ARTICLE

SBC-SHAP: Increasing the Accessibility and Interpretability of Machine Learning Algorithms for Sepsis Prediction

Daniel Walke,[a,b,*] Daniel Steinbach,[c,d] Thorsten Kaiser,[e] Alexander Schönhuth,[f] Gunter Saake,[b] David Broneske,[g,†] and Robert Heyer[f,h,†]

Background: Sepsis is a life-threatening condition that is one of the major causes of death worldwide. Early detection of sepsis is required for fast initialization of an appropriate therapy. Complete blood count data containing information about white blood cells, platelets, hemoglobin, red blood cells, and mean corpuscular volume could serve as early indicators. However, previous approaches are limited by their interpretability (i.e., investigating the influence of feature values on individual predictions) and accessibility (i.e., easy accessibility for clinicians without programming expertise).

Methods: We developed a graph-based approach for training machine learning (ML) algorithms to incorporate time-series information for prediction based on complete blood count data. Additionally, we investigated the effect of integrating different ratios to a healthy reference measurement to improve the performance of the previously published ML model. Finally, we developed a web application based on our approaches to increase accessibility.

Results: While it was irrelevant how exactly the ratio was formed, our approach increased the sensitivity at 80% specificity across all ML models from up to 78.2% to up to 82.9% on an internal dataset (i.e., same tertiary care center) and from 65.4% to 73.4% on an external dataset (i.e., independent tertiary care center) for prospective time-series analysis. Additionally, we propose SBC-SHAP (https://mdoa-tools.bi.denbi.de/sbc-shap), a web application that visualizes the sepsis risks and individual interpretations of several ML classifiers.

Conclusions: We are confident that this tool will increase the interpretability and accessibility of ML models for predicting sepsis based on complete blood count data. SBC-SHAP is open-sourced on https://github.com/danielwalke/sbc_app.

[a]Bioprocess Engineering, Otto von Guericke University, Magdeburg, Germany; [b]Database and Software Engineering Group, Otto von Guericke University, Magdeburg, Germany; [c]Institute of Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics, Leipzig University Hospital, Leipzig, Germany; [d]Medical Informatics Center—Department for Clinical AI and Translational Medicine, University of Leipzig Medical Center, Leipzig, Germany; [e]University Institute for Laboratory Medicine, Microbiology and Clinical Pathobiochemistry, OWL University Hospital of Bielefeld University, Detmold, Germany; [f]Faculty of Technology, Bielefeld University, Bielefeld, Germany; [g]German Center for Higher Education Research and Science Studies (DZHW), Hannover, Germany; [h]Multidimensional Omics Analyses Group, Leibniz-Institut für Analytische Wissenschaften—ISAS—e.V., Dortmund, Germany.

*Address correspondence to this author at: Bioprocess Engineering, Otto von Guericke University, Universitätsplatz 2, Magdeburg 39106, Germany. E-mail daniel.walke@ovgu.de.

[†]Shared last authors.

Received January 20, 2025; accepted June 02, 2025.

https://doi.org/10.1093/jalm/jfaf091

IMPACT STATEMENT

We highly increased the interpretability and accessibility of machine learning classifiers for predicting sepsis based on complete blood count data to enable faster detection of sepsis in the clinical routine.

INTRODUCTION

Sepsis is one of the major causes of death worldwide (1). An early detection of this life-threatening condition is important for an appropriate treatment with antibiotics (2). The inflammatory response is mediated by the release of cytokines from neutrophil granulocytes and macrophages (3). Complete blood count data such as white blood cell (WBC) count, red blood cell count (RBC), platelet count (PLT), hemoglobin (HGB), and mean corpuscular volume (MCV) may serve as readily accessible indicators for sepsis (4). Steinbach et al. proposed a machine learning model (RUSBoost) based on complete blood count data to achieve an area under the receiver operating curve (AUROC) of 0.872 on an internal (i.e., hold-out dataset from the same tertiary care center) test and 0.805 on an external (i.e., separate independent tertiary care center) validation dataset (2). Then, Walke et al. extended this work by using graph neural networks (GNNs) to incorporate time-series information (5). Therefore, the time series of each patient is represented as a graph structure. The graph structure for a single patient contains for each measurement a node (attached node features: age, sex, and complete blood count data), and nodes are connected by edges based on their measurement time (Fig. 1A–C). Specifically, we connected all measurements of a patient with all previous measurements (i.e., graph structure based on patient history). Using GNNs to learn on these time-series graphs, they achieved an AUROC of up to 0.900 for prospective sepsis classification (AUROC: 0.834 on external data) and up to 0.958 for retrospective analysis (AUROC: 0.952 on external data) (5). Although these studies showed promising results for sepsis analysis solely with complete blood count data, they have 2 main limitations.

Limitation 1: Interpretability

Most machine learning models, especially deep learning models like GNNs, struggle with providing in-depth interpretability for individual predictions that can indicate how specific feature values (e.g., WBC counts) contribute to the prediction (e.g., increased or decreased risk). While there exist frameworks [e.g., SHapley Additive exPlanations (SHAP) (6)] for increasing the interpretability of ensemble-based machine learning algorithms like RUSBoost, these approaches cannot be applied on a feature level to GNNs that achieved better predictions (5). Existing frameworks for increasing interpretability of GNNs only focus on feature importance (e.g., overall importance of WBC counts) (7, 8) or structural importance (e.g., contribution of other measurements in a time series) (9, 10) but do not explain how impactful a feature's value is for the risk.

Limitation 2: Accessibility

Using trained machine learning models requires programming knowledge, model expertise, and sufficient time for setting up the pipelines (11). Therefore, clinicians and physicians often cannot use these models in their clinical routines. While there exist some applications for sepsis prediction, they are either not open-sourced (12), do not provide a user-friendly web interface (13), or require many input parameters (14). Additionally, we have not found a tool that provides visualizations for the local interpretability of obtained

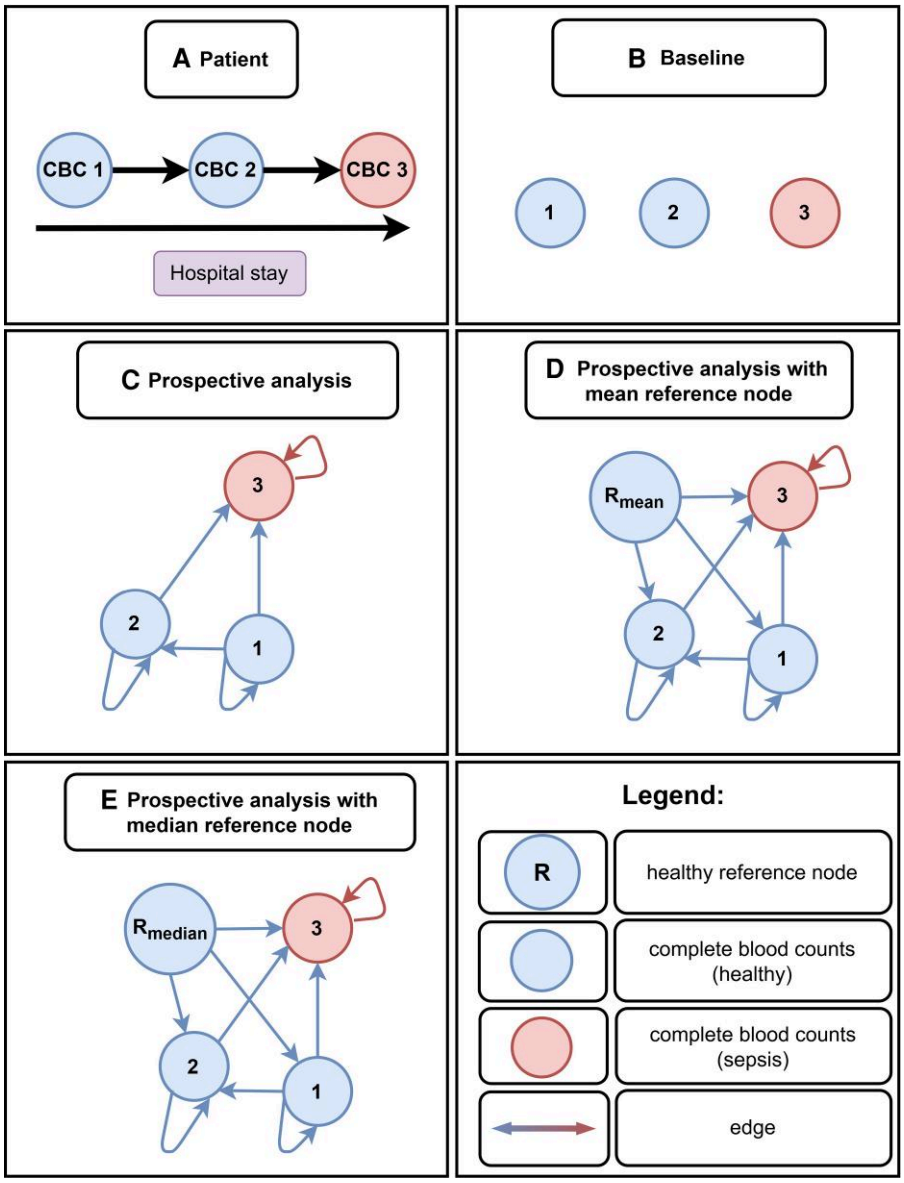

Fig. 1. Graphs for time-series analysis on complete blood count data (5). Each patient can have one or multiple complete blood count measurements (nodes) (A). These measurements contain information about the patient's age, sex, and measured HBG, RBCs, WBCs, MCV, and PLT. Additionally, each measurement is labeled as "control" (blue node) or "sepsis" (red node). In the simplest case, machine learning classifiers can be trained on these node measurements independently from each other (B). Additionally, nodes can also be connected based on their measurement times (i.e., each measurement of a patient is connected to all previous measurements) to perform a prospective analysis; i.e., nodes can analyze previous measurements (C). However, some nodes do not have time-series information (i.e., there is only one measurement available). Therefore, we added a reference node based on the mean (D) or median (E) of control complete blood count measurements. This reference node should serve as a comparison for the current measurement to a healthy reference measurement.

sepsis predictions, which further complicates the applicability in a clinical context.

In this study, we propose a technique for increasing the sensitivity of machine learning classifiers for the prediction of sepsis based on complete blood count data. This technique allows the inspection and visualization of contributions of specific feature values on individual predictions (limitation 1). Thereby, we achieve full compatibility of our graph machine learning approach with the SHAP framework to return highly interpretable predictions. Additionally, we evaluate strategies to introduce a reference node (i.e., a reference to a physiologically complete blood count measurement in each patient's history) as an additional comparison to a healthy control measurement (Fig. 1D and E). Finally, we propose SBC-SHAP, an easily accessible web application for the classification of sepsis based on complete blood count data (limitation 2). As input, SBC-SHAP only requires a unique patient identifier, age, sex, complete blood count data, and their measurement times. It will then return the sepsis risk based on input data and will visualize the contribution of specific feature values to the sepsis risk [SHAP values (6)]. Additionally, SBC-SHAP provides predicted sepsis risks based on multiple machine learning models [i.e., XGBoost (15), random forest (16), decision tree (17), and logistic regression (18)]. We are not aware of any other models that only require complete blood count information while being interpretable and easily accessible via the browser (https://mdoa-tools.bi.denbi.de/sbc-shap).

MATERIALS AND METHODS

First, we introduce details regarding the accessibility and deployment of our application in the "Deployment" section. Then, we introduce the steps for our graph-based approach and machine learning training ("Graph-Based Feature Engineering and Training"). This simplification allows us to get highly interpretable results for individual predictions. Finally, we introduce the implementation of our web application, SBC-SHAP ("Web Application for Increased Accessibility").

Deployment

Users can easily access the web application via https://mdoa-tools.bi.denbi.de/sbc-shap and only need to submit information about the patients' age, sex, an anonymous identifier, measurement time, and complete blood count values (HBG, WBCs, RBCs, MCV, and PLT). We do not store or write any information from requests on our servers. Additionally, we provide an easy setup for local installations. Therefore, users only need to install and start Docker Desktop (https://www.docker.com/products/docker-desktop/), download our docker compose file (https://github.com/danielwalke/sbc_app/blob/main/docker-compose.yml), navigate into the directory of the downloaded "docker-compose.yml" file, and run "docker-compose up -d" in the terminal. Then, the application can be accessed under "localhost:3000/sbc-shap" in a browser. Note that the application startup requires some time (a couple of minutes). A complete tutorial for SBC-SHAP and its local setup is provided in our GitHub repository (https://github.com/danielwalke/sbc_app) and on the web application itself (https://mdoa-tools.bi.denbi.de/sbc-shap).

Graph-Based Feature Engineering and Training

Use Cases. We define 3 use cases for the sepsis classification based on complete blood count data. First (baseline), we set up common machine learning classifiers that do not incorporate any time-series information as baseline (Fig. 1B). As a second use case (prospective analysis), we incorporate time-series information by constructing a directed graph to generate new features and use these features to train machine learning classifiers (Fig. 1C). The third use case (prospective analysis

with reference node) builds upon the second use case by adding an additional reference node [mean (Fig. 1D) or median (Fig. 1E) of complete blood count measurements].

Preprocessing and Feature Engineering. We used the dataset from Steinbach et al. (2) available at https://doi.org/10.5281/zenodo.10781419 (19) for the training and evaluation. Details about the dataset are listed in the Supplemental Material (Supplemental Note 1). Preprocessing and construction of graphs for time-series analyses are based on Steinbach et al. (2) and Walke et al. (5), respectively (Fig. 1A). Features containing time-series information were incorporated by appending new features describing the difference (diff.) to the mean of all previous measurements (Fig. 2B). For the third use case, we additionally investigate using the quotient (quot.) of complete blood count measurements to the mean of all previous

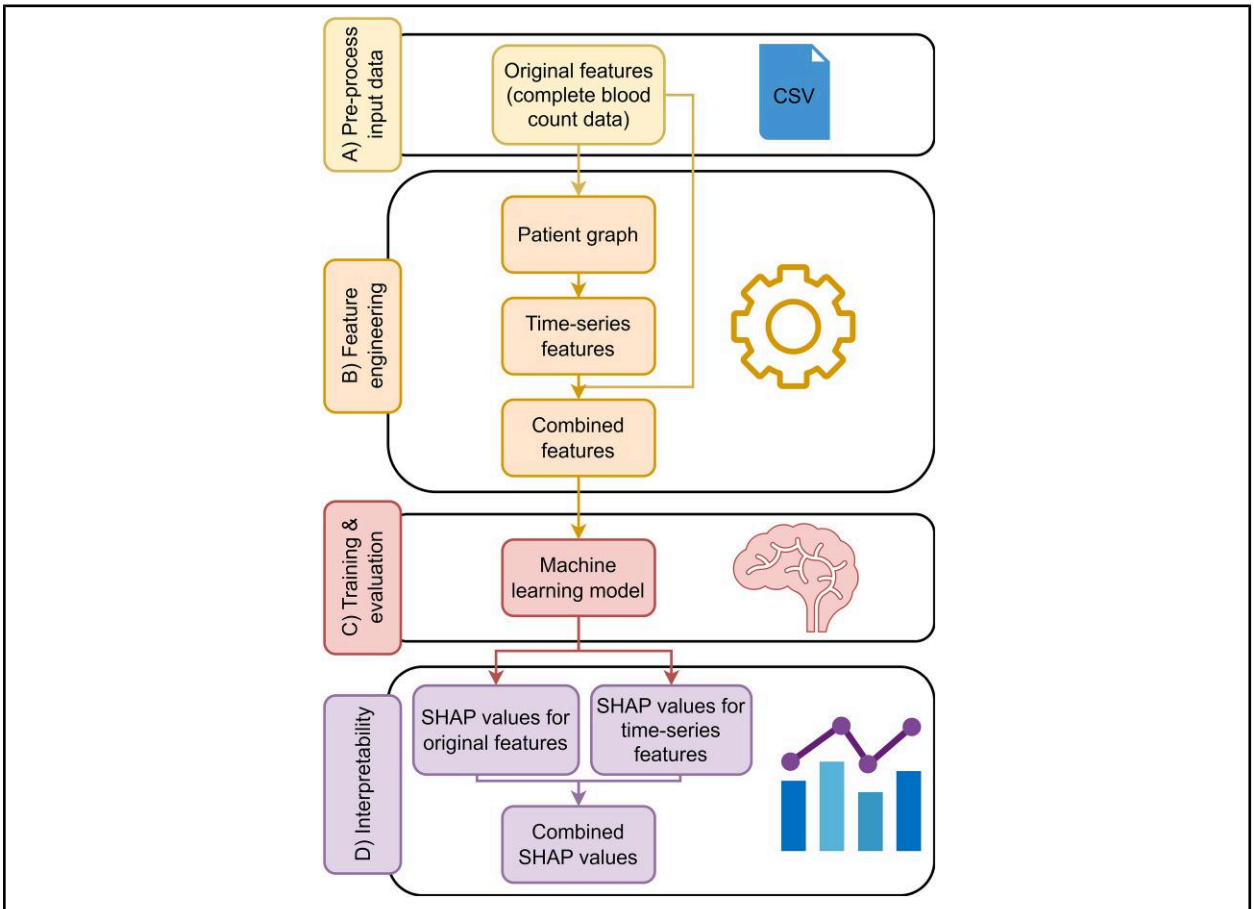

Fig. 2. Overview of the overall workflow. (A), First original data containing information about the patient's complete blood count data is preprocessed according to Steinbach et al. (2); (B), Then, data is transformed into patient graphs according to Walke et al. (5) based on time-series information. Time-series information is generated based on this patient graph using a feature engineering technique (i.e., neighborhood aggregation based on connected complete blood count measurements). Resulting time-series features are combined with the original features; (C), The combined features are passed on to a machine learning model for training and are evaluated based on AUROC and sensitivity at 80% specificity; (D), Finally, we used SHAP (6) to increase the interpretability of the trained model.

measurements including the reference node (calculated by the mean and median of all control measurements in the training set). These newly generated features from the time series are concatenated with the original features and used as a new, transformed input for a machine learning classifier (Fig. 2C).

Training and Evaluation. As machine learning classifiers, we used XGBoost (15), random forest (16), decision tree (17), and logistic regression to evaluate our approach across diverse machine learning algorithms. We determined the AUROC across all classifiers for consistency with the work from Steinbach et al. (2) and Walke et al. (5). Additionally, we assessed the sensitivity at 80% specificity to precisely show the implications for clinicians. Since there is always a trade-off between specificity and sensitivity, we set a fixed threshold of 80% for the specificity to guarantee a consistent threshold across all models while ensuring a relatively low false positive rate (20%). Further details regarding training and tuning are listed in the Supplemental Material (Supplemental Note 2).

Interpretability. Finally, we applied the SHAP framework (6) on each classifier to receive highly interpretable results that show the contribution (i.e., SHAP values) of specific feature values to the predictions (i.e., sepsis risk). A positive SHAP value indicates a shift of the prediction toward sepsis and vice versa. The returned SHAP values were then separated into 2 parts: one part for the influences of feature values from the current measurement (i.e., original features) and the other part for the influences of feature values from time-series information. The sum of both SHAP values is used to investigate the combined (overall) influence (Fig. 2D). For a global interpretation of the model, we averaged the absolute SHAP values over all samples in the internal dataset for the original features and time-series features. We used SHAP to understand and visualize how our trained machine learning models make specific decisions. Additionally, SHAP values can show by their magnitude which features have higher or lower impacts on the predicted sepsis risk. Besides SHAP, we visualized partial dependence plots to investigate the influence of a feature on the predictions across the entire dataset (see Supplemental Note 5).

Web Application for Increased Accessibility

We decided to implement a web application because it does not require a download, which might represent another usage barrier. While the web application is publicly available (https://mdoa-tools.bi.denbi.de/sbc-shap), it is also possible to set up a local instance to protect privacy requirements ("Deployment"). The web application consists of a frontend and a backend (Fig. 3). The frontend aims to visualize all results including features, sepsis risks, and SHAP values as interpretation mechanisms. The backend aims to make all computationally expensive computations (i.e., the machine learning predictions and calculation of SHAP values). Additional details regarding the implementation of the frontend and backend are listed in Supplemental Note 3. The complete source code is open-sourced at https://github.com/danielwalke/sbc_app/tree/main.

RESULTS

In this section, we describe the results of our approach for classifying sepsis based on time-series data while retrieving interpretable results ("Simplification for Increased Interpretability"). Additionally, we show the different functions of our proposed web application, called SBC-SHAP, to increase accessibility ("Web Application for Increased Accessibility").

Simplification for Increased Interpretability

Machine Learning Evaluation. First, we show the results of our approach for analyzing time-series data. We represented each time series of a patient

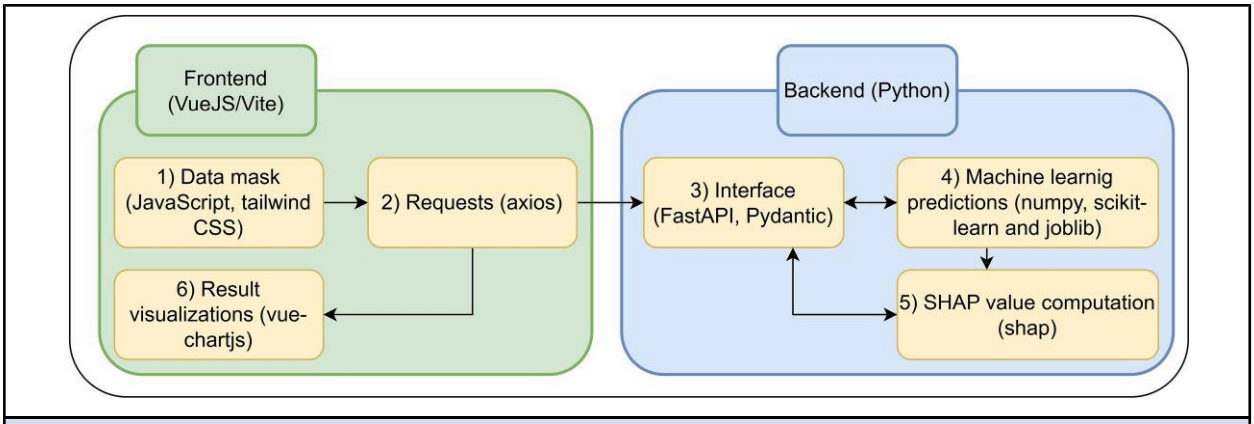

Fig. 3. Architecture of the SBC app. The SBC app consists of a frontend (green) and a backend (blue). First data is imported on the frontend via file import or the data input mask (1). Then, data is transferred with axios (2) to our FAST-API backend via HTTPs (3). Data is passed to the machine learning classifiers to make predictions (4) and then passed to the SHAP explainer to calculate the SHAP values (5). After the computations of sepsis risks and SHAP values, the results are transferred back to the frontend via FAST-API and axios for the visualization in the frontend (6).

as a graph to connect the complete blood count measurements with previous measurements. Additionally, we integrated a reference node (mean or median of all healthy complete blood count measurements in the training set) in each time series. To incorporate the time-series feature, we either used the diff. or the quot. to previous time-series measurements (with and without reference node) as aggregation. As a baseline, we trained the same machine learning algorithms without incorporating time-series information. The classification performance (Fig. 4) improved on machine learning algorithms using our prospective approach without a reference node (internal AUROC up to 0.889, external AUROC up to 0.839) compared to the baseline models (internal AUROC up to 0.875, external AUROC up to 0.818). Our approach achieved a sensitivity at 80% specificity of up to 81.4% and 72.1%, while the baseline only achieved up to 78.2% and 65.4% on the internal and external dataset, respectively. The integration of a reference node (mean node) further increased the classification performance (internal AUROC up to 0.896, external AUROC up to 0.841) when using the difference to all previous time-series measurements of this patient case. This further increased the sensitivity at 80% specificity to up to 82.9% and 73.7% on the internal and external dataset, respectively. However, the comparison of different reference nodes (mean or median) and aggregation strategies (diff. or quot.) revealed only slight differences in the classification performance, indicating that it is less important which reference nodes or aggregation strategy are used.

Interpretability. Afterwards, we applied the SHAP framework (6) to retrieve the contributions of specific features to predictions. We only compare the combined SHAP values across different approaches (baseline, our approach, and our approach with reference nodes) for the XGBoost classifier (Fig. 5) because it achieved the highest classification performance on the internal dataset (Fig. 4). Higher mean absolute SHAP values indicate a higher importance, and vice versa. A detailed overview of all other SHAP value plots is listed in Supplemental Figs. 1–3. While WBCs, HGB, and PLT were the most important features for the baseline model, age has become one of the 3 most important features instead of HBG for

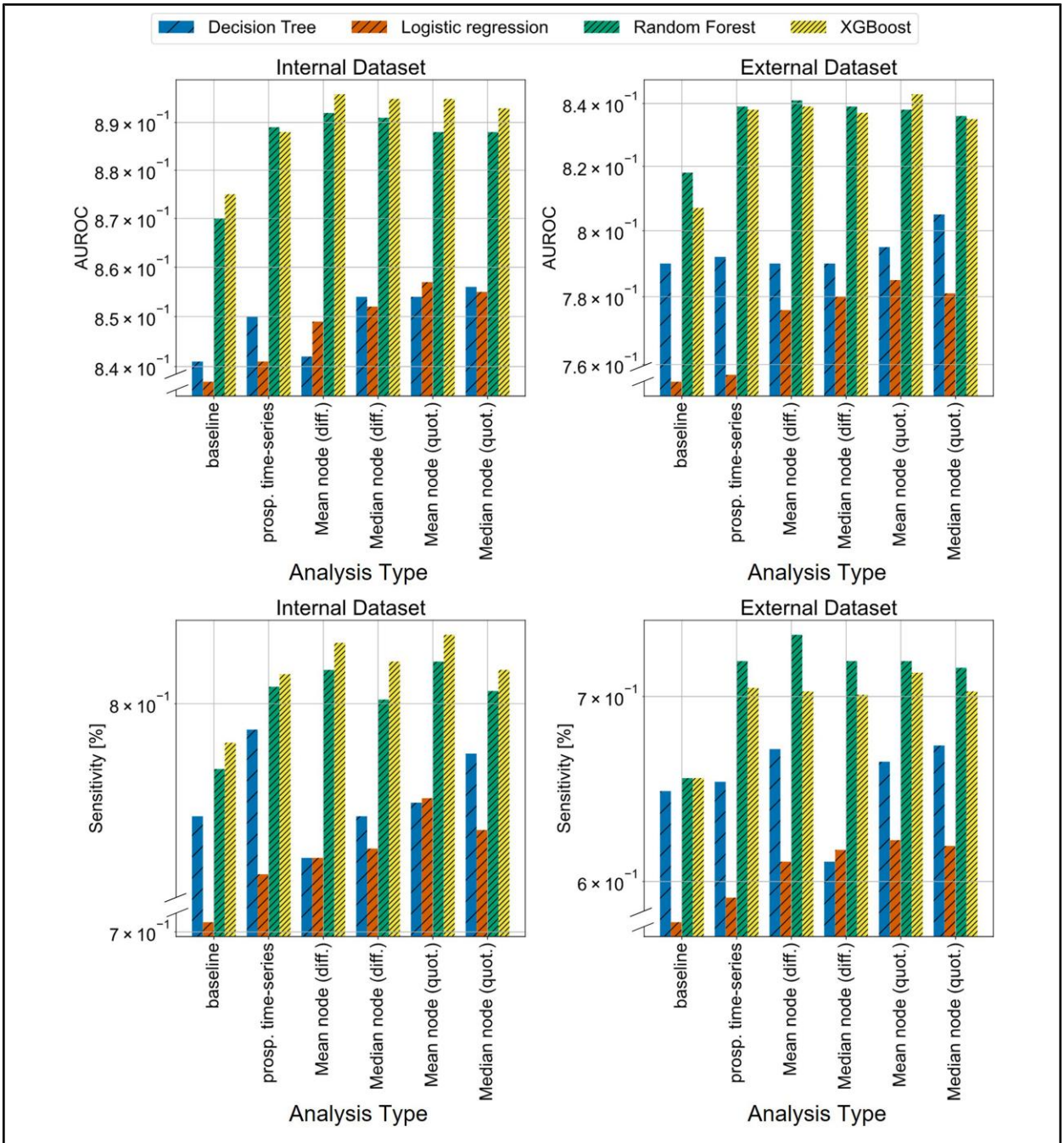

Fig. 4. **Results (AUROC and sensitivity at 80% specificity) of our approaches and the baseline models. As test data, we used an internal (Leipzig) and external (Greifswald) dataset. We compared the classification performance across logistic regression, decision tree, random forest, and XGBoost with the baseline models compared to our approaches using only time-series information from previous measurements (prosp. time series) and using a reference node (mean or median of control measurements) with different aggregation strategies, i.e., diff. or quot. to previous measurements. Higher values indicate a better classification performance.**

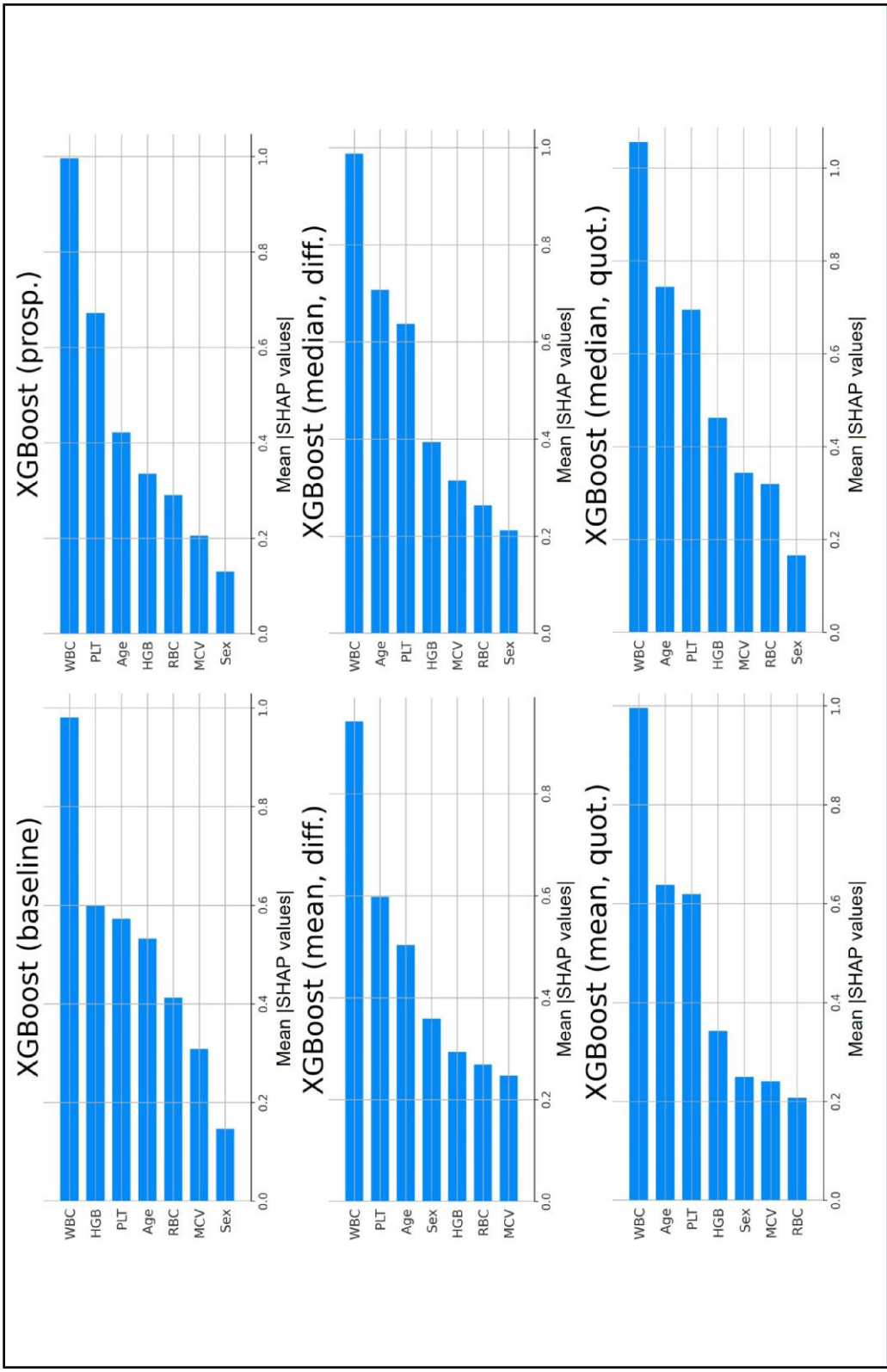

Fig. 5. Mean absolute SHAP values for measuring the overall feature importance across all approaches for the XGBoost algorithm. Higher values indicate a higher importance and vice versa. We visualized the feature importance of WBCs, HGB, PLT, RBCs, MCV, age, and sex. While the baseline model mainly used the features WBC, HGB, and PLT, the other approaches that incorporate time-series information used mainly WBC, PLT, and age for the prediction.

the models using our approach. However, across all approaches, WBCs remained the most important feature for the analysis. Additionally, the different reference nodes (mean or median) and aggregation strategies (difference or quotient) lead to slightly changed feature importance across the less important features (RBCs, MCV, and sex). This change compared to the baseline models indicates that slight modifications in the classification mechanisms can result in much higher classification performance in the models. Therefore, it is especially important to investigate the influence on individual (local) predictions to get highly interpretable results ("Web Application for Increased Accessibility").

Web Application for Increased Accessibility

We developed an open-source (https://github.com/danielwalke/sbc_app/) web application (SBC-SHAP) available at https://mdoa-tools.bi.denbi.de/sbc-shap for increasing the accessibility of our proposed machine learning models. It provides the classification of complete blood count data (Fig. 6) using the algorithms proposed in "Simplification for Increased Interpretability." Users need to input a unique patient identifier, age, sex, complete blood count information, and the measurement time (Fig. 6A). Then, they can decide whether they want to do a prospective analysis or a retrospective analysis (Fig. 5B) and submit their data (Fig. 6C). As a result, the sepsis risk as a value between 0% and 100% is returned based on the classifiers' confidence (see Supplemental Note 4 for further details) in the prediction (Fig. 6D). SHAP values are visualized after clicking on the magnifier symbol (Fig. 6E) and show the contribution of each feature value to the predicted sepsis risk (Fig. 6F). SHAP values visualize the influence of individual feature values (e.g., high WBC count) on the prediction (i.e., higher or lower sepsis risk) for improving the local interpretability of those black-box models. For example (Fig. 6F), a high age shifts the classification to "sepsis" (positive SHAP value visualized as a red bar) and physiological WBC counts to "control" (negative SHAP value visualized as a blue bar). Thereby, users can gain a better understanding of the underlying predictions and can learn from the algorithm. We are convinced that good explainability and visualization reduce concerns and enable practitioners to use the algorithms optimally for clinical decision-making and to learn the more targeted interpretation of machine learning models in laboratory medicine. Furthermore, users can also filter results based on their input data (e.g., patients older than 50) and their results (e.g., sepsis risk above 80%) to easily investigate the dataset. This filtering might help in detecting patterns between input data and the returned sepsis risks.

The default classification algorithm for all samples is XGBoost because it achieved the highest classification performance among all investigated models (Fig. 4). Additionally, users can investigate the predictions and SHAP values of other classifiers (logistic regression, decision tree, and XGBoost) by clicking on the list symbol (Fig. 6G). Thereby, they can compare the results across different classifiers to investigate different strengths and weaknesses among the trained machine learning models.

DISCUSSION

In this section, we discuss our results and show existing limitations of our approaches for improving interpretability ("Improved Interpretability") and accessibility ("Improved Accessibility"). Finally, we introduce 2 example use cases to improve the understanding of our tool and the associated SHAP values ("Example Use Cases").

Improved Interpretability

We developed a graph-based feature engineering approach to incorporate time-series information of complete blood count data. This approach increased the AUROC from 0.875 on the internal dataset and 0.818 on the external dataset to up to 0.896 and 0.841, respectively. Additionally, this

Fig. 6. Web application SBC-SHAP for increasing the usability and interpretability of machine learning algorithms for sepsis classification. (A), First, users input complete blood count data from patients into the respective fields; (B), Then, they can choose a prediction type, i.e., either prospective or retrospective analysis; (C), Next, they can submit their data to the trained random forest; (D), Sepsis risks for each complete blood count measurement are returned; (E), By clicking on the details icon on the right (magnifier symbol), the influences of individual features (F) to the prediction (SHAP values) are visualized as colored bar charts (blue indicates a direction toward the prediction "control" and red indicates a direction toward the prediction "sepsis"); (G), Finally, the predictions of other models and their SHAP values can be investigated by clicking on the list icon on the right.

approach even outperformed GNNs from previous work with an AUROC of up to 0.890 on the internal and 0.820 on the external dataset (5). Besides the improved classification performance, our approach is compatible with the SHAP framework, which enables local interpretability of predicted sepsis risks to investigate how specific feature values contribute to the predicted sepsis risk. While GNNs also provide some interpretability mechanisms, they only investigate how specific neighboring nodes (10) or subgraph structures (9) influence the prediction or investigate the importance of specific nodes, edges, or features (7). Therefore, it is not possible to investigate how a specific feature value contributes to the prediction similar to the SHAP framework (6).

We also evaluated our approach with different machine learning models (logistic regression, decision tree, random forest, and XGBoost) and achieved a superior performance across all models. However, in general, ensemble-based models like random forests (AUROC up to 0.892) and XGBoost (AUROC up to 0.896) achieve a better classification performance compared to more shallow machine learning models like decision trees (AUROC up to 0.856) and logistic regressions (AUROC up to 0.857). Since it is difficult to investigate feature contributions on predictions across an entire dataset, we plotted the partial dependence of features for all proposed models (see Supplemental Note 5, Supplemental Fig. 5). While across all models similar trends emerge in how features contribute to the sepsis risk (e.g., an increase in WBC count increases the sepsis risk), the different models differ in their decision boundaries (i.e., how they predict sepsis), which can lead to divergent results (i.e., predicted sepsis risks) across the same complete blood count measurement.

However, our approach still has some limitations. First, further information [e.g., body temperature (20), procalcitonin (21) or C-reactive protein concentrations (22)] could further increase the sensitivity of machine learning classifiers. However, those features must be readily accessible through routine tests to avoid introducing bias from clinicians' suspicions when ordering specialized diagnostics. While SHAP provides intuitive explanations for predictions and enables users to investigate how specific feature values contribute to the prediction, it also has some limitations. For example, it assumes feature independence, which can distort attributions since features are often correlated in the real world (e.g., correlation between HBG and RBCs). Additionally, global interpretations to investigate how a specific feature influences the prediction across the entire dataset is not possible, which is why we integrated partial dependence plots in our analysis (see Supplemental Note 5). In most cases, sepsis labels appear toward the end of each time series, introducing a bias in the dataset that may be exploited by the underlying machine learning models. Currently, our approach is limited to the prediction of a sepsis risk. The incorporation of further syndromes and diseases could further enhance the applicability of our approach. Finally, it might be interesting to investigate further preprocessing steps like clustering to further improve the classification performance (23).

Improved Accessibility

We successfully developed a user-friendly web application for predicting sepsis solely based on complete blood count data that is easily accessible under https://mdoa-tools.bi.denbi.de/sbc-shap. The application provides diverse functionalities including sorting, filtering, multiple classification algorithms, customizable sensitivity, and visualization of SHAP values for increasing local interpretability. Nevertheless, SBC-SHAP has some limitations. First, it is not clear whether displayed sepsis risks and the visualization of SHAP values are intuitive and helpful for clinicians. A survey regarding the usability of SBC-SHAP might be required as an evaluation. Furthermore, it remains uncertain how additional features (e.g.,

procalcitonin) could be integrated without significantly increasing the system's complexity and reducing the responsiveness of the web application.

Example Use Cases

As an example of identifying sepsis cases, a clinician might have measured complete blood count information of a patient (ID: 584590 in test data) during their hospitalization (Supplemental Fig. 6A). Then, the clinician only needs to enter "https://mdoa-tools.bi.denbi.de/sbc-shap" in a browser, add complete blood count information in the input fields (or import information as CSV), and submit their data. The first measurement (age: 86, sex: male, HBG: 7.5 mmol/L, WBC: 8.1 Gpt/L, RBCs: 4.4 Tpt/L, MCV: 80.5 fl, PLT: 200 Gpt/L) indicates a low sepsis risk with 13.8%. However, after the second measurement, HGB (6.6 mmol/L), RBC (3.77 Tpt/L), and PLT (148 Gpt/L) decreased and MCV (81.7 fl) and WBC (10.2 Gpt/L) increased, which correctly increased the sepsis risk to 50.2%. The main reasons for the increased sepsis risk are the increased SHAP values for PLT (SHAP value increased from −0.14 to +0.80) and WBC (−1.48 to 0.30).

As an example of correcting values with prior knowledge, a clinician might have measured the complete blood count information of a patient (ID: 583818, age: 81 years, sex: male, HGB: 3.7 mmol/L, WBC: 8 Gpt/L, RBC: 2.58 Tpt/L, MCV: 72.9 fl, PLT: 299 Gpt/L). According to SBC-SHAP (Supplemental Fig. 6B), the predicted sepsis risk would be over 50%, mainly due to the low HGB (high positive SHAP values of 1.4) and RBC values (high positive SHAP values of 0.47). However, the clinician might have some additional prior information (e.g., a bleeding event might have reduced the HGB and RBC level). After correcting the values to physiological values [e.g., HGB: 8 mmol/L (24), RBC: 5 Gpt/L (25)] by changing the respective input fields, the predicted sepsis risk decreases to 2.4% (Supplemental Fig. 6C).

CONCLUSION

We developed a new graph-based machine learning approach for the classification of sepsis based on complete blood count data. Our approach increased the sensitivity at 80% specificity of machine learning models for sepsis classification from up to 78.2% to 82.9% on the internal dataset (i.e., same tertiary care center) and from up to 65.4% to 73.4% on the external (i.e., independent tertiary care center) validation dataset. Additionally, our approach enables graph machine learning to be fully compatible with the SHAP framework (6) to make our predictions highly interpretable. We found slight changes in the overall feature importance across our approaches indicating that even slight modifications can still substantially increase the classification performance. Finally, we provide an open-source, easily accessible web application, called SBC-SHAP (https://mdoa-tools.bi.denbi.de/sbc-shap). Users can investigate sepsis risks for individual complete blood count measurements and can investigate how specific feature values contribute to predicted sepsis risks. Additionally, filtering options allow the filtering of data with and without results for diverse use cases (e.g., filtering based on sepsis risks). Furthermore, predicted sepsis risks can be compared across different machine learning models (logistic regression, decision tree, random forest, XGBoost) to investigate the strengths and weaknesses of individual machine learning models. We are confident that this tool sets the foundation for the development of a more generic user interface that enables the evaluation and exploration of all kinds of machine learning models independent of the use case (e.g., integration of more features in new machine learning models or machine learning models from other domains).

SUPPLEMENTAL MATERIAL

Supplemental material is available at *The Journal of Applied Laboratory Medicine* online.

Nonstandard Abbreviations: WBC, white blood cell; RBC, red blood cell; HGB, hemoglobin; AUROC, area under the receiver operating characteristic curve; GNN, graph neural networks; SHAP, SHapley Additive exPlanations; PLT, platelets; MCV, mean corpuscular volume.

Author Contributions: *The corresponding author takes full responsibility that all authors on this publication have met the following required criteria of eligibility for authorship: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved. Nobody who qualifies for authorship has been omitted from the list.*

Daniel Walke (Conceptualization-Equal, Formal analysis-Equal, Investigation-Lead, Methodology-Equal, Software-Lead, Validation-Equal, Visualization-Lead, Writing—original draft-Lead, Writing—review & editing-Equal), Daniel Steinbach (Conceptualization-Equal, Data curation-Lead, Methodology-Supporting, Software-Supporting, Visualization-Supporting, Writing—review & editing-Equal), Thorsten Kaiser (Conceptualization-Equal, Data curation-Equal, Methodology-Supporting, Writing—review & editing-Equal), Alexander Schönhuth (Conceptualization-Equal, Visualization-Supporting, Writing—review & editing-Equal), Gunter Saake (Funding acquisition-Equal, Project administration-Equal, Supervision-Supporting, Writing—review & editing-Equal), David Broneske (Conceptualization-Equal, Methodology-Supporting, Project administration-Equal, Supervision-Equal, Validation-Supporting, Visualization-Supporting, Writing—review & editing-Equal), and Robert Heyer (Conceptualization-Equal, Funding acquisition-Equal, Methodology-Supporting, Project administration-Equal, Supervision-Equal, Validation-Supporting, Visualization-Supporting, Writing—review & editing-Equal)

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form*.

Research Funding: This research was funded by the German Research Foundation under the project "Optimizing Graph Databases Focusing on Data Processing and Integration of Machine Learning for Large Clinical and Biological Datasets" (project number 463414123; grant numbers HE 8077/2-1, SA 465/53-1).

Disclosures: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

Acknowledgments: We would like to thank Kay Schallert (Multidimensional Omics Analyses Group, Leibniz-Institut für Analytische Wissenschaften—ISAS—e.V., Bunsen-Kirchhoff-Straße 11, 44139 Dortmund, kay.schallert@isas.de) for his valuable support in the deployment process, which greatly facilitates the accessibility of our tool.

REFERENCES

1. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease study. Lancet 2020;395:200–11.
2. Steinbach D, Ahrens PC, Schmidt M, Federbusch M, Heuft L, Lübbert C, et al. Applying machine learning to blood count data predicts sepsis with ICU admission. Clin Chem 2024;70:506–15.
3. Delano MJ, Ward PA. The immune system's role in sepsis progression, resolution, and long-term outcome. Immunol Rev 2016;274:330–53.
4. Agnello L, Giglio RV, Bivona G, Scazzone C, Gambino CM, Iacona A, et al. The value of a complete blood count (CBC) for sepsis diagnosis and prognosis. Diagnostics (Basel) 2021;11:1881.
5. Walke D, Steinbach D, Gibb S, Kaiser T, Saake G, Ahrens P, et al. Edges are all you need: potential of medical time series analysis with graph neural networks. PLoS ONE. Forthcoming.
6. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. 2017. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach (CA): Curran Associates Inc. p. 4768–77. https://dl.acm.org/doi/10.5555/3295222.3295230 (Accessed June 2025).
7. Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. 2019. GNNExplainer: generating explanations for graph neural networks. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EA, Garnett R, editors. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019; Vancouver (Canada): Curran Associates, Inc. p. 9240–51. https://proceedings.neurips.cc/paper_files/paper/2019/file/d80b7040b773199015de6d3b4293c8ff-Paper.pdf (Accessed June 2025).
8. Luo D, Cheng W, Xu D, Yu W, Zong B, Chen H, Zhang X. 2020. Parameterized explainer for graph neural network.

In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems 33: NeurIPS 2020; Virtual: Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/hash/e37b08dd3015330dcbb5d6663667b8b8-Abstract.html (Accessed June 2025).

9. Perotti A, Bajardi P, Bonchi F, Panisson A. GRAPHSHAP: explaining identity-aware graph classifiers through the language of motifs. In: 2023 International Joint Conference on Neural Networks (IJCNN); Gold Coast (Australia): IJCNN. p. 1–8. https://ieeexplore.ieee.org/document/10191053 (Accessed June 2025).
10. Akkas S, Azad A. 2024. GNNShap: scalable and accurate GNN explanation using shapley values. In: Proceedings of the ACM Web Conference 2024; Singapore (Singapore): Association for Computing Machinery. p. 827–38. https://dl.acm.org/doi/10.1145/3589334.3645599 (Accessed June 2025).
11. Crankshaw D, Sela G-E, Zumar C, Mo X, Gonzalez JE, Stoica I, et al. 2020. InferLine: latency-aware provisioning and scaling for prediction serving pipelines. In: SoCC '20: Proceedings of the 11th ACM Symposium on Cloud Computing; Virtual: Association for Computing Machinery. p. 477–91. https://dl.acm.org/doi/abs/10.1145/3419111.3421285 (Accessed June 2025).
12. Whalen K. PowerPoint presentation. https://dihi.org/wp-content/uploads/2020/02/Sepsis-Watch-One-Pager.pdf (Accessed December 2024).
13. GitHub. ikoghoemmanuell/sepsis-prediction-with-ML-and-FastAPI. 2024. https://github.com/ikoghoemmanuell/Sepsis-Prediction-with-ML-and-FastAPI?tab=readme-ov-file (Accessed December 2024).
14. GitHub. abhishek-parashar/sepsis-prediction. 2024. https://github.com/abhishek-parashar/sepsis-prediction?tab=readme-ov-file (Accessed December 2024).
15. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco (CA): Association for Computing Machinery. p. 785–94. https://dl.acm.org/doi/10.1145/2939672.2939785 (Accessed June 2025).
16. Ho TK. Random decision forests. 1995. In: Proceedings of 3rd International Conference on Document Analysis and Recognition; Montreal (Canada): Institute of Electrical and Electronics Engineers. p. 278–82. https://ieeexplore.ieee.org/document/598994 (Accessed June 2025).
17. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. 1st Ed. New York (NY): Chapman and Hall/CRC; 2017. https://doi.org/10.1201/9781315139470 (Accessed June 2025).
18. Cox DR. The regression analysis of binary sequences. J R Stat Soc Series B Stat Methodol 1958;20:215–32.
19. Gibb S, Ahrens P, Steinbach D, Schmidt M, Kaiser T. sbcdata: laboratory diagnostics from septic and non-septic patients used in the AMPEL project. 2024. https://doi.org/10.5281/zenodo.10781419 (Accessed June 2025).
20. Drewry AM, Fuller BM, Bailey TC, Hotchkiss RS. Body temperature patterns as a predictor of hospital-acquired sepsis in afebrile adult intensive care unit patients: a case-control study. Crit Care 2013;17:R200.
21. Wacker C, Prkno A, Brunkhorst FM, Schlattmann P. Procalcitonin as a diagnostic marker for sepsis: a systematic review and meta-analysis. Lancet Infect Dis 2013;13:426–35.
22. Póvoa P, Almeida E, Moreira P, Fernandes A, Mealha R, Aragão A, et al. C-reactive protein as an indicator of sepsis. Intensive Care Med 1998;24:1052–6.
23. Mondal R, Ignatova E, Walke D, Broneske D, Saake G, Heyer R. Clustering graph data: the roadmap to spectral techniques. Discov Artif Intell 2024;4:7.
24. Nebe T, Bentzien F, Bruegel M, Fiedler GM, Gutensohn K, Heimpel H, et al. Multizentrische ermittlung von referenzbereichen für parameter des maschinellen blutbildes/multicentric determination of reference ranges for automated blood counts. LaboratoriumsMedizin 2011;35:3–28.
25. Gulati GL, Hyun BH. The automated CBC. A current perspective. Hematol Oncol Clin North Am 1994;8:593–603.