medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Generalized Prediction of Hemodynamic Shock in Intensive Care Units

Aditya Nagori[1,2,3], Pradeep Singh[1], Sameena Firdos[1], Vanshika Vats[1], Arushi Gupta[1], Harsh Bandhey[1], Anushtha Kalia[4], Arjun Sharma[4], Prakriti Ailavadi[6], Raghav Awasthi[1], Wrik Bhadra[1], Ayushmaan Kaul[1], Rakesh Lodha[5], Tavpritesh Sethi[1,5*]

1. Indraprastha Institute of Information Technology Delhi, 110020, Delhi, India
2. CSIR-Institute of Genomics and Integrative Biology, New Delhi, 110007, India
3. Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, 201002, India
4. Cluster Innovation Centre, University of Delhi
5. All India Institute of Medical Sciences, Department of Pediatrics, New Delhi, 110029, India
6. Netaji Subhas University Of Technology, Delhi

Abstract

Early prediction of hemodynamic shock in the ICU can save lives. Several studies have leveraged a combination of vitals, lab investigations, and clinical data to construct early warning systems for shock. However, these have a limited potential of generalization to diverse settings due to reliance on non-real-time data. Monitoring data from vitals can provide an early real-time prediction of Hemodynamic shock which can precede the clinical diagnosis to guide early therapy decisions. Generalization across age and geographical context is an unaddressed challenge. In this retrospective observational study, we built real-time shock prediction models generalized across age groups (adult and pediatric), ICU-types, and geographies. We developed generalizable models on publicly available eICU dataset, which is externally validated on cohorts derived from more than 0.23 million patient-hours of vitals data from a pediatric ICU in New Delhi and 1 million patient-hours vitals data from the adult ICU MIMIC-III database. Out of 208 hospitals data of eICU, we found 156 eligible for cohort building and split this data hospital-wise in a 5 fold training-validation-test set. Our model predicted hemodynamic shock 8 hours in advance

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

with AUROC of 86 %(SD= 1.4) and AUPRC of 93% (SD =1.2). Our models identified 92% of all the shock events more than 8 hours in advance. Upon external validation on the MIMIC-III cohort, it achieved an AUROC of 87 %(SD =1.8), AUPRC 92 %(SD=1.6). External validation of our models on New Delhi's Pediatric SafeICU data achieved an AUROC of 87 %(SD =4) AUPRC 91% (SD=3). Therefore, our models can guide early therapy decisions to save lives, reduce false alarms and address the generalizability gap. Our data and algorithms are publicly available as a pre-configured Docker environment at https://github.com/SAFE-ICU/ShoQPred.

Keywords: Hemodynamic Shock, Intensive Care, Deep Learning, Predictive Modeling, Software

Introduction

Shock is one of the most common complications in patients admitted to the ICUs with incidence as high as 33%[1]. Hypovolemic, cardiogenic and septic shock are all characterized by altered hemodynamics[2], hence also referred as hemodynamic shock (HS). The mortality rate in patients who develop shock in ICU is as high as 34% in developing countries[3], triggered by a cascade of poor blood perfusion, inadequate oxygen availability to vital organs and multiple organ failure. Early identification is critical for aggressive management[4,] improved patient outcomes and mortality reduction[5,6,7,8] However a delay in early identification can compound the risk of mortality and organ failure. These delays can be attributed to a few observations as follows: firstly, there is a plethora of information being generated inside an ICU. For physicians and nurses to observe the data on a real-time basis is overwhelming. Information overload has worsened the physician's workload, contributed towards burnout and increased the chances of errors[9]. Studies have found an association between burnout and jeopardised care[10,11,12]. Thus, removing the risk of informational overload using Machine learning based assistive tailored models can prevent the risk of

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

suboptimal care[9]. Patient care also gets compromised as a result of poor strength of nurses or ICU staff. Studies have found lowering the bed-to-nurse ratio or ICU staffing can result in poor patient outcomes[13,14]. Therefore, the artificial intelligence based algorithm can help in monitoring and prioritizing the deteriorating patients. It's very important to reduce the information burden by building high performance decision models which can recognise the distinguishing patterns to assist the ICU staff with actionable and interpretable insights for decision making. Hemodynamic instability is one of the initial signs related to manifestation of further deterioration, therefore, we built a decision support system which can produce early warning scores with high recall and precision.

Secondly, development of models for early warning monitoring support systems should have feasibility of deployment to the same setting as well to other settings. Literature is full of models relying upon information from various sources and tested retrospectively [15,16,17,18], their prospective deployment to the same other settings seems to be suspicious. There is an underlying hospital to hospital variability in data integration and collection which hinders the deployability and hence the performance. Thus, feasibility of model deployment should be taken care of well during the model development.

In this work, we developed real-time models for early identification of shock using high resolution vitals time series data using hand-engineered time-series features, machine learning and deep learning approaches. Vitals time-series are routinely generated at a much higher resolution than hourly nursing notes, hence have the potential to forecast critical outcomes[19 20 21]. Our Safe-ICU data warehouse[22] with more than 0.23 million hours of patient physiological time-series vitals data allowed us to validate on pediatric age groups.

High resolution vitals time-series data has shown potential for multicenter generalization[21] but a very few studies have been conducted to evaluate generalizability of AI models for the ICU[15,21,23]. In our knowledge, none of the studies evaluated the potential of models learnt on

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

the adult population and generalized to pediatric one. One of the key challenges to do so is the dependency on a large set of clinical measurements and difference in age based criteria. These challenges can be overcome by using high resolution physiological vitals time-series data generated through monitoring sensors and transfer learning[21, 24] .

In this work, we reported machine learning based prediction models which take readily available time-series vitals data to forecast shock status 0 to 8 hour ahead of its onset. We do this by using hand engineered time-series features derived from the raw signals to build these models (Figure 1). There is a high percentage of patients who developed shock in the ICU. We built cohorts around patients who developed shock after 7 hours of ICU admission. Artificial intelligence algorithms have the potential to generalize and scale therefore we evaluated our models for their ability to generalize on MIMIC-III adult ICU and SafeICU pediatric ICU data using transfer learning approach. Thus, the aim of our work was to utilize readily available ICU time-series data to build robust parsimonious predictive models for the onset of Hemodynamic shock and evaluate the potential for generalization of these models to another setting.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

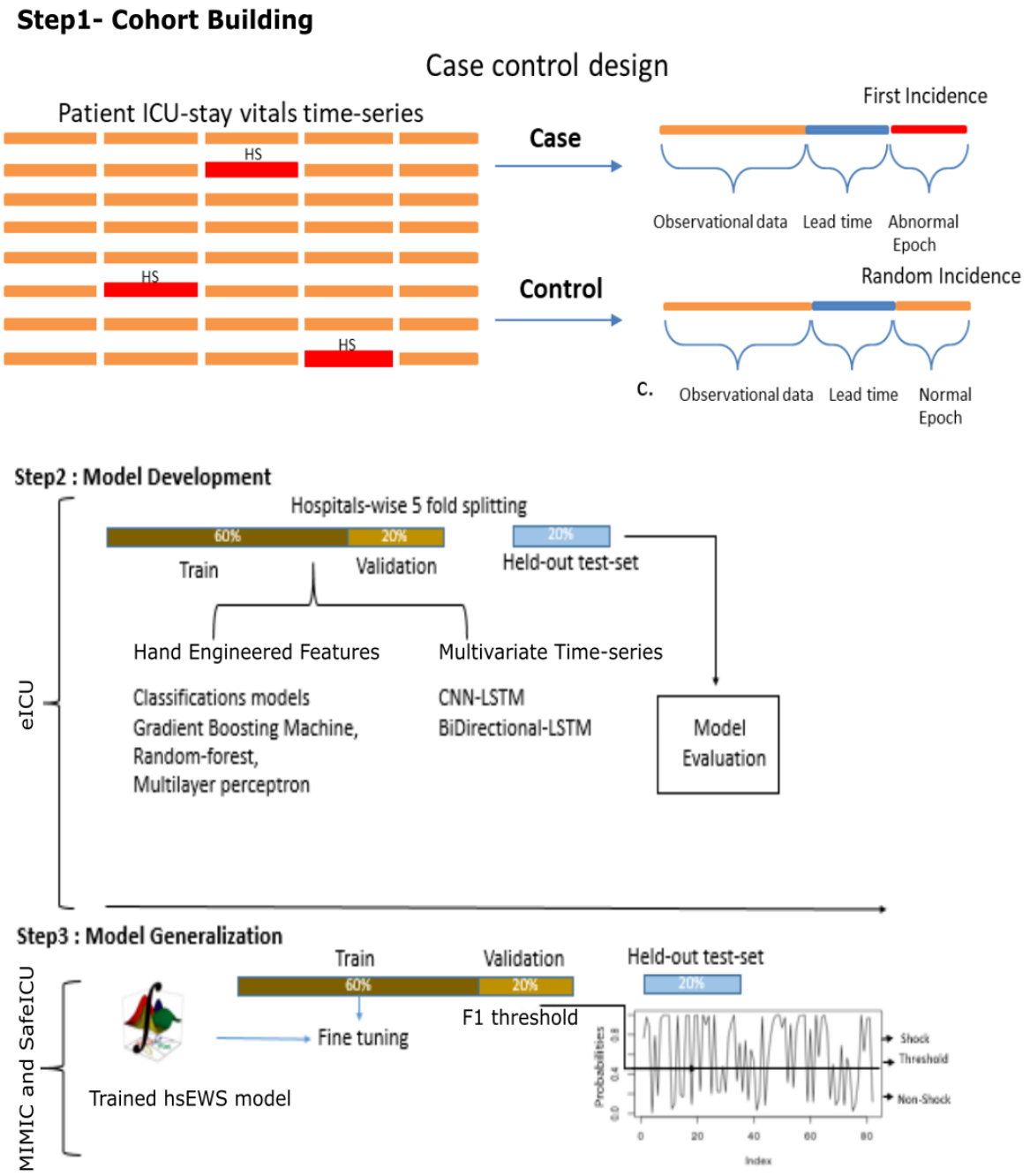


Figure 1: Summary of the pipeline for generalizing the prediction of hemodynamic shock. Step 1 shows Cohort Building. Step 2 shows the model building for the prediction of shock 0-8 hours in advance and evaluation. Step 3 shows the generalization of learned models on MIMIC-III data and Indian settings's SafeICU data.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Results

The vitals periodic data of eICU data was pre-processed according to inclusion and exclusion criteria, resulting in 16,246 admissions (Supplementary figure 1a). The median length of stay of patients in the extracted cohort was 3.1 days. The median length of stay with hemodynamic shock was found to be 15 hours. As expected, our data show that mortality is positively correlated with the fraction of length of stay with hemodynamic shock (pearson's correlation $r = 0.89$, $p = 4.8*10^{-4}$) (Supplementary figure 2). MIMIC-III and SafeICU (New Delhi) data served as external validation sets for adult and pediatric settings respectively. MIMIC-III matched subset contains 22,247 numeric data files corresponding to 10,269 subject ids with vital time-series data. Subject-wise data is segregated for 17,294 ICU stays using ICU in-time and out-time information from the ICUSTAY table (Supplementary figure 1b). Inclusion and exclusion criteria and case-control cohort extraction criteria were the same as eICU (Supplementary figure 1b, 1c). The mean and standard deviation values of the physiological variables used in the modeling are listed in table1.

Model development and performance on eICU data.

We trained and tested models on the eICU's hospital ID-wise 5 fold data splits. We found the Gradient boosting classifier model as the best performing model (Supplementary figure 2). We termed our final model hemodynamic shock early warning system (hsEWS), which generates prediction every 5 minutes and interpretable SHAP values for the observational time window. Depending upon the presence and absence of arterial blood pressure, we created two types of models, called hsEWS-invasive and hsEWS-non-invasive, respectively. The hsEWS-invasive model utilizes Arterial blood pressures (Systolic and Diastolic) achieved an AUROC of 86% and AUPRC of 93% (figure2 a. b.). hsEWS was found to identify 92% of all the hemodynamic shock events with an overall precision of 81% (figure2

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

b). The hsEWS-non-invasive model achieved an AUPRC of 90% and AUROC of 80%, and the overall recall was 90% with a precision of 79%. Therefore, for the subjects for whom arterial blood pressure is not being measured, a high recall hsEWS-non-invasive model can be applied. The model's performance declined with increasing lead-time; however, the recall remained above 89% for all lead times.

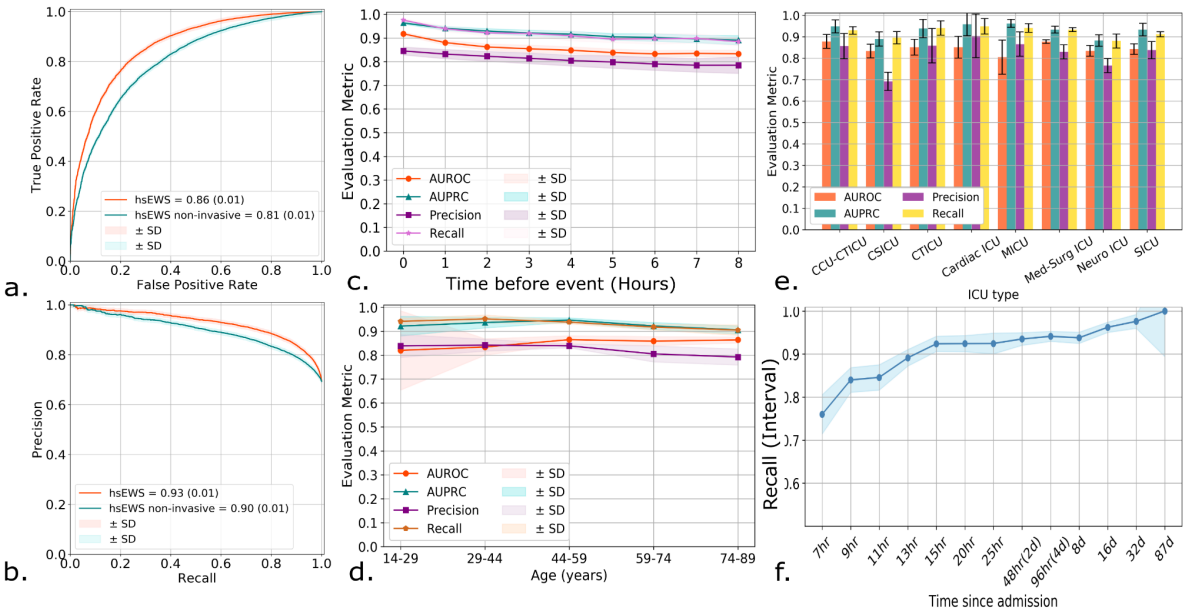


Figure2: a.) AUROC for the prediction of Hemodynamic shock in the next 5 min to 8 hours. b.) AUPRC for the prediction of Hemodynamic shock in the next 5min to 8 hours. c.) Model performance for different lead time or time before the event. d.) Model performance for different age groups, eICU data contain adult age groups. e.) Model performance in different unit stay types (CCU – Coronary care unit, CSICU- Cardiac surgery ICU, CTICU- Cardiothoracic ICU, MICU-Medical ICU, SICU- Surgical ICU). f.) Recall of the model for each time-interval shown on the x-axis for the prediction made during different lengths of stays in the ICU.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Model performance for different age groups, unit types and time-since admission.

The model performance remained consistent for different age groups with a slight decline in higher age groups above 60. However, AUROC increased for age groups above 60, and the model recall remained above 90% with precision above 79% for all age groups (figure 2d.). We also evaluated performance for different types of ICU stays unit types. All the ICU types were found to have recall above 90% except for Cardiac Surgery ICU and Neurological ICU, the precision among these ICUs also declined below overall precision(figure 2e.). We also observed that recall of the model increases with length of stay, and the model started predicting the event with 90% recall after the 13th hour of stay.

Interpreting the Feature Contribution in the Model Performance on test sets.

A total of 3,970 hand-engineered time-series features were narrowed down to 2,120 important predictors using Boruta. This includes 51 classes of features listed in Supplementary Table S1. After developing the models, we computed SHAP values for interpreting the feature contribution on test predictions. We plotted the summary SHAP values of the top 15 contributors among all in figure 3a. The heart rate minimum and arterial systolic blood pressure minimum were found as the top predictors. The heart-rate minimum is positively correlated to its SHAP values (figure3.b). While the arterial systolic blood pressure minimum is negatively correlated to its SHAP values (figure 3c), we also found that aggregated linear trend was positively correlated to SHAP values, indicating an increased respiratory rate has an influence on the prediction of future hemodynamic shock. We also found features in the frequency domain using continuous wavelet transform (Cwt) coefficients, absolute energy, quintiles, mean, autocorrelation, sum of recurring values, as important, complete list of selected features is presented in the Supplementary Table S1.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

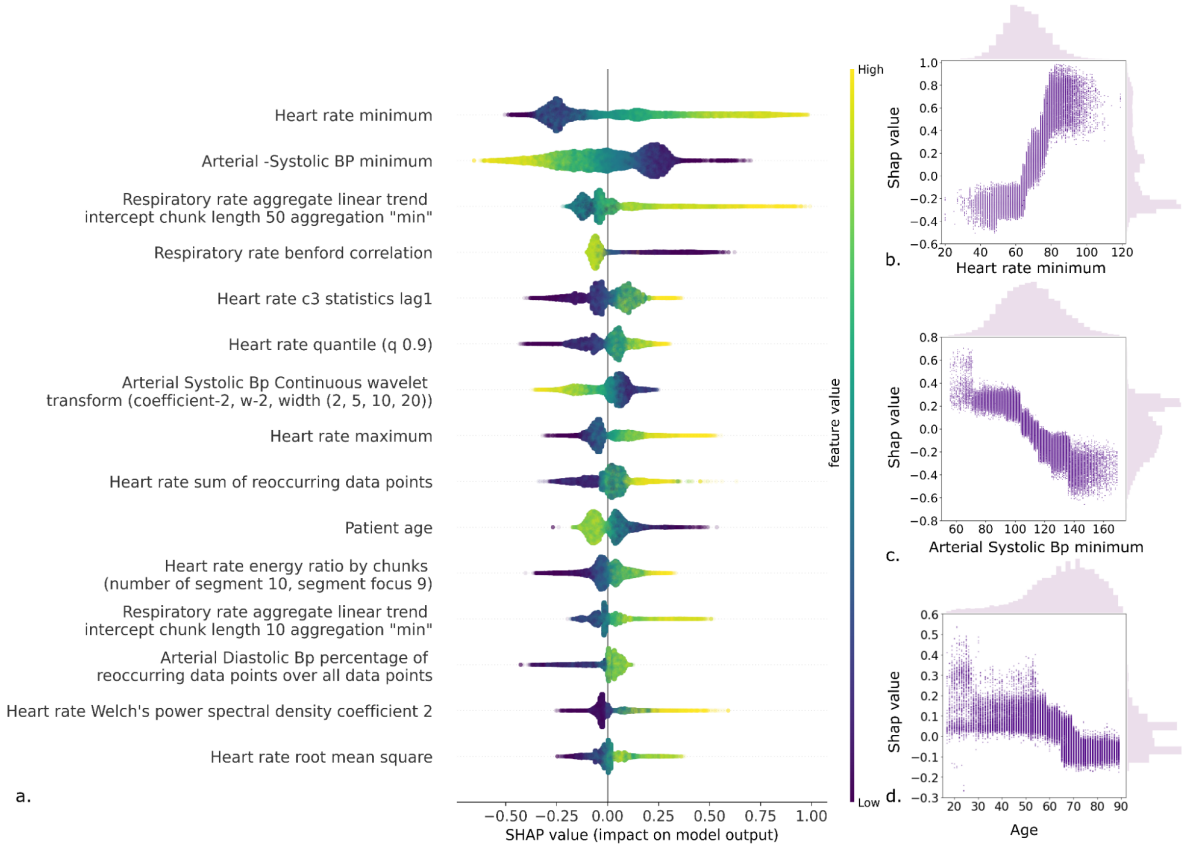


Figure3 a.) SHAP values of top 15 hand engineered time-series features (in descending order) derived from the 5 fold test set on eICU data. Violin plot shows the distribution of SHAP values, color map represents the feature value from low to high b.) A scatter plot of Heart rate minimum and corresponding SHAP values, showing positive correlation between the two. c.) A scatter plot of Arterial Systolic Blood pressure minimum and corresponding SHAP values, showing negative correlation d.) A scatter plot of Patient age and corresponding SHAP values, showing negative correlation.

Performance of external validation of hsEWS model.

For external validation, we used two datasets, MIMIC-III (adult age groups) and SafeICU data (pediatric age group). We found that with retraining on the new settings data, both in the case of MIMIC and SafeICU results improved over standalone performance and direct testing. With retraining hsEWS achieved an AUPRC of 93% on MIMIC (figure 4.a) and 91% on SafeICU (figure 4.b) and an AUROC of 87% both on MIMC and SafeICU (figure4 c, d).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

With direct testing (without retraining), hsEWS achieved an AUC of 78% and AUROC of 89% on MIMIC data. However, it performed poorly for direct testing on SafeICU pediatric data (figure 3b, 3d). As criteria for the pediatric hemodynamic shock is based on the age-specific cutoff, which differs from the adult population criteria which could be one of the reasons for the poor performance. However, retraining of hsEWS on SafeICU pediatric training data resulted in an AUROC of 87% and AUPRC of 91% on 5 fold test-set (figure 3d). Overall, model recall remained consistent for all the lead-time, except for retraining on MIMIC, in which a decline in the recall is seen with lead-time (figure 4e). However, the retraining improved the precision on both MIMIC and SafeICU data cohorts by 10-15%, as shown in figure 4f.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

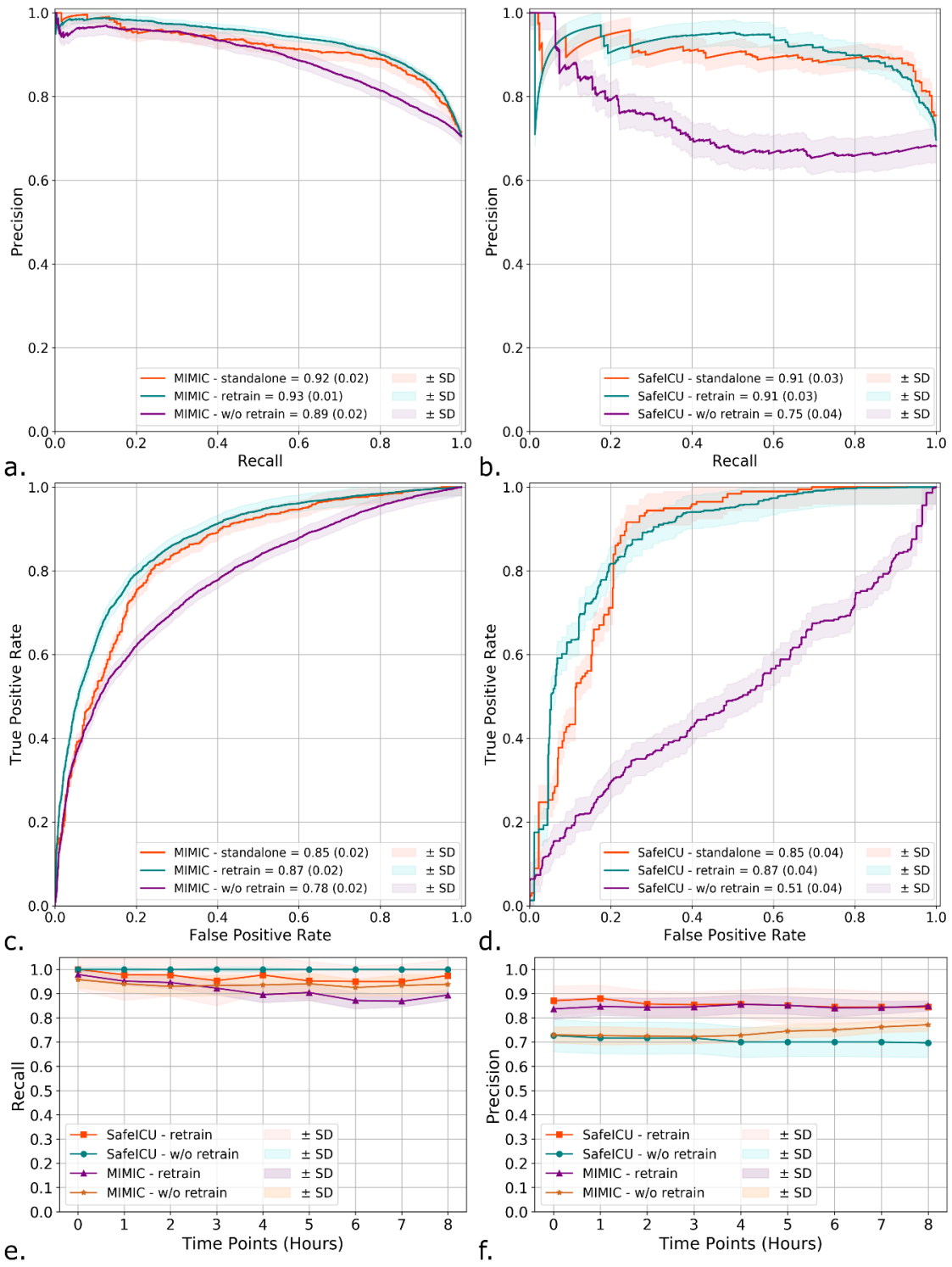


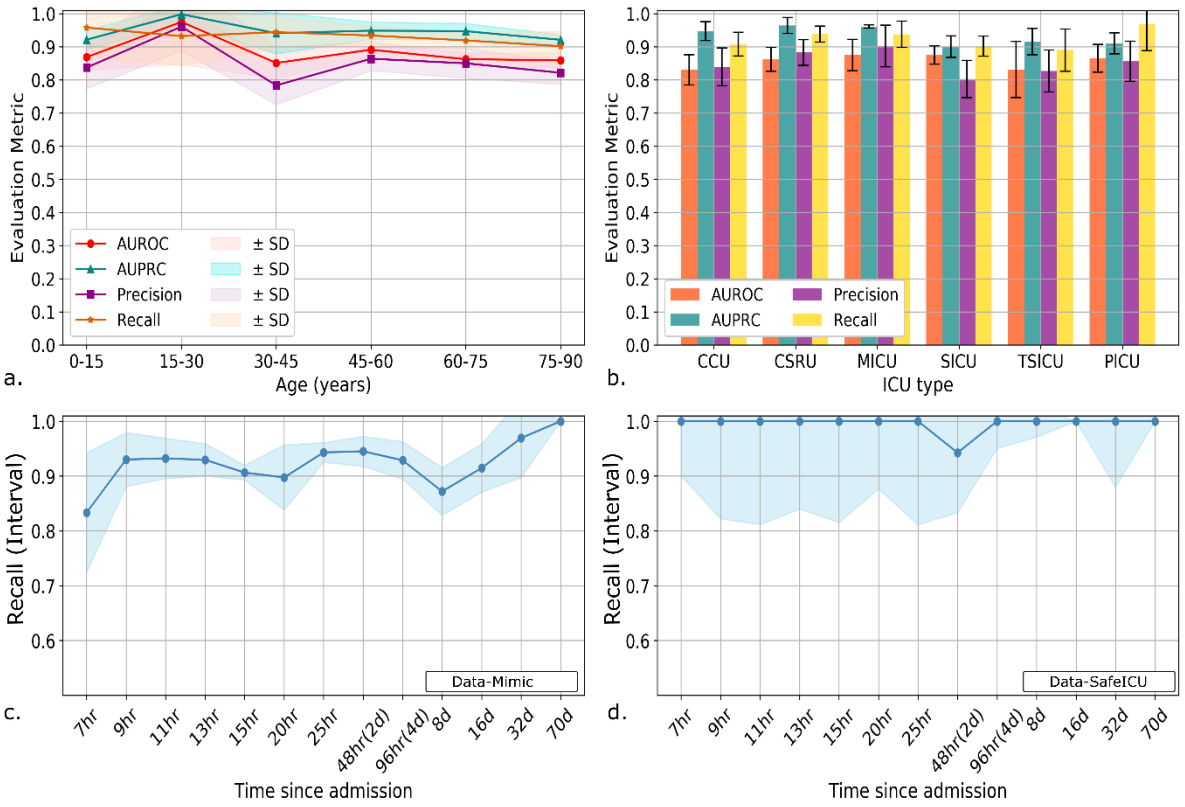
Figure 4 a.) AUPRC on MIMIC data for standalone, retraining and without retraining (w/o retrain) of hsEWS model. b.) AUPRC on SafeICU data for standalone, retraining and without retraining (w/o retrain) of hsEWS model. c.) AUROC on MIMIC data for standalone, retraining and without retraining (w/o retrain) of hsEWS model. d.) AUPRC on SafeICU data

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

for standalone, retraining and without retraining (w/o retrain) of hsEWS model e.) Recall of retrained and without retrained hsEWS model on SafeICU and MIMIC data. e.) Precision of retrained and without retrained hsEWS model on SafeICU and MIMIC data.

Performance of hsEWS retrained model on different age groups, new setting's ICU types, and with time since admission.

hsEWS retrained on MIMIC is tested on MIMIC's test-sets and hsEWS retrained on SafeICU is tested on SafeICU test-sets. Models showed a consistent recall above 90% on all age groups with a precision greater than 81% except for the 30-40 age group. Models showed a consistently higher performance recall across all ICU types (figure 5b.), where precision remained 80% and above across all ICU types. In the MIMIC cohort, performance with time since admission reached an interval recall above 90% from the 9th hour of stay, while in SafeICU, it remained consistently high with a recall of 100% from the beginning.



medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure5: a.) Model performance parameters for different age groups (SafeICU -pediatric age group is embedded along with MIMIC-III (Adult age groups) data) b.) Performance of retrained hsEWS model in different first care units. c.) hsEWS retrained on MIMIC data, model Recall for "time since admission" d.) hsEWS retrained on SafeICU data, model Recall for "time since admission"

Discussion

Development of machine learning based early warning systems for ICUs is not new[15,16,17,18]. However, a majority of these models have not made their way to successful deployments in the ICUs due to challenges of generalizability, interpretability and workflow integration[25 26]. Further, it is generally expected that deep learning will outperform handcrafted features given sufficiently large data. Our research addresses these two challenges and shows for the first time that hemodynamic shock can be predicted in ICUs with models trained on patients with completely different demographic, sociocultural and geographic factors. Further our research opens the way to building interpretable models using handcrafted features which outperformed state of the art deep learning models in predicting shock.

Clinically, the correlation between increased mortality in late onset shock is indicative of a global under-recognition of this major killer in the ICUs despite reports suggesting a higher odds of mortality for late onset of shock[27,28]. We also found that a 10% increase in the time fraction of ICU stay spent in hemodynamic shock led to an estimated 6% increase in mortality rate. Our model Recall was found to increase for the late onset of shock (Figure 2e, 5c, 5d), so our model can reduce the mortality rate associated with late onset of hemodynamic shock.

In this work, we predicted hemodynamic shock using hand-engineered time-series features derived from 5 physiological time-series. The choice of these time-series was made due to their availability at bedside across different age groups, geographies and resource limiting

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

settings. It is often seen that many subjects are not monitored for arterial blood pressure, therefore many models are not generalizable to those subjects. We therefore also trained a non-invasive hsEWS model which does not require invasive blood pressure measurements. Every hour of delays in the detection of shock contributes to the increased risk of mortality in patients admitted to the ICU[6, 8]. A reliance on a large number of variables required to produce results hinders the model transfer for generalizability. In literature, a model with 36 variables to predict the onset of shock with an AUC of 82% has been previously reported[17]. Our model on pediatric ICU surpassed the performance reported in literature by 5% with a reduced number of variables required. These models have limitations of collecting a large number of inputs which is not feasible for every minute resolution. We in this work used the readily available data to combat the issues related to large numbers of inputs. Our model uses the 5 vitals signals data which is routinely monitored electronically. We have made use of time-series of the patient, which ensures to supply patterns in the patient's physiology over a window of time unlike a single time-point value. The hand-engineered time-series features were computed to represent the dynamics of the patient physiology described in Supplementary Table1.

We tested our model across age-groups, ICU-types for different hospitals across the USA and a pediatric ICU in India. Our assessment showed a consistent performance across these groups. We observed that with increase in the lead time the performance decreases, however our overall recall is 92% at prevalence of 83%.

In recent years, transfer learning has assumed an increasing importance especially in the low data regimes[29, 21, 30]. We transferred the hsEWS model to MIMIC-3 setting and SafeICU setting and show that transferring models trained on eICU data to MIMIC and SafeICU helped improve the overall AUC and precision while maintaining an overall higher recall above 90% (figure 5e and 5f).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Threshold computation was done to account the population prevalence. Our model can be generalized by learning new population based probability thresholds.

There are few limitations of our study. Our models are developed and tested on the retrospective data and there is a need for testing the models prospectively. We have created an easy to use docker container which works well with streaming real-time data from monitoring. All the data were taken from a running ICU therefore almost all of the patients were receiving some form of medication and might have come with a history of shock which is not taken into account. However, this does not affect the prediction potential of our model since we made sure to include only subjects who developed shock in the unit. Also the training window or observational and lead-time data are free from any incidence of hemodynamic shock.

Finally, we believe that the future directions should synergize hand-engineered and deep learning based predictions for improving patient outcomes. This will especially hold true for multimodal data including signals, text, tabular and imaging data. While many handcrafted features in our work are straightforward to interpret, others may still need research for clinical interpretation. Here, we have offset this limitation by using Shapley Additive exPlanation (SHAP) value based model interpretation where we have computed SHAP values corresponding to each feature. These values can represent the influence of the feature on the prediction.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

Dataset description & preprocessing

This study utilized two datasets available in the public domain along with our pediatric data resource (SAFE-ICU). The eICU data[31] is a publicly available data containing more than 800 million observations for 1,92,751 patient admissions to 208 hospitals across the USA. It contains 12.2 million patient hours of vitals periodic data at the 5-minute resolution. Our inclusion criteria specified ICU stay >= 7.5 hours, absence of shock within the first seven hours of ICU admission, less than 10% missingness of data and presence of arterial blood pressure measurements. The second dataset "Medical Information Mart for Intensive Care (MIMIC)"[32]. MIMIC dataset hosted by physionet.org website as a publicly available data resource. Datasets are de-identified and available for analysis as per the approval by MIT institutional review boards (IRBs) documented on the website. These data were collected between 2001 to 2012 at Beth Israel Deaconess Medical Center (BIDMC). We have used MIMIC III v1.4 which was released on September 2nd 2016. 22,247 numeric records that have been time aligned and matched with 10,282 MIMIC III clinical database records, were used for generating the Mimic Shock cohort. These data are summarized at 1 minute resolutions and merged using subject Ids. Further subject records were splitted into respective ICU stays based on the in-time and out-time given in the "ICUSTAYS" table of the clinical data of MIMIC-III database.

Third dataset, SafeICU data resource[22](Sepsis advanced forecasting engine ICU Database) is an in-house ICU data resource built at Pediatric ICU of All India Institute of Medical Science, New Delhi. Ethical permission was sought and was approved by the Institutional review board (IEC/NP-211/08.05.2015). SafeICU data collected between February, 2016 to January 2020 were used for constructing a final validation across the continent cohort.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Imputation - All the time-series data was imputed using Kalman filter imputation[33] up to 80% missingness, however in the final training observation windows were selected based on the "not more than 10% imputation criteria". Shock index was not imputed.

Epoch and Cohort Generation.

We created 30 minute epochs for dense labelling of the time-interval as shock or no-shock. The labels for shock were derived using shock-index (SI). A time of onset of Shock ($t_{(shock)}$) is defined as the starting time of an epoch where the median shock-index was greater than 0.7. We took time-series data of 420 minutes of the five signals; heart-rate (HR), systolic arterial blood pressure (Sys-Abp), diastolic arterial blood pressure (Dia-Abp), respiratory rate, oxygen saturation (SpO2). Lead time of 0 to 8 hour at 1 hour interval prior to the first occurrence of a 30-minute shock window was taken. Training data is further filtered based on 10% or less imputation. These scores can be computed for every-timestamp present in the numeric data, thus precisely labels the onset of the condition. For the pediatric ICU data we labelled the data using age adjusted cutoff for shock index[34].

Model Development and Evaluation.

eICU data cohort was used to train models to predict the next 0-8 hour shock status. Hospital IDs were randomly divided in 5 fold cross validation sets. Further train-validation and tests were chosen according to hospital ID. So every time the model was tested on new 20% hospitals data. This was done in order to optimize models for generalization on external settings. Further the Hand-engineered time-series features and raw signals data were transformed to z-scores for rescaling in order to facilitate model fitting. All the hyperparameters such as number of estimators, max features, epoch numbers, batch size, number of trees, and F1 threshold were optimized on the validation set. Final results were reported as the median and Standard deviation on 5 fold test sets.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Hand engineered times-series features extraction and selection.

Five time-series signals; heart-rate (HR), systolic arterial blood pressure (Sys-Abp), diastolic arterial blood pressure (Dia-Abp), respiratory rate, oxygen saturation (SpO2) were used to extract hand engineered features. These features, also termed as hand-engineered time-series (HETS) features were extracted using the tsfresh python package[35]. This includes wavelet transform coefficients, Fourier transform coefficients, discriminative power, linear trends, recurrent value based features etc. Python library "tsfresh" was used to calculate the features on the time-series data of the cohort. A total of 3970 features were extracted using tsfresh python package. Further variable selection was performed using Boruta which is a feature selection algorithm, implemented through R package "Boruta"[36]. Boruta selected features are listed in Supplementary Table S1.

Modeling on hand engineered time-series features.

We built three models on hand-engineered time-series features, 1. Gradient boosting classifiers[37] , 2. Random-Forest models[38] using sklearn (version 2.1.0) library in python 3.6. 3. Multilayer-perceptrons were trained on the hand-engineered time-series features selected after running feature selection algorithm Boruta. Hyperparameters tuning was performed using grid search for optimizing model performance parameters on the validation set.

Modeling on raw signals.

LSTM (long short term memory) network is a type of recurrent neural network (RNN) which is capable of learning sequence information for temporal prediction problems[39]. As there will be lags of unknown duration between important events in a time series, LSTM are useful in time series data to process, classify and predict.

We tested two LSTM based architecture, first CNN-LSTM, it contains convolutional neural networks followed by LSTM which is connected to a two dense layer perceptron to produce probability scores. Second, we used Bi-Directional LSTM[40] model, the model uses a

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

time-distributed layer followed by Bi-Directional LSTM layer which is further connected to three dense layers with tanh as activation function on first two layer of dense and softmax at last layer to get probabilities. The models were fitted in python 3.6 using Keras (version 2.3.1), Tensorflow (version 2.1.0). For static features age and gender, we used a single layer perceptron in parallel to the LSTM model, the weights of the last layer of LSTM model and single layer perceptron were concatenated. Finally, a multilayer perceptron with softmax activation was used to generate probability scores on the concatenated time-series and weighted static variable features.

hsEWS and hsEWS-non-invasive model.

The best model is selected from different classifiers mentioned above using AUROC and AUPRC. Selected model was used for further evaluation; this model generates a score between 0 and 1 if found greater than a threshold attributed to future Shock and non-shock risk, the model termed as hsEWS We computed this threshold on the validation set. The test is kept untouched.

SHAP value computation for model interpretability.

SHAP (SHapley Additive exPlanations) is a method which assigns each feature an importance value based on game theoretic principles[41]. The SHAP values represent the influence of features on model prediction. We computed SHAP value distribution of each feature for all the testsets. All the testsets were combined to plot the model explanation in terms of top 15 important features and their relationship with SHAP values.

Model generalization.

We evaluated our models on two external validation sets. For this, we used MIMIC-III and SafeICU (Pediatric) data. We performed similar pre-processing with slight data specific variation and cohort building as done for eICU. The higher resolution MIMIC Data and safeICU data were brought to 5 minute resolution to match eICU resolution. We tested three

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

scenarios of external validation. First, a standalone model, where the Model is trained on the new setting's data itself. Second, in which model was retrained on new setting's training data and tested on the 5 fold test sets. Third, we tested the hsEWS model directly on external validation data. For data specific prevalence based probability threshold, we used the validation set of the external validation setting. We choose the threshold over the precision recall curve which gives the maximum F1 score, in order to optimize precision and recall.

Acknowledgements

This work was supported by the Wellcome Trust/DBT India Alliance Fellowship IA/CPHE/14/1/501504 awarded to Tavpritesh Sethi. We thank Dr. Anurag Agrawal, Director, CSIR-Institute of Genomics and Integrative Biology for his valuable suggestions while preparing the manuscript. We also thank Mr. Varun Prakash and Mr. Anil Sharma for the technical support provided at PICU, AIIMS, and New Delhi.

The funders had no role in the execution of this study or the interpretation of the results.

Author contributions:

AN, TS designed the study. AN, PS, TS, RL were involved in acquiring data. AN, PS performed pre-processing and developed the cohorts, AN, PS, SF, AG involved in feature computation, feature selection AN, SF, PS, AG, AK, AS Developed models and performed model generalization, AN, VV formal analysis and visualized the results, AN, VV, RA involved in statistical analysis. WB, AMK performed multi-feature model experiments, AN, TS interpreted the data and wrote the first draft of the report. TS, AN, RL revised the report critically for important intellectual content. HB, PA, AN involved in the Docker App development.

Conflicts of Interest and Source of Funding: The authors declare that they have no competing interests. This work was supported by the Wellcome Trust/DBT India Alliance Fellowship IA/CPHE/14/1/501504 awarded to Tavpritesh Sethi. Mr. Aditya Nagori is

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

supported by CSIR-GATE fellowship. Mr. Pradeep Singh is supported through the Indo-Israel collaborative research grant received by Dr. Tavpritesh Sethi and Dr. Rakesh Lodha.

References

1. Vincent, J.-L. & De Backer, D. Circulatory shock. *N. Engl. J. Med.* **369**, 1726–1734 (2013).

2. Hendy, A. & Bubenek-Turconi, Ş.-I. The Diagnosis and Hemodynamic Monitoring of Circulatory Shock: Current and Future Trends. *J. Crit. Care Med.* **2**, 115–123 (2016).

3. Yealy, D. M. *et al.* Recognizing and managing sepsis: what needs to be done? ??? (2015). doi:10.1186/s12916-015-0335-2

4. Armen, S. B. *et al.* Improving Outcomes in Patients With Sepsis. *Am. J. Med. Qual.* **31**, 56–63 (2016).

5. Be1. Berger, T. et al. Shock index and early recognition of sepsis in the emergency department: pilot study. West. J. Emerg. Med. 14, 168–174 (2013).rger, T. *et al.* Shock index and early recognition of sepsis in the emergency department: pilot study. *West. J. Emerg. Med.* **14**, 168–174 (2013).

6. Goal-directed, E., Group, C. & Province, Z. The effect of early goal-directed therapy on treatment of critical patients with severe sepsis/septic shock: a multi-center, prospective, randomized, controlled study. *Zhongguo Wei Zhong Bing Ji Jiu Yi Xue* **22**, 331–334 (2010).

7. Herget-Rosenthal, S., Saner, F. & Chawla, L. S. Approach to Hemodynamic Shock and Vasopressors. *Clin J Am Soc Nephrol* **31. Herget**, 546–553 (2008).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

8. Bai, X. *et al.* Early versus delayed administration of norepinephrine in patients with septic shock. *Crit. Care* **18**, (2014).

9. Furlow, B. Information overload and unsustainable workloads in the era of electronic health records. *Lancet Respir. Med.* **8**, 243–244 (2020).

10. Papazian, L., Sylvestre, A. & Herridge, M. Should all ICU clinicians regularly be tested for burnout? Yes. *Intensive Care Medicine* **44**, 681–683 (2018).

11. Hall, L. H., Johnson, J., Watt, I., Tsipa, A. & O'Connor, D. B. Healthcare staff wellbeing, burnout, and patient safety: A systematic review. *PLoS ONE* **11**, e0159015 (2016).

12. Kim, M. H. *et al.* Burnout and self-reported suboptimal patient care amongst health care workers providing HIV care in Malawi. *PLoS One* **13**, (2018).

13. Jung, M. *et al.* The effect of bed-to-nurse ratio on hospital mortality of critically ill children on mechanical ventilation: a nationwide population-based study. *Ann. Intensive Care* **10**, 159 (2020).

14. Clarke, S. P. & Donaldson, N. E. *Nurse Staffing and Patient Care Quality and Safety*. *Patient Safety and Quality: An Evidence-Based Handbook for Nurses* (Agency for Healthcare Research and Quality (US), 2008).

15. Hyland, S. L. *et al.* Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat. Med.* **26**, 364–373 (2020).

16. Henry, K. E., Hager, D. N., Pronovost, P. J. & Saria, S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci. Transl. Med.* **7**, (2015).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

17. Potes, C. *et al.* A clinical prediction model to identify patients at high risk of hemodynamic instability in the pediatric intensive care unit. *Crit. Care* **21**, (2017).

18. Nemati, S. *et al.* An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Crit. Care Med.* **46**, 547–553 (2018).

19. Desautels, T. *et al.* Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med. Informatics* **4**, e28 (2016).

20. Ghassemi, M. *et al.* A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. in *Proceedings of the National Conference on Artificial Intelligence* **1**, 446–453 (AI Access Foundation, 2015).

21. Mao, Q. *et al.* Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open* **8**, 17833 (2018).

22. Sethi, T. *et al.* Validating the tele-diagnostic potential of affordable thermography in a big-data data-enabled ICU. in *ACM International Conference Proceeding Series* **Part F1276**, (2017).

23. Brajer, N. *et al.* Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission. *JAMA Netw. open* **3**, e1920733 (2020).

24. Karmakar, C., Saha, B., Palaniswami, M. & Venkatesh, S. Multi-task transfer learning for in-hospital-death prediction of ICU patients. in *Proceedings of the Annual*

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS **2016-October**, 3321–3324 (Institute of Electrical and Electronics Engineers Inc., 2016).

25. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine* **17**, 1–9 (2019).

26. Wang, F. & Preininger, A. AI in Health: State of the Art, Challenges, and Future Directions. *Yearbook of medical informatics* **28**, 16–26 (2019).

27. Sakr, Y. *et al.* Early- versus late-onset shock in European intensive care units. *Shock* **28**, 636–643 (2007).

28. Roman-Marchant, O., Orellana-Jimenez, C. E. A., De Backer, D., Melot, C. & Vincent, J. L. Septic shock of early or late onset: Does it matter? *Chest* **126**, 173–178 (2004).

29. Desautels, T. *et al.* Using Transfer Learning for Improved Mortality Prediction in a Data-Scarce Hospital Setting. *Biomed. Inform. Insights* **9**, 117822261771299 (2017).

30. Weiss, K., Khoshgoftaar, T. M. & Wang, D. D. A survey of transfer learning. *J. Big Data* **3**, 1–40 (2016).

31. Pollard, T. J. *et al.* The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 1–13 (2018).

32. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

33. Li, Q., Li, R., Ji, K. & Dai, W. Kalman filter and its application. in *Proceedings - 8th International Conference on Intelligent Networks and Intelligent Systems, ICINIS 2015* 74–77 (Institute of Electrical and Electronics Engineers Inc., 2016). doi:10.1109/ICINIS.2015.35

34. Acker, S. N. *et al.* Shock index, pediatric age-adjusted (SIPA) is more accurate than age-adjusted hypotension for trauma team activation. in *Surgery (United States)* **161**, 803–807 (2017).

35. Christ, M., Braun, N., Neuffer, J. & Kempa-Liehr, A. W. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing* **307**, 72–77 (2018).

36. Kursa, M. B. & Rudnicki, W. R. Feature selection with the boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).

37. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* (2001).

38. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

39. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).

40. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**, 2673–2681 (1997).

41. Lundberg, S. M., Allen, P. G. & Lee, S.-I. *A Unified Approach to Interpreting Model Predictions*.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

.**Table1** | Cohort characteristics

Variables	Median(sd)		Statistical Test
	Non-Shock	**Shock**	**Negative log10 P-value**
Age (years)	67.5(13.4)	64.29 (14.62)	29.53(w)*
Arterial Systolic Blood pressure, mm Hg	139.74 (11.97)	135.15 (10.98)	32.8(w)*
Arterial Diastolic Blood pressure, mm Hg	59.8 (6.36)	60.55 (5.78)	4(w)*
Heart rate, per min	67.42 (4.6)	74.31 (4.4)	216.82(w)*
Respiratory rate, per min	16.58 (3)	18.4 (2.98)	88.31(w)*
Oxygen Saturation	97.53 (1.31)	97.41 (1.25)	0.22(w)
Gender (Female %)	40%	37%	2.15(c)*
Length of stay (days)	1.18(3.05)	4.29(7.8)	
length of stay with hemodynamic shock	0	15hr	
Mortality rate	1%	11%	

medRxiv preprint doi: https://doi.org/10.1101/2021.01.07.21249121; this version posted May 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Values are median (SD) unless indicated, *significance level at negative log10 p-value >= 1.3, Wilcoxon (W) rank sum test (non-parametric) or Student's t-test (t) (parametric) were used after testing for the normality assumption. c - Chi-squared test of proportions.