Explainable AI to Predict Bloodstream Infection in the Intensive Care Unit

Louisa Edwards
le7fp@virginia.edu
Department of Computer Science
University of Virginia
Charlottesville, Virginia, USA

Christopher C. Moore
ccm5u@virginia.edu
Division of Infectious Diseases and International Health
Department of Medicine
University of Virginia
Charlottesville, Virginia, USA

N. Rich Nguyen
nn4pj@virginia.edu
Department of Computer Science
University of Virginia
Charlottesville, Virginia, USA

ABSTRACT

This paper uses an explainable machine-learning method to address the challenge of diagnosing bloodstream infections (BSI), infectious diseases caused by bacterial or fungal microorganisms in the blood. These infections can lead to sepsis, a life-threatening condition, and cause increased mortality, longer hospital stays, and higher treatment costs. Central venous catheters, used extensively in intensive care units (ICU) for administering medication, fluids, and nutrition, are a primary source of BSI. Early detection of BSI is crucial for improved clinical outcomes; however, current methods using blood cultures have limitations such as long processing time, risk of contamination, and low negative predictive value. Machine learning models have been developed for early BSI detection to overcome these challenges. However, the complexity of these models often limits their utility, as their decision-making process is difficult to explain and hence hard to trust in clinical settings. To this end, we explore the concept of explainable artificial intelligence (AI) and its potential to diagnose BSI. We further present our results from applying a technique known as local interpretable model-agnostic explanations (LIME) to our best predictive models, suggesting a potential path towards creating trustworthy and understandable machine learning models for BSI detection.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Supervised learning by classification**; *Cross-validation*; • **Applied computing** → **Health informatics**.

KEYWORDS

deep learning, bloodstream infection, data mining, explainable AI

1 INTRODUCTION

Bloodstream infections are infectious diseases defined by the presence of viable bacterial or fungal microorganisms in the bloodstream, later demonstrated by the positivity of one or more blood cultures [14]. The body's response to bloodstream infection can lead to sepsis, which is a leading cause of global mortality [5]. Even for patients who do not experience sepsis, bloodstream infection is associated with increased mortality, longer hospital stays, and higher treatment costs [9]. Patients in the intensive care unit are at especially high risk of BSI because of the frequent need for catheters and their already critical condition [7].

In the ICU setting, physicians order blood cultures based on their analysis of many aspects of a patient's physiological condition, including lab results, body temperature, heart rate, and many other features. Early detection of bloodstream infection and adequate antimicrobial treatment is associated with improved clinical outcomes, especially for patients at risk for septic shock [4]. However, blood cultures take several days to process, are subject to the risk of contamination, and often have a low negative predictive value, making bloodstream infection difficult to identify clinically. To address this issue, researchers have developed predictive models that use machine learning to detect bloodstream infections.

Studies have demonstrated promising results predicting the presence of bloodstream infection [6] and identifying pathological signatures of infection [15]. However, many state-of-the-art models developed are so complex that their decision-making architecture cannot be *explained*. This lack of transparency diminishes clinician trust in the reliability of such models and discourages their use in high-risk settings. In the field of machine learning, explainability refers to the ability of a model to justify its outcomes and assist users in rationalizing its predictions [11]. Put another way, explainability is the ability to explain a model's behavior in human terms. Explainability is distinct from, but closely related to, interpretability, which refers to the ability to see and understand the inner mechanics of a model [13]. In many cases, some level of explainability can be achieved even if a machine learning model is not fully interpretable [10].

2 RELATED WORK

Many of the most powerful predictive models are so-called "black box" models that are neither interpretable nor explainable. In clinical applications, lack of explainability is a barrier to trust, and therefore to model adoption. To use a model in a decision-making process that affects their patients' health, clinicians must be able to trust, understand, and justify its predictions to themselves, their colleagues, and their patients [1]. As additional regulations are passed into law, explainability is also increasingly important for legal compliance. The General Data Protection Regulation (GDPR), a comprehensive law on data privacy in the European Union, requires companies that use algorithmic decision-making tools to provide meaningful information about the process involved [2].

2.1 Characteristics of Explainability

As machine learning continues to become more widespread in high-risk areas, demand for explainable models will continue to grow.

Before discussing what explainability means in the clinical context, we will break down the dimensions of explainable AI in its timing, scope, specificity, and target audience.

Timing. Methods for explainability can be categorized into three distinct groups based on when they aim to generate an explanation. Pre-training explainability attempts to create explainability before data is input into a model. This entails understanding and describing the data used to train a model, based on the understanding that the output of a model is largely dependent on the data it trains on. Pre-training explainability includes exploratory data analysis and feature engineering using methods such as imputation, feature splitting, and variable transformations. This type of explainability is generally embedded into the modeling process. Pre-training methods can also be used to identify and mitigate pre-training bias but are less able to explain model output. The second category of explainability methods, in-training (ante-hoc) explainability, refer to methods that integrate explainability into the structure of the model itself. Finally, post-training (post-hoc) explainability methods extract explanations to describe a trained model or prediction [12].

Scope. Explanations can generally apply to one of two scales: local or global. Local explanations apply to a specific sample or prediction, while global explanations are valid for a set of samples, or for the whole data set [12]. Local explanations are common in models in which predictions are made sample by sample, and knowledge is not stored from one prediction to the next. Global explanations are most common in models with long-term dependencies, in which knowledge is shared from one sample to another.

Specificity. Model-specific explainability methods are applicable to a certain type or architecture of model, while model-agnostic techniques are those that can be applied to any machine learning model. Pre-training techniques are generally model agnostic, while ante-hoc and post-hoc methods are often, but not exclusively, model-specific.

Target Audience. While explainable methods can be designed for a variety of audiences, the most common distinction is that between the developer, who is generally interested in technical insights that can explain an entire model, and the user, who is primarily concerned with context-driven insights that explain how a specific insight was generated. In addition to varying in scope, explainable methods designed for different audiences must align with the technical experience of the target audience [12].

2.2 Our Approach

Studies on explainability in clinical practice have found that clinicians view explainability as a means of justifying their decision-making in the context of the model's prediction [11]. Clinicians want to understand the features that lead to the model's decision - referred to in machine learning as feature importance - so that they can evaluate how the prediction aligns with the current standard of care. While metrics such as accuracy, specificity, and sensitivity are critically important, clinicians are willing to accept inaccurate predictions if they can understand why and in which contexts a model falls short. Real-world application - that is, successful prediction with real patients - is important in promoting continued use of predictive models, demonstrating the role that user interaction plays in model trustworthiness. Returning to the elements of explainable AI described above, in the context of BSI, local explanations are preferred over global ones, as clinicians are most concerned with explaining the output of the model for a specific patient. Because the target audience is the clinician rather than the developer, the focus is on a context-driven explanation rather than a technical one. With these factors in mind, we implemented LIME, an algorithm that provides local, model-agnostic post-hoc explainability that clinicians can understand [8].

Table 1: Positive, negative, and no blood culture patients

	total	negative	no-culture	positive
No. of patients	50,216	13,676	34,998	1,542
Percentage of total	100%	27.2%	69.7%	3.07%

3 METHODS

In developing models to predict BSI, we built and tested a variety of deep neural networks, including recurrent neural networks (RNNs), a class of neural networks designed to capture long-term dependencies, and convolutional neural networks (CNNs), a class of neural networks designed for image processing that use convolutional layers to capture patterns in data. All of our models take as input multi-variate time series data from the hours leading up to a blood culture and output a prediction probability representing whether the patient is positive or negative for bloodstream infection. By comparing the prediction to the ground-truth label from a patient's blood culture, our models can recognize and learn from patterns in the data. Ground truth labels also allow us to evaluate model performance on validation and test data.

3.1 Data

We used multivariate time series data sourced from the University of Virginia (UVA) Electronic Health Record. Our data includes 50,216 unique ICU patients, 363,552 unique time steps, and 38 clinically relevant features, including both lab results and vital signs. We defined an "episode" of bloodstream infection to include hourly data in the 48-hour window leading up to a blood culture. In addition to patients with positive and negative blood cultures, we included random 48-hour periods from patients who did not have a blood culture drawn. These patients served as controls and were labeled as negative, ensuring that the model's output was not conditional on the presence of a blood culture. Our response variable had two classes: a positive class containing patients with positive blood cultures, and a negative class containing patients with negative blood cultures as well as patients who did not have a blood culture drawn. See Table 1 for a breakdown of the count and percentages of positive, negative, and no blood culture patients in our dataset.

After obtaining the raw data in CSV files, we developed a preprocessing pipeline to convert it into episodes of bloodstream infection. Before doing so, we removed any patients whose blood culture contained a common contaminant. Our processing pipeline then

Table 2: Performance statistics for best CNN and GRU models

	Precision	Recall	AUROC
CNN	0.599	0.503	0.821
GRU	0.559	0.251	0.744

split data into episodes, combined them into one multidimensional dataset, and converted the result into TensorFlow Dataset format.

We used outlier cutoff values from a study by Zimmet et al [16]. Rather than imputing missing data, we masked missing values so that the model did not use them as input. This decision was based on the assumption that missing values in our dataset are not missing at random. If a patient is missing lab results for a given lab test, for example, it suggests that the clinician believed the test was unnecessary based on the patient's condition at the time.

We shuffled the combined dataset and split it into training, test, and validation sets with a 70:20:10 ratio. We used a batch size of 64 episodes for model training, testing, and validation. Cutoff values for outliers, alternative feature names, and human-readable documentation were stored in a separate file.

3.2 Adaptation of LIME

Local interpretable model-agnostic explanations (LIME) is a model-agnostic method that provides local explainability [8]. As a perturbation based approach, LIME works by perturbing data points and feeding them into the model. The model outputs are then weighed as a function of proximity to the original data. A simple, interpretable model, such as linear regression or a decision tree, is trained on the perturbed data; this model can then be interpreted, providing insight into the black box model.

4 EXPERIMENTAL RESULTS

We achieved our best model performance with convolution neural networks. Although CNNs were designed to process image data, they have also performed well on multivariate time series classification tasks. CNNs pass a weighted filter called a convolutional layer over the input data and then use a pooling layer to combine the output of the filters into encodings. By using multiple convolutional layers on top of each other, CNNs can recognize complex patterns in data.

We also experimented with gated recurrent networks (GRUs) [3], a form of recurrent neural network that uses a gating mechanism to mitigate the vanishing gradient problem that occurs in standard RNNs. By maintaining hidden states and allowing previous outputs to be used as inputs, RNNs can capture long-term dependencies in time series data.

4.1 Model Performance

From an explainability perspective, RNNs are intuitively favorable given the time series nature of our data. RNNs are designed to model sequential information. In contrast, CNNs are designed to process visual information. Because CNNs interpret data as images, columns of data next to each other are interpreted to be similar to each other. While this is true in the case of image pixels, the order of columns in time series data is not generally ordered this way, unless it has been feature engineered to do so. CNNs can offer powerful performance, but RNNs offer the potential for more explainability in the context of time series data. We implemented LIME on our best GRU and CNN models. Model statistics are shown in Table 2, and results from LIME are discussed in detail in the following section.

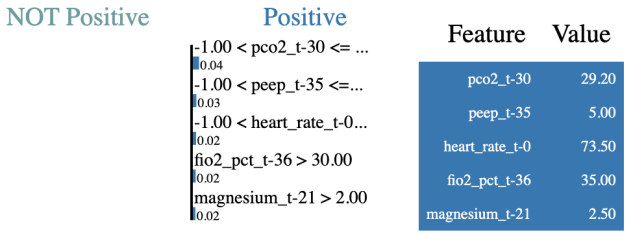


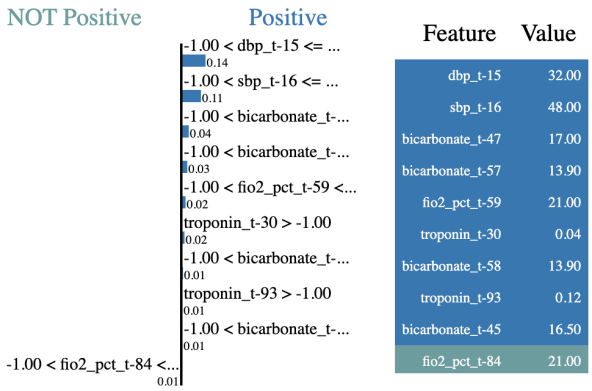
Figure 1: LIME output for a positive patient. Our CNN correctly predicted positive with probability of 1.0.

4.2 Clinical Explainability

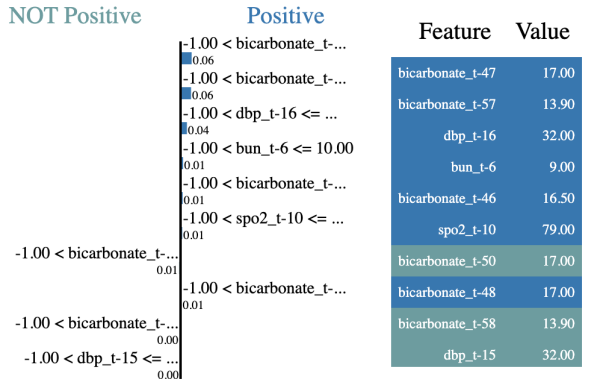
Figure 1 shows the output of LIME for a patient that our CNN correctly predicted as positive. The features listed are ranked in order of importance (i.e. how important they were to the model's decision); the corresponding values of each feature are listed in the value column. For example, the first row of the chart lists pc02_t_30 and 29.30, indicating that the most important feature to the model's prediction was partial pressure of carbon dioxide at time step 30, which had a value of 29.30. While LIME can be run on any number of features, the output in this example has been limited to the five features to improve readability and to isolate the most important features.

The features on the "Positive" side of the chart contribute to a prediction that the patient is positive, while the features on the "NOT Positive" side contribute to a negative prediction. In this case, all five of the most important features indicated a positive prediction; the model predicted correctly positive with a probability of 1.0. When provided in conjunction with the model's prediction, this type of output allows doctors to better understand what went into the model's decisions. By isolating the most important features and their values, it also allows doctors to evaluate whether the model may be picking up on a specific condition that the patient is already known to have (ex: kidney failure, elevated blood pressure, etc.) or a bloodstream infection. Ultimately, this information allows for improved trust in the model, better insight into when the model may be incorrect, and better clinical decision-making.

Figure 2 compares the output of LIME from our CNN and GRU on the same positive patient. While the outputs share many of the same features (dbp, bicarbonate, and tronponin), the list and order of features is different, indicating that the models made their predictions differently. Most notably, the CNN correctly predicted positive with a probability of 1.0, while the GRU predicted positive with a probability of 0.43. Given the performance difference between these models, it is not surprising that the CNN had a more accurate prediction.



(a) LIME output for our CNN on another positive patient. The model correctly predicted positive, with a probability of 1.0.



(b) LIME output for GRU on the same positive patient. The model incorrectly predicted negative, with a positive probability of 0.43.

Figure 2: LIME output for another positive patient.

4.3 Missing Data

Missing data is a serious concern in our dataset. Regardless of how good a model is, it will perform poorly if there is not sufficient data for a given patient. The same is true of LIME.

Figure 3 displays LIME output for our CNN from yet another positive patient. All of the values in the chart are -1, indicating that they are missing and have been masked. LIME interprets the masking as actual values and is unable to provide a relevant prediction. In a clinical setting, predictive model would only be used if sufficient data on the patient was present. Nonetheless, this feedback can be valuable: seeing missing values listed as the most important values to the prediction suggests that the model does not have sufficient data and that the prediction is unreliable. Ultimately, it underscores the importance of data quality to model performance and explanability.

5 CONCLUSION

Building on existing research, our work demonstrates that deep learning has the potential to predict bloodstream infection in at-risk ICU patients correctly. Moreover, post-hoc explainability methods such as LIME can provide local explainability, giving clinicians additional information about why a prediction was made, which physiological features the model focuses on, and whether the model can be trusted.

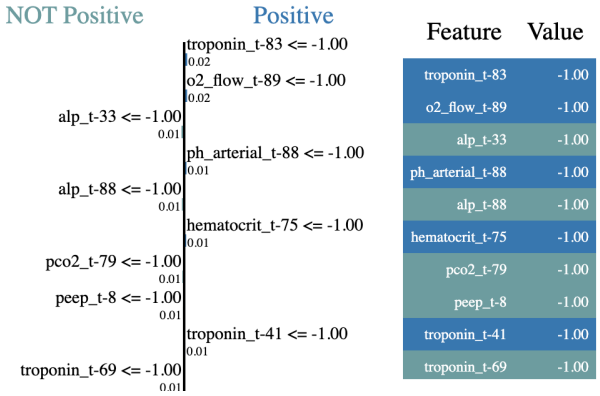


Figure 3: LIME output for a positive patient with lots of missing data. Our CNN incorrectly predicted negative, with a positive probability of only 0.07.

ACKNOWLEDGMENTS

This work was supported by a University of Virginia Global Infectious Diseases Institute Seed Grant and a University of Virginia Center for Engineering in Medicine Seed Grant Program, which were both awarded jointly to CCM and NRN.

REFERENCES

[1] Chaddad A. 2023. Survey of Explainable AI Techniques in Healthcare. *Sensors (Basel)* 23, 2 (2023), 634.
[2] Selbst A. 2017. Meaningful information and the right to explanation. *International Data Privacy Law* 7, 4 (2017), 233–242.
[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
[4] Timsit JF. et al. 2020. Bloodstream infections in critically ill patients: an expert statement. *Intensive Care Med* 46, 2 (2020), 266–284.
[5] Singer M. et al. 2016. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 315, 28 (2016), 801–810.
[6] K.C. Pai et al. 2021. An Artificial Intelligence Approach to Bloodstream Infections Prediction. *Journal of Clinical Medicine. 2021* 10, 13 (2021), 2901.
[7] Gahlot R. et al. 2014. Catheter-related bloodstream infections. *International Journal of Critical Illness Injury Science* 4, 2 (2014), 162–167.
[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[9] Kristina E. Rudd et al. 2020. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet* 395, 10219 (2020), 200–211.
[10] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1 (2019), 206–215.
[11] Tonekaboni S. et al. 2019. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *Cornell University arXiv* (2019).
[12] Rojat T. et al. 2021. Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey. *Cornell University arXiv* (2021).
[13] Vishwarupe V. et al. 2022. Explainable AI and Interpretable Machine Learning: A Case Study in Perspective. *Procedia Computer Science* 204 (2022), 869–876.
[14] C. Viscoli. 2016. Bloodstream Infections: The peak of the iceberg. *Virulence* 7, 3 (2016), 248–251.
[15] Zoabi Y. et al. 2021. Predicting bloodstream infection outcome using machine Learning. *Sci Rep* 11 (2021).
[16] Alex N. Zimmet et al. 2020. Pathophysiologic Signatures of Bloodstream Infection in Critically Ill Adults. *Critical care explorations* 2, 10 (2020).

A LIST OF FEATURES

Feature abbreviations and corresponding names:

age: Age
albumin: Albumin
alp: Alkaline Phosphate
alt: Alanine Transaminase
ast: Aspartame Aminotransferase
bicarbonate: Bicarbonate
bun: Blood Urea Nitrogen
calcium: Calcium
chloride: Chloride
co2: Carbon Dioxide
creatinine: Creatinine
dbp: Diastolic Blood Pressure
fio2_pct: Fraction of Inspired Oxygen
glucose: Glucose
hematocrit: Hematocrit
hemoglobin: Hemoglobin
heart_rate: Heart Rate
lactic_acid: Lactic Acid
magnesium: Magnesium
o2_flow: Oxygen Flow Rate
pco2: Partial Pressure of Carbon Dioxide
peep: Positive End-expiratory Pressure
ph_arterial: Arterial Blood Gas
phosphorus: Phosphorus
po2: Partial Pressure of Oxygen
potassium: Potassium
protime_inr: Prothrombin Time
ptt: Partial Thromboplastin Time
platelet_count: Platelet Count
resp_rate: Respiratory Rate
sbp: Systolic Blood Pressure
sodium: Sodium
spo2: Oxygen Saturation
temp: Core Body Temperature
total_bilirubin: Total Bilirubin
total_protein: Total Protein
troponin: Troponin
wbc: White Blood Cell Count