INTRODUCTION
Bloodstream infections (BSI) and associated sepsis are the leading causes of death worldwide. Prompt treatment with targeted antibiotics affects both the financial impact and the clinical outcome of bloodstream infection. However, the current standard-of-care, which depends on blood culture-based diagnosis, is often unable to provide such a fast response. Several novel methods allowing rapid microbial identification (RMI) are employed in Tel Aviv Sourasky Medical Center (TASMC) and in hospitals worldwide.

In the current project we took another step forward and set out to utilize machine learning tools in order to identify the bacteremic patients whose condition is more likely to deteriorate. Identifying these patients at an early stage, will allow the clinical team to prioritize their treatment, and select the ones that will undergo RMI instead of the traditional lengthy identification routines.

The model relies on extracting features from different modalities available in Electronic Health Records (EHRs). Using these EHRs we constructed a machine learning model that predicted patient mortality with high accuracy. Our framework also provides ranking of the most important features that were used to define the decisions. This interpretation will enable us to build an explainable scoring system that can be utilized by clinicians.

MATERIALS & DATA RESOURCES
Medical records were collected by the Tel Aviv Sourasky Medical Center (TASMC) data center, using EHRs from patients admitted to TASMC between 2013-2020. Medical records include EHRs, demographic features, quantitative clinical measurements and previous medical history in TASMC.

METHODS
This study uses a supervised LightGBM classifier as the predictor model. LightGBM is a high-performance machine learning algorithm that benefits from great interpretability potential due to its recursive tree-based decision system. In contrast, internal model mechanisms of black-box modelling strategies are typically difficult to interpret.

To identify the principal features driving model prediction, SHAP (SHapley Additive exPlanations) values were calculated. These values are suited for complex models such as artificial neural networks and gradient-boosting machines.

The performance models were evaluated by assessing the classification accuracy (ratio of true predictions over all predictions), the precision, sensitivity/recall and F1 scores, and area under the Receiver Operating Characteristic Curve (auROC).

RESULTS
The results show that the model can accurately identify the outcome of patients, regardless of their original diagnosis upon hospital admission. Notably, the performance of the external test set is also accurate, which suggests that the model captures the key biomarkers of patient mortality. Our framework also provides ranking of the most important features that were used to define the decisions. As expected, age was the most predicting feature. Measurements such as red blood cell distribution width (RDW), albumin and creatinine levels and were also key features in predicting mortality. This interpretation will enable us to build an explainable scoring system that can be utilized by clinicians.

Figure 1. Training and testing data sets distribution data filtering process. a. Shows data distribution over time. b. A diagram showing the initial dataset and different filtering processes.

Figure 2. ROC curves of the predictive model. The blue line reflects training and testing via cross-validation. The orange line reflects testing the model on the prospective dataset.

Figure 3. SHapley Additive exPlanations (SHAP) beeswarm plot for mortality prediction showing SHAP values for the most important features of the model. Features in the summary plots (y-axis) are organized by their mean absolute SHAP values. Each point corresponds to an individual person in the study. The points position on the x-axis shows the impact that feature has on the classifier’s prediction for a given individual. Values of those features (i.e. age) are represented by their color.