Budget Constrained Machine Learning for Early Prediction of Adverse Outcomes for COVID-19 Patients

Sam Nguyen
Lawrence Livermore National Laboratory
Ryan Chan
Lawrence Livermore National Laboratory
Jose Cadena
Lawrence Livermore National Laboratory
Braden Soper
Lawrence Livermore National Laboratory
Paul Kiszka
ProMedica Health System
Lucas Womack
ProMedica Health System
Mark Work
ProMedica Health System
Joan Duggan
University of Toledo College of Medicine and Life Sciences
Steven Haller
University of Toledo College of Medicine and Life Sciences
Jennifer Hanrahan
University of Toledo College of Medicine and Life Sciences
David kennedy
University of Toledo College of Medicine and Life Sciences
Deepa Mukundan
University of Toledo College of Medicine and Life Sciences
Priyadip Ray (ray34@llnl.gov)
Lawrence Livermore National Laboratory

Research Article

Keywords: COVID-19, machine learning, outcomes

Posted Date: June 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-593801/v1

Abstract

Background: Machine learning (ML) based risk stratification models of Electronic Health records (EHR) data may help to optimize treatment of COVID-19 patients, but are often limited by their lack of clinical interpretability and cost of laboratory tests. We develop a ML based tool for predicting adverse outcomes based on EHR data to optimize clinical utility under a given cost structure. This cohort study was performed using deidentified EHR data from COVID-19 patients from ProMedica Healthcare in northwest Ohio and southeastern Michigan.

Methods: We tested performance of various ML approaches for predicting either increasing ventilatory support or mortality and the set of model features under a budget constraint was optimized via exhaustive search across all combinations of features.

Results: The optimal sets of features for predicting ventilation under any budget constraint included demographics and comorbidities (DCM), basic metabolic panel (BMP), D-dimer, lactate dehydrogenase (LDH), erythrocyte sedimentation rate (ESR), CRP, brain natriuretic peptide (BNP), and procalcitonin and for mortality included DCM, BMP, complete blood count, D-dimer, LDH, CRP, BNP, procalcitonin and ferritin.

Conclusions: This study presents a quick, accurate and cost-effective method to evaluate risk of deterioration for patients with SARS-CoV-2 infection at the time of clinical evaluation.

Full Text

This preprint is available for download as a PDF.

Tables

Please see the supplementary files section to view the tables.

Figures

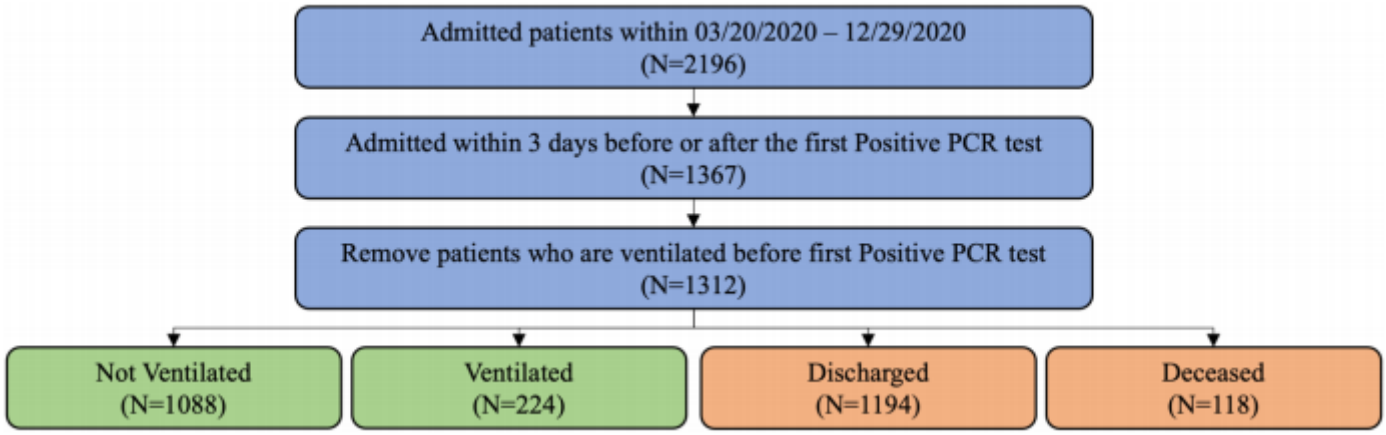

Figure 1

A flowchart of study enrollment. Composite ventilation related outcomes are represented by green boxes and Mortality related outcomes are represented by red boxes.

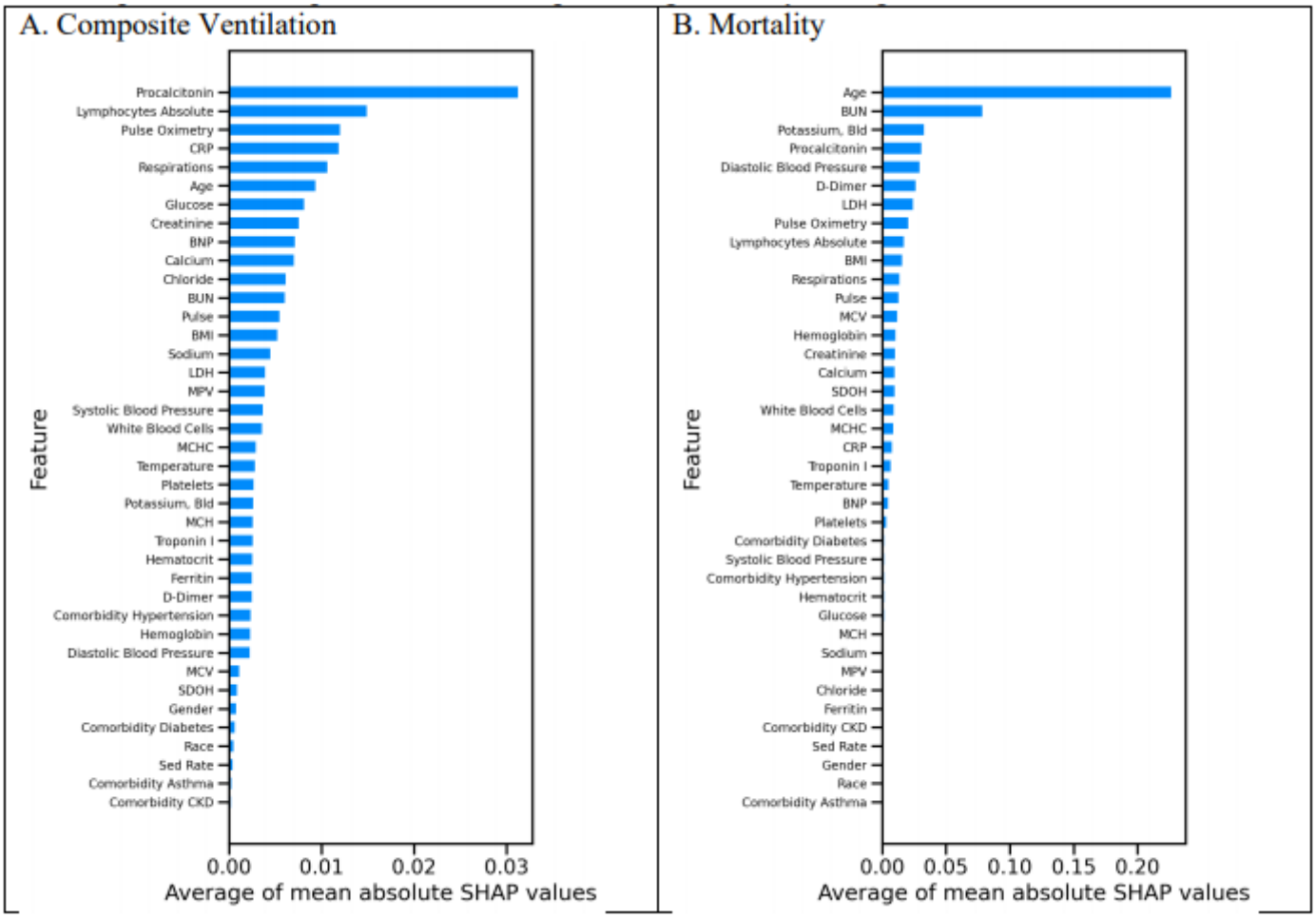

Figure 2

Feature importance of a trained XGBoost model. The higher the absolute SHAP value, the greater the contribution of the feature to the predicted outcome. The SHAP values are averaged over 5 folds of test splits that span the whole dataset. A. The top three most important features for predicting ventilation are Procalcitonin, Lymphocytes Absolute and Pulse Oximetry. B. The top three most important features for predicting mortality are Age, BUN and Potassium.

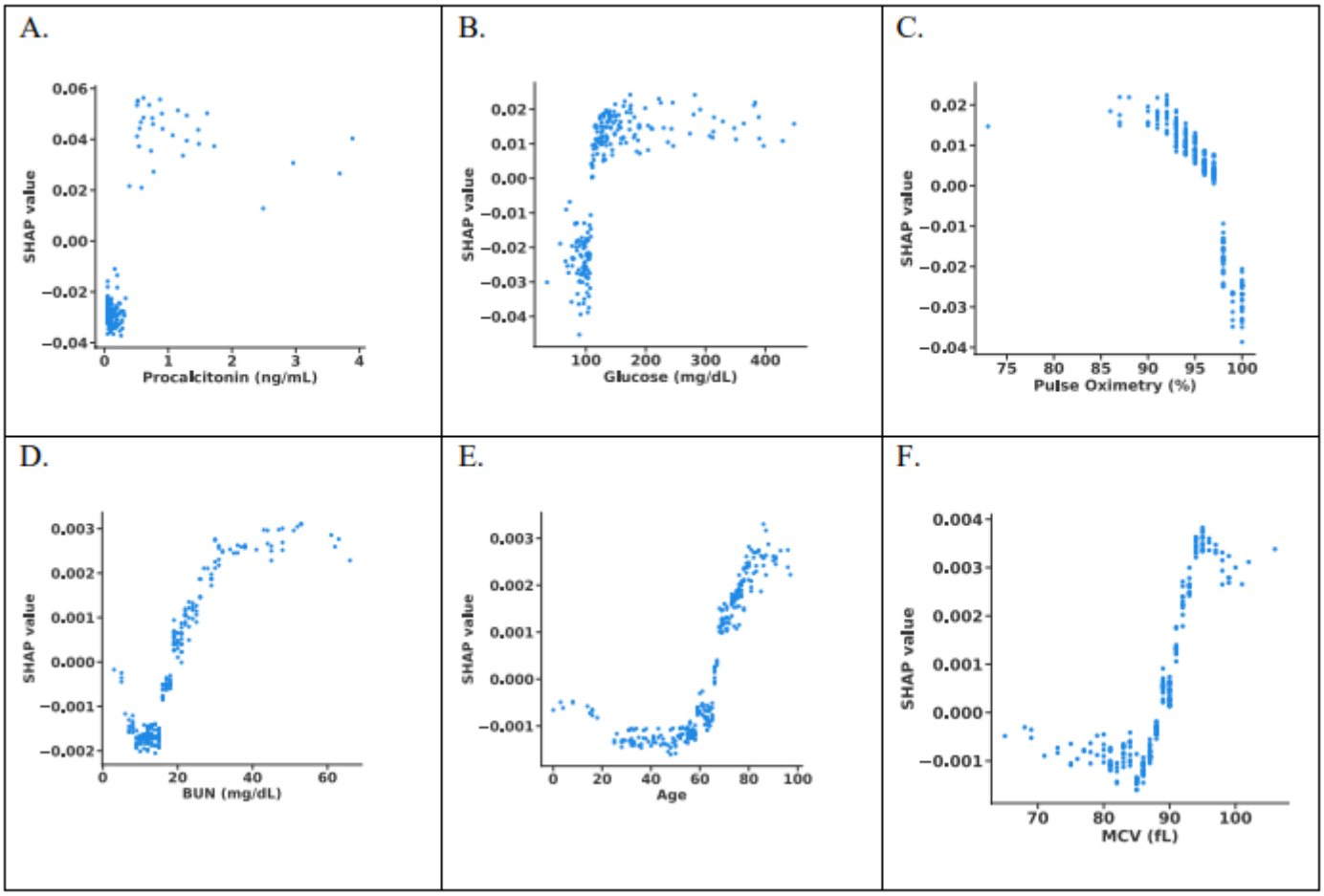

Figure 3

Scatter plots of SHAP values versus unnormalized values for selected features. A-C. The top three most significant features for predicting composite ventilation. D-F. The top three most significant features for predicting mortality.

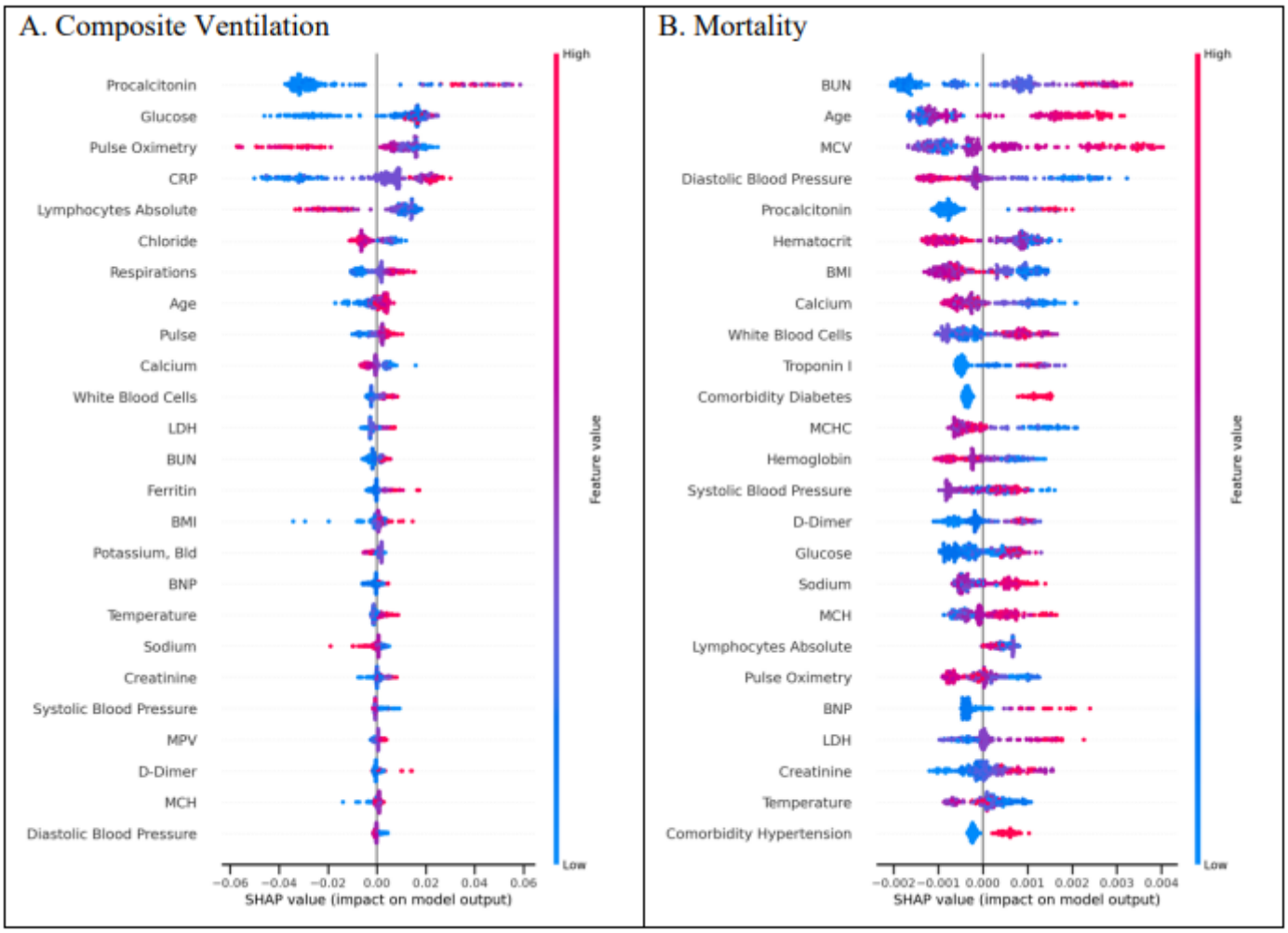

Figure 4

SHAP scatter plot for prediction using all features. Positive SHAP values imply the corresponding feature was indicative of ventilation/death. Negative SHAP values imply the corresponding feature was indicative of no ventilation/discharged. A zero SHAP value implies the feature has no impact on the predicted outcome. The normalized range of value of features are color-coded.

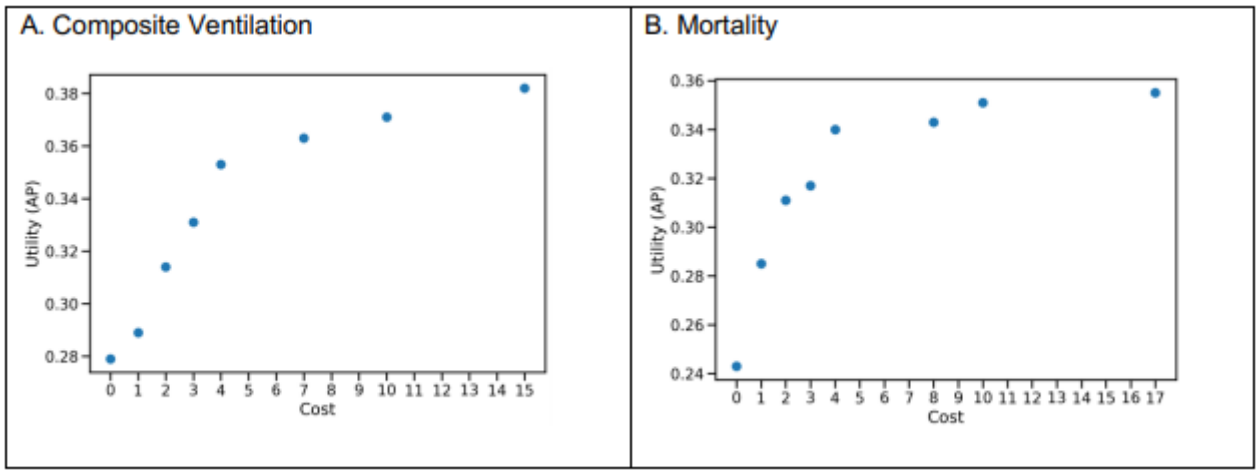

Figure 5

Visualization of Total cost versus Utility. Although an increase in budget allows more clinical feature groups in our selection, this does not guarantee an increase in performance. The performance maximizes when the total costs of features is 15 for composite ventilation and 17 for mortality, and any additional feature does not increase utility.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.pdf