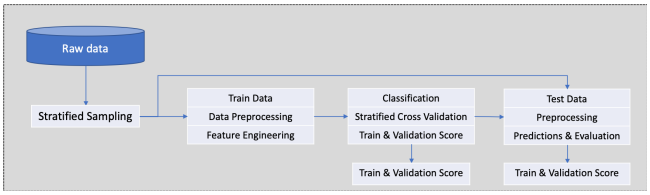

Figure 2. Validation Schema

It's worth noting here that there is a tricky issue about sampling methods on the data we are dealing with. The provided data is labeled per observation per hour per patient, which may lead to think that one sample is equal to one observation. This thinking will lead to a huge data leak during the modeling process between patient recording. We can call it Inter-leak. The definition of the sample should be the whole recording collected from each patient. Still, this approach does stop the inter-leak but can't stop the leak between observation of the same recording.

In our experiment, we assume that the Intra-leak will not have a significant impact on the training and validation process.

3.4. Cut-off Value Optimization

The decision of this alarm is not only based on the prediction of the risk probability but also, it's based on what level of risk the alarm should be triggered. A good classifier with a bad triggering mechanism will do poor during the deployment stage. This applies to all similar problems characterized by serious class unbalance and time-sensitive predictions and requires early detection within a specific time interval. For instance, Customer Churn Prediction in banking and insurance businesses. This type of problem is extremely similar to the sepsis problem that we are discussing in this paper. In this direction, in order to mitigate the sepsis risk effectively we need to:

1. Predict as accurately as possible the risk of sepsis development for a subject. This is evaluated by the AUC score.
2. Make sure we trigger the alarm ON-TIME for the right subject. This relies principally on the defined threshold of sepsis.

To be successful in the second evaluation step, we propose the validation schema described in the figure 3 below to build a robust machine learning model.

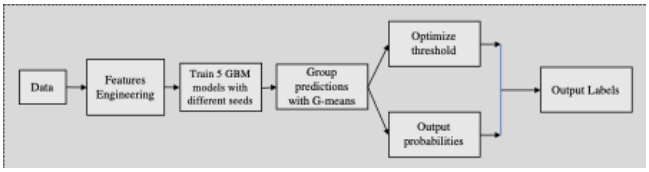

Figure 3. Proposed Machine Learning Pipeline.

3.4.1. Alert System

The process to define the AUC threshold stars by comparing the out-of-fold predictions with actual values and calculate the number of false-positive and false-negative according to different thresholds values. The optimal threshold that allows minimum false positive and false negative is chosen.

As described above, our training process is not subject-wise but the prediction is. In other words, during the training, we don't distinguish between each subject and we look at the training data as one subject or one table and each timestep as an observation independently. On the other hand, during the prediction happens for each subject separately.

The question here is should we apply this process on a subject level or the whole training dataset similarly to the training? To answer this, there are options to explore:

1. Timesteps observations across the training data.
2. Timesteps observations across the subject record.

The first option evaluates the thresholds by looking at the observations in the training data regardless of the subject record they belong to. This will be more focusing on how early the sepsis was detected if any.

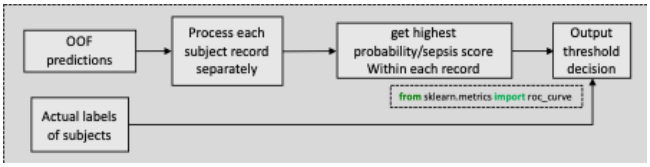

Figure 4. Approach No.1 that defines threshold by looking at subjects as data-point by aggregation.

The second option will aggregate the sepsis scores by each record and will analyze the threshold by looking at each subject as one observation. This option focuses more on whether we could detect the right septic patients or not regardless of how early this happened.

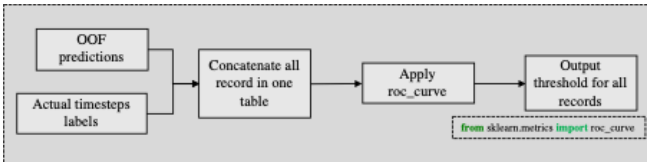

Figure 5. Approach No.2 that defines threshold by looking at timesteps in each record as data-point and regardless of the subject.

From both approaches, we can see there is a difference in the counting of the false positive and false-negative ratios. This impacts the false alarm and sepsis misdetection rates and the normalized utility score at the end. The figure below describes the workflow of our experiment is evaluating both approaches.

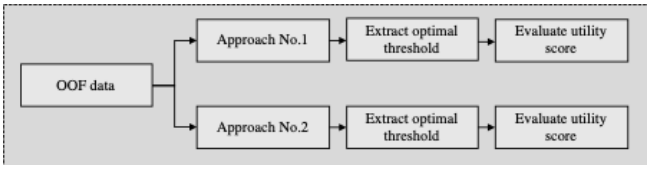

Figure 6. Experiment workflow. OOF means out-of-fold data collected during the cross-validation process

3.4.2. Results: optimal cut-off

To observe the impact of each method, we can compare them regarding three metrics: utility score, false alarm, misdetection rate.

From the obtained results, we can conclude that the Approach No.1 is better and we should define the optimal cut-off by considering one Timestamp observations as an

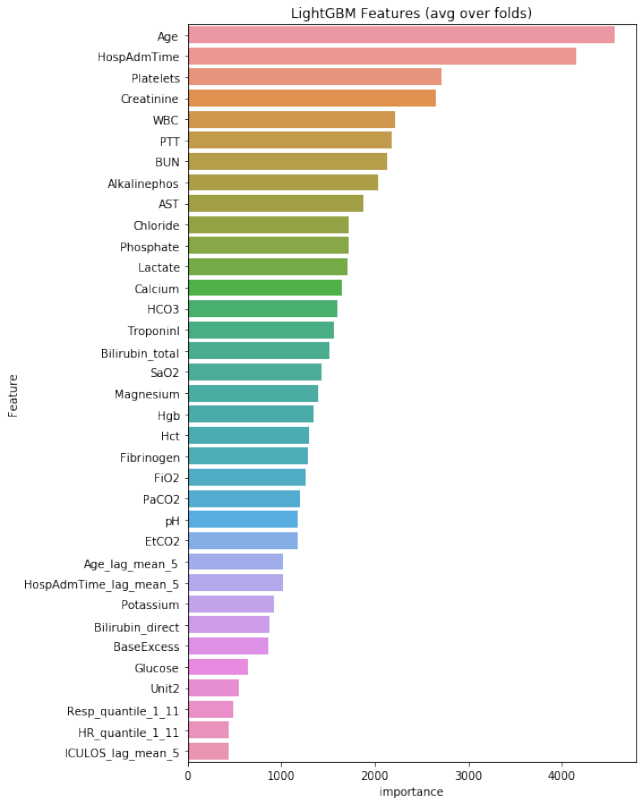

Figure 9. Features Importance is defined based on information gain during the training process of Light-GBM

element of the whole training data and not an observation isolated within a patient record. It's worth noting that a better schema to find the optimal threshold can be explored. While this can be a solid schema to empirically define the optimal threshold, there is still a risk of over-fitting, which, in our estimation, can be handled by performing this experiment multiple times using the same model but with different seed in the hyper-parameters. This can help to reduce variance and over-fitting.

nb of test subjects	Approach	Threshold	False alarm	Misdetection	Utility
10084	No.1	33.22%	26.09 %	27.271 %	31.6%
	No.2	40.65%	27.271 %	17.735 %	32.9%

Figure 7. Comparison of performance scores for both threshold optimization approaches.

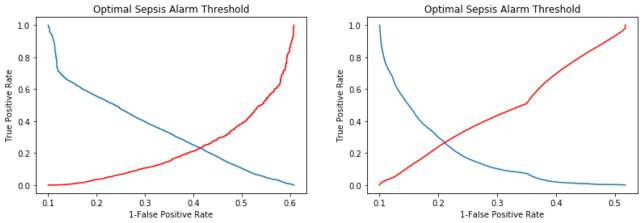

Figure 8. Optimal Cut-off is defined using the ROC curve. Approach No.1 (right) and Approach No.2 (left)

Form the experimental results, we conclude that the optimal cut-off should defined by analysing subjects as data points rather considering all the sub-labels within the records of each subject.

3.5. Relevant Factors

Concerning the importance of the features, the figure 9 exhibits the most relevant features to detect sepsis according to the Light-GBM. We observe the Age is one the most important factor to define sepsis risk. This is very correlated with the state of the art about sepsis risk estimation. Besides to age, length of stay and other characteristics of Blood Cells are very deterministic to define sepsis risk.

3.6. Discussion & Conclusion

In this paper, we proposed an efficient and fast sepsis early detection system that requires less information with cheaper computational cost and provides great prediction capabilities. Our model can help in diagnosing sepsis in the hospital at least 6 hours before a human doctor save up to 31.6% sepsis patients from fatal death. This 75 000 lives saved from sepsis if such an algorithm is implemented in each US hospital [14]. We believe such work can help in making more progress in saving

It will be much easier to implement it in the medical devices with lower computational capacity.

References

[1] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation 12 1997;9:1735–80.

[2] Umscheid C, Betesh J, VanZandbergen C, Hanish A, Tait G, Mikkelsen M, French B, Fuchs B. Development, implementation, and impact of an automated early warning and response system for sepsis. Journal of Hospital Medicine 09 2014;10.

[3] Kumar A, Roberts D, Wood K, Light B, Parrillo J, 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. Critical care medicine 07 2006;34:1589–96.

[4] Frost R, Newsham H, Parmar S, Gonzalez-Ruiz A. Impact of delayed antimicrobial therapy in septic itu patients. Critical Care 09 2010;14:1–2.

[5] Hotchkiss R, Moldawer L, Opal S, Reinhart K, Turnbull I, Vincent JL. Sepsis and septic shock. Nature Reviews Disease Primers 06 2016;2:16045.

[6] MA R, Josef C JR, Shashikumar SP WM, Nemati S CG, A S. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. Critical Care Medicine An International Journal 09 2019;in press.

[7] Zhang Y, Lin C, Chi M, Ivy J, Capan M, Huddleston J. Lstm for septic shock: Adding unreliable labels to reliable predictions. 12 2017; 1233–1242.

[8] Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems 01 2012;25.

[9] Futoma J, Hariharan S, Heller K. Learning to detect sepsis with a multitask gaussian process rnn classifier 06 2017;.

[10] Morteza Zabihi Serkan Kiranyaz MG. Sepsis prediction in intensive care unit using ensemble of xgboost models: the physionet/computing in cardiology challenge 2019. Critical Care Medicine An International Journal 09 2019;in press.

[11] Fan SL, Miller N, Lee J, Remick D. Diagnosing sepsis – the role of laboratory medicine. Clinica Chimica Acta 07 2016;460.

[12] Menezes B, Araújo F, Amorim F, Santana A, Soares F, Souza J, Araújo M, Santos L, Rocha P, Gomes M, Neto O, Júnior P, Amorim A, Biondi R, Ribeiro R. Comparison of demographics and outcomes of patients with severe sepsis admitted to the icu with or without septic shock. Critical Care 11 2013;17:P48.

[13] Anghel A, Papandreou N, Parnell TP, Palma AD, Pozidis H. Benchmarking and optimization of gradient boosted decision tree algorithms. CoRR 2018;abs/1809.04559. URL http://arxiv.org/abs/1809.04559.

[14] et al. RC. Prevalence, underlying causes, and preventability of sepsis-associated mortality in us acute care hospitals 02 2019;URL https://doi.org/10.1001/jamanetworkopen.2018.7571.

Address for correspondence:

Soufiane CHAMI
Biomedical Engineering Research Complex - UND
soufiane.chami@und.edu