Interpretation of Artificial Intelligence Algorithms in the Prediction of Sepsis

Induparkavi Murugesan[1], Karthikeyan Murugesan[2], Lingeshwaran Balasubramanian[3], Malathi Arumugam[4]

[1]Father Muller Medical College, Mangalore, India, [2]Georgia Institute of Technology, Atlanta, USA, [3]Sri Ramakrishna Engineering College, Coimbatore, India, [4]Data Scientist, Bangalore, India

Abstract

Despite the rise of Artificial Intelligence (AI) algorithms and their applications in various fields, their utilizations in high-risk fields like healthcare and finance is limited because of the lack of interpretability of their inner workings. Some algorithms are interpretable, but not accurate, whereas some produce accurate results and not decipherable. Research is underway to explore the possibilities to interrogate an AI system, and ask why it makes certain decisions. This paper aims to investigate the decision-making process by AI algorithms in the prediction of sepsis based on patients' clinical records.

We were ranked 59 in the PhysioNet/Computing in Cardiology Challenge 2019 and the utility score obtained on the full test set is 0.131, and our team name was ARUL.

1. Introduction

Sepsis is a Systemic Inflammatory Response Syndrome (SIRS) secondary to infection. When associated with acute organ failure, it leads to severe morbidity and mortality [1]. The effectiveness of antibiotic therapy rapidly decreases after the onset of sepsis. Studies show that machine learning algorithms are better than existing scoring systems for the early prediction of sepsis [2,3].

Machine learning and deep learning AI models are widely being used in computer vision, cryptography, natural language processing, anomaly detection, personalised advertisements and recommendation systems. Irrespective of so many applications, their usage in high-risk fields like healthcare is limited because of a lack of interpretability which makes patient-specific, immediate interventions difficult. Even the developers of these models cannot fully explain how these models make predictions.

Recent studies [4] in the area of interpretation of AI algorithms show that algorithms like Random Forest, eXtreme Gradient Boosting (XGB), Gradient Boosting trees, and Deep Learning models can be interpreted like linear models. Rozet et al. [5] explored machine learning models in determining the key factors for the prediction of stress in an individual. Du et al. [6] have employed explanatory techniques to interpret the prediction of cancer from radiological data. Through this paper we explored the possibilities of explaining the decisions made by the AI algorithm in the prediction of sepsis from patients' clinical data base with the help of TreeSHAP algorithm. This algorithm is used to interpret the tree-based ensemble machine learning model XGB. The dataset for this analysis was provided by PhysioNet/Computing in Cardiology Challenge 2019[6]. The data set contained forty clinical features as independent variables and the dependent variable is Sepsislabel. Each patient's covariates included demographics, vital signs, and laboratory values. The features which made the highest contribution global predictions as well as local predictions, were analyzed visually for their importance. The TreeSHAP algorithm used to extract these visuals is commonly used for post-analysis of AI-based algorithms because of its consistency [7].

2. Interpretability

The TreeSHAP algorithm was used in this study for interpreting the XGBoost AI model. The SHapley Additive exPlanations (SHAP) values calculate the impact of a feature by comparing what a model predicts with and without that feature. However, since the order in which a model looks at these features can influence its predictions, this is done in all possible sequences. If the algorithm is trained on 1000 records and 40 features, then shap values table will have the same dimensions. The mean of shap values over all the records are taken and the feature with the maximum mean scores first as important feature. Each entry for a record and feature will be different due to the interaction among the features in their respective records. So, the features work in groups and collectively describe a whole. The Shapely values which explain the interpretability of the algorithms are based on game theory [8]. Testimonials from game theory say that equitable allocation of profits leads to unique results for feature contribution methods in machine learning models. The SHAP values explain the output of these tree-based models as sum of the contributions of each feature and in every possible order.

 ISSN: 2325-887X DOI: 10.22489/CinC.2019.332

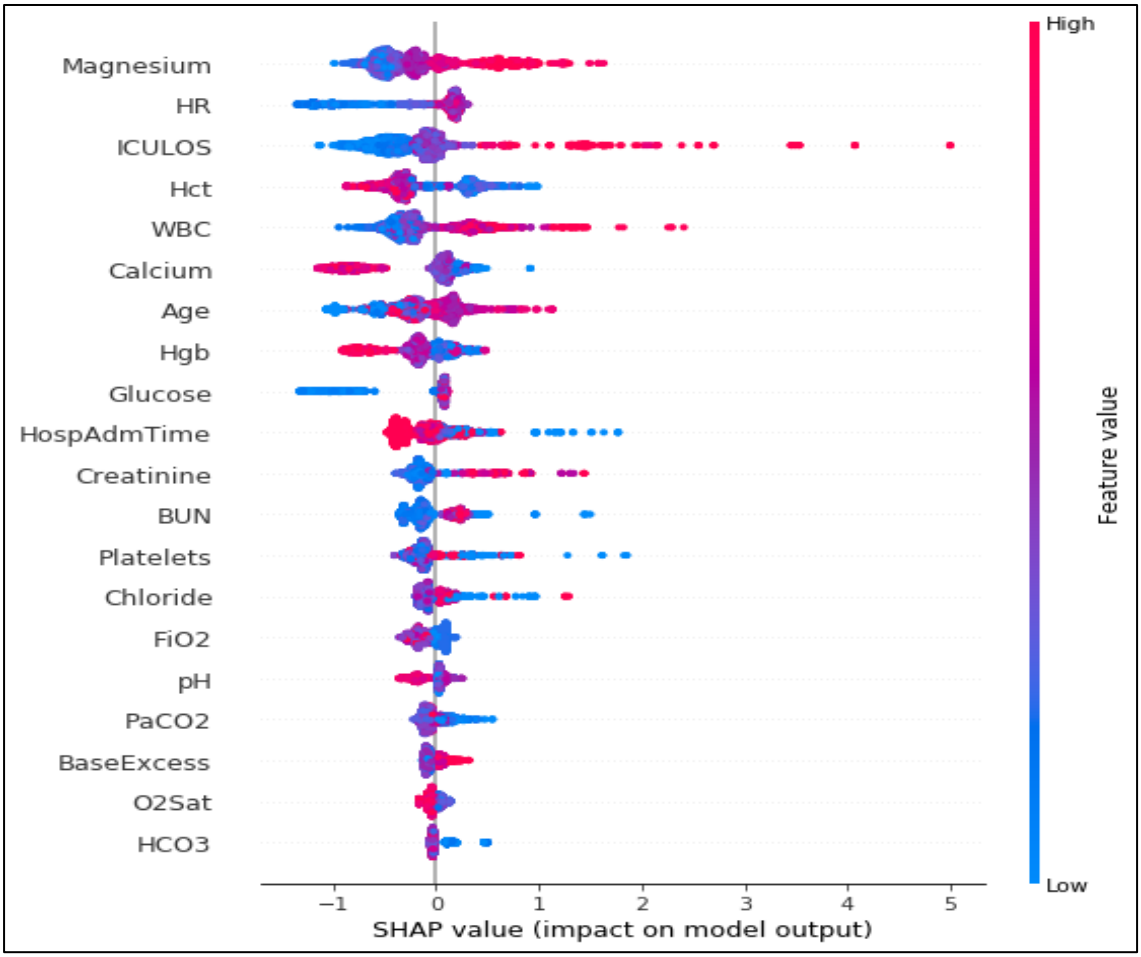

Figure 1. Summary plots show the most important features in the AI model's decision-making process. This plot shows the SHAP values for every feature (along the x-axis) in units of, log odds of prediction of sepsis. The colour of each dot implies the value of that feature. If the intensity of colour is high, the value of that feature is large; otherwise, it is low.

Interpretability of AI models can be achieved both locally and globally. Local interpretability returns comprehensive details of how an individual prediction was made with the AI algorithm. Global interpretability helps to understand the entire structure of the model.

3. Interpretability with Summary plots

The intuition behind the summary plot is that it takes all the features and shows which features propel the model's capability in predicting sepsis. Global interpretability can be obtained through summary plots.

Based on the summary plot in Figure.1, the most important feature for predicting sepsis is the serum magnesium level. The 2nd most powerful indicator of risk is the Heart Rate (HR). Basically, the summary plot tells us which features are most important. The color maps the value of a feature to the outcome in risk. Every individual is represented as one dot on each row. The x-coordinate of the dot is the impact of that feature on the model's prediction for the individual. The color of the dot represents the value of that feature for the individual. Dots that do not fit are stacked (there are 4000 records in this example). Since the XGBoost model has a logistic loss, the x-axis has units of log-odds (TreeSHAP explains the change in the margin output of the model). The features are sorted in descending order by mean shapely values. The features, BUN and platelets are not important globally but very important for a subset of individuals. The coloring by feature value shows many patterns. For example, having a higher hemoglobin (Hgb) level lowers the chances of getting sepsis. Since the shapely values of a feature is calculated by taking the average marginal contribution of the feature over all possible combinations of features, the time taken will be large to get these values, if number of records and features are very high.

4. Dependency plots

Individual variable importance or dependency plots are popular amongst statisticians for model explanations. SHAP dependence plots show the consequence of a single (or two) feature over the entire dataset. A dependence plot is a scatter plot that shows the effect of a single feature on the predictions made by the model. They plot a feature's value vs. the SHAP value of that feature across many samples. Each dot is a single prediction (row) from the dataset. The x-axis is the value of the feature. The y-axis is the SHAP value for that feature, which represents how much the feature's value changes the output of the model for that sample's prediction. The colour corresponds to a second feature that may have an interaction effect with the feature that has been plotted. The second feature is selected automatically. The interaction effect between this other feature and the feature used for plotting shows up as a distinct vertical pattern of colouring. The vertical dispersion of SHAP values at a single feature value is driven by interaction effects, and another feature can be chosen for coloring to highlight possible interactions.

In Figure 2, the dependence plot for Heart Rate is given and the feature, WBC is chosen to show interaction effects between them. SHAP value of that feature is plotted against the value of that feature for all the examples in a dataset. Since SHAP values represent a feature's contribution to a change in the model output, the plot in Figure 2 represents the change in the prediction of sepsis as HR changes. SHAP value of that feature is plotted against the value of the feature for all records. In this example, the risk of sepsis is a maximum when the heart rate (HR) lies between 100 and 140 beats/minute. Here, coloring by WBC shows that the HR has less impact on the prediction of sepsis for patients with high WBC count. Similarly, when the HR is lower and magnesium levels are lower, the possibility of sepsis is lower too.

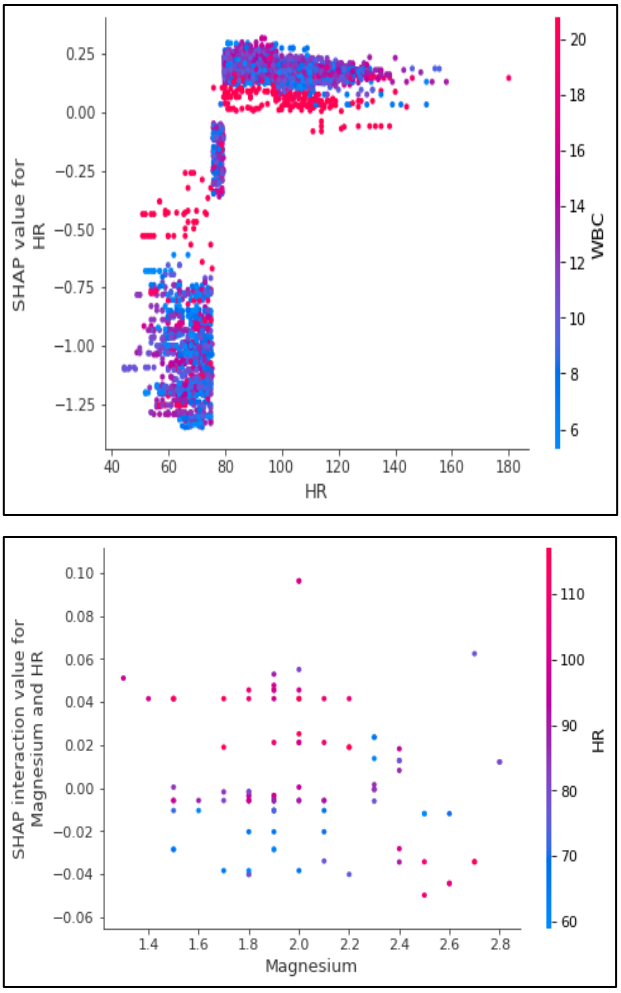

Figure 2. Dependency plots. SHAP value of that feature is plotted against the value of the feature for all records. HR has less impact on the prediction of sepsis for patients with high WBC count. The possibility of sepsis is lower when both HR and Magnesium level are low

5. Local Interpretability

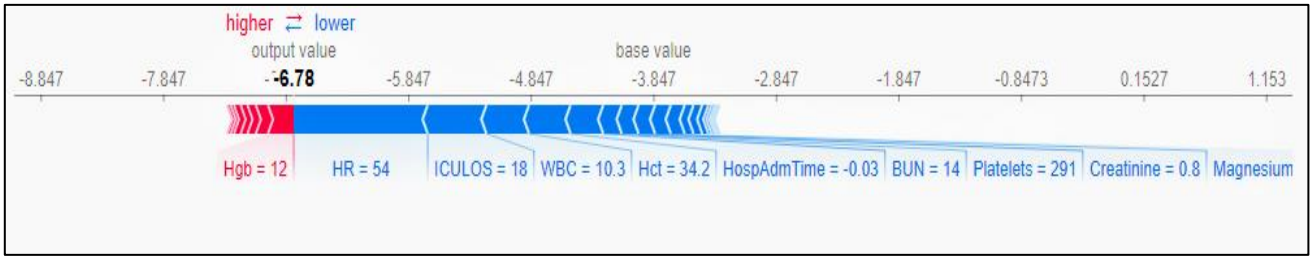

Figure 3. Force plot depicting feature contribution towards a single prediction

In this study, SHAP value of a feature for a single record prediction is obtained by the contribution of that feature towards the prediction. Figure.3, the force plot shows the features in the model, responsible to predict the patient's risk of developing sepsis and shows features contributing to pushing the model output from the base value to the actual output. Features forcing the prediction higher are shown in red, while those forcing the prediction lower are in blue.

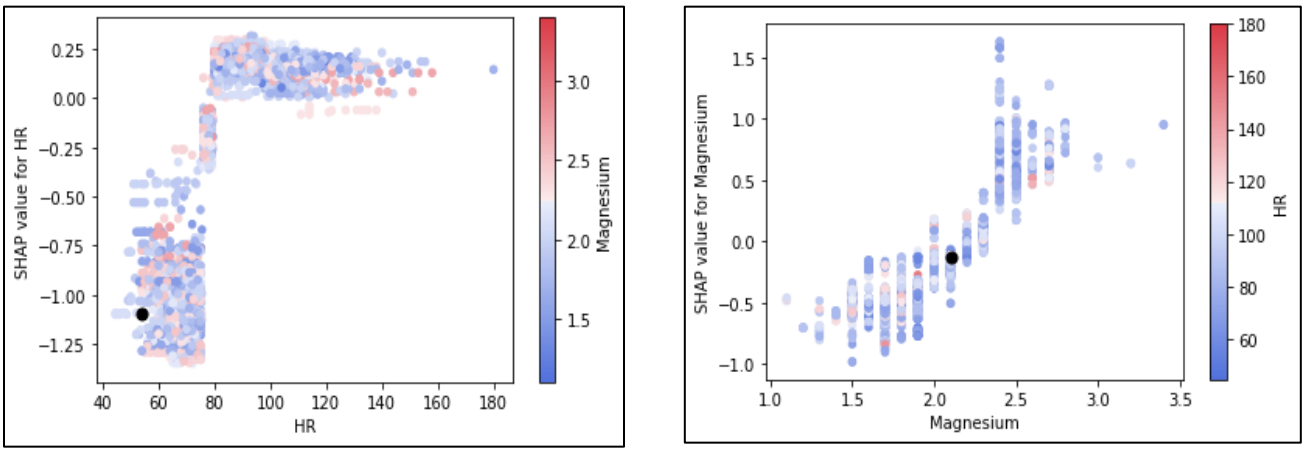

Figure 4. Global and local interpretable plots

6. Global and Local Interpretability

A record is selected and highlighted with a black circle to explain the combined effects global and local interpretability in a single graph. The following figures show the patterns for global and local interpretability. The algorithm chose serum magnesium and HR as the top two most important features for the prediction of sepsis. So, these two features are chosen to derive global and local interpretability and plotted in the Figure 4.

Conclusion

Trust is an important factor in establishing a healthy relationship between humans and machines. These results have been obtained based on 4000 records. The summary plot feature importance values change when the data set size is increased or decreased. Similarly, SHAP algorithm applied on different algorithms produce different results. Now, the interpretation of AI algorithms is made easy with the TreesSHAP algorithm. In this study, an ensemble AI model XGBoost is trained on a medical data set for the prediction of sepsis. Interesting patterns of both kinds, linear and nonlinear are discovered and analyzed for their influence in predictions through visualizations**.**

References

[1] American college of chest physicians/society of critical care medicine consensus conference: Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Crit. Care' Med. 20:864-874

[2] Mao, Q., Jay, M., Hoffman, J.L., et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open 2018;8: e017833.

[3] Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An interpretable machine learning model for accurate prediction of sepsis in the ICU. Critical Care Medicine, 46(4), 547–553.

[4] Y. Lou, R. Caruana, and J. Gehrke. Intelligible models for classification and regression. In KDD, 2012.

[5] Rozet. A., Kronish, I.M., Schwartz, J.E. Using machine learning to derive just-in-time and personalized predictors of stress: observational study bridging the gap between nomothetic and ideographic approaches. Journal of Medical Internet Research, J Med Internet Res. 2019 Apr 26;21(4): e12910

[6] Reyna, M.A., Josef, C., Jeter, R., Shashikumar, S.P., Westover, B., Nemati, S., Clifford, G.D., Sharma, A. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology 2019; Critical Care Medicine 2019; In press

[7] Lundberg, S.M., Lee, S. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 4765--4774, 2017

[8] Lipovetsky, S., Conklin, M. Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry 17.4 (2001): 319-330

Address for correspondence

Induparkavi Murugesan
Father Muller Medical College, Mangalore, India.
induparkavi@gmail.com