Generalized Prediction of Shock in Intensive Care Units using Deep Learning

Aditya Nagori[1,2,3], Anushtha Kalia[4], Arjun Sharma[4], Pradeep Singh[3], Harsh Bandhey[3], Prakriti Ailavadi[6], Raghav Awasthi[3], Wrik Bhadra[3], Ayushmann Kaul[3], Rakesh Lodha[5], Tavpritesh Sethi[3,5*]

1. CSIR-Institute of Genomics and Integrative Biology, New Delhi, 110007, India,
2. Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, 201002, India
3. Indraprastha Institute of Information Technology Delhi, 110020, Delhi, 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

Shock is a major killer in the ICU and machine learning based early predictions can potentially save lives. 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 and continents. More than 1.5 million patient-hours of novel data from a pediatric ICU in New Delhi and 5 million patient-hours from the adult ICU MIMIC database were used to build models. We achieved model generalization through a novel fractal deep-learning approach and predicted shock up to 12 hours in advance. Our deep learning models showed a receiver operating curve (AUROC) drop from 78% (95%CI, 73-83) on MIMIC data to 66% (95%CI, 54-78) on New Delhi data, outperforming standard machine learning by nearly a 10% gap. Therefore, better representations and deep learning can partly address the generalizability-gap of ICU prediction models trained across geographies. 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 termed as hemodynamic shock (HS). The mortality rate in patients who develop shock in ICU is 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] In this work, we developed real-time models for early identification of shock using high resolution vitals time series, deep learning and standard machine 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.[9,10] However, their use in predicting shock is not yet explored. Our Safe-ICU data warehouse[11] with more than 1.5 million hours of patient physiological time-series vitals data, laboratory investigation records, treatment charts, doctors and nurse assessment charts allowed us to build and validate deep learning based shock

prediction models which could generalize across continents. Representation learnt through deep neural networks have shown potential to improve the sepsis prediction model performance in ICU.

Previous studies have built models that used laboratory data and blood reports. One of the studies in Pediatric patients used 36 variables to predict the onset of hemodynamic shock and achieved an AUC of 82%[12]. Another study on the adult population predicted septic shock with an AUC of 83% at a median lead time of 28.2 hours before the onset using 54 EHR features[13]. Hyland et al predicted circulatory shock using 112 variables in a full and 16 variables in the lite model, the difference in performance is marginal and the performance of the models falls as the lead time increases[14]. Most of these models involved collecting a large set of data variables at high frequency, however, multicentric model validation has been achieved by using a minimum set of vitals as the predictors for the sepsis prediction[14,]. High resolution vitals time-series data has shown potential for multicenter generalization[15] but a very few studies have been conducted to evaluate generalizability of AI models for ICU[13, 14,15]. In our knowledge, none of the studies evaluated the potential of models learnt on 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 in both the populations at high frequency, this challenge can be overcome by using high resolution physiological vitals time-series data generated through monitoring sensors. In this work, we reported machine learning based prediction models which take readily available time-series vitals data to forecast shock status 3 to 12 hour ahead of its onset. We do this by using state-of-art- deep learning models and non-linear time-series features to build these models. There is a high percentage of the patients who developed shock in ICU. We built cohorts around patients who developed shock after spending a considerable amount of time in ICU. Artificial intelligence algorithms have potential to generalize and scale therefore we evaluated our models for their ability to generalize on our pediatric ICU data. We do this by making a comparison between deep-learning algorithms which are capable of extracting features automatically and hand-engineered features based models. We have used a battery of models whichever worked best at a given a prediction lead time we integrated that into our app. 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 HS and evaluate the potential for generalization of these models to another setting.

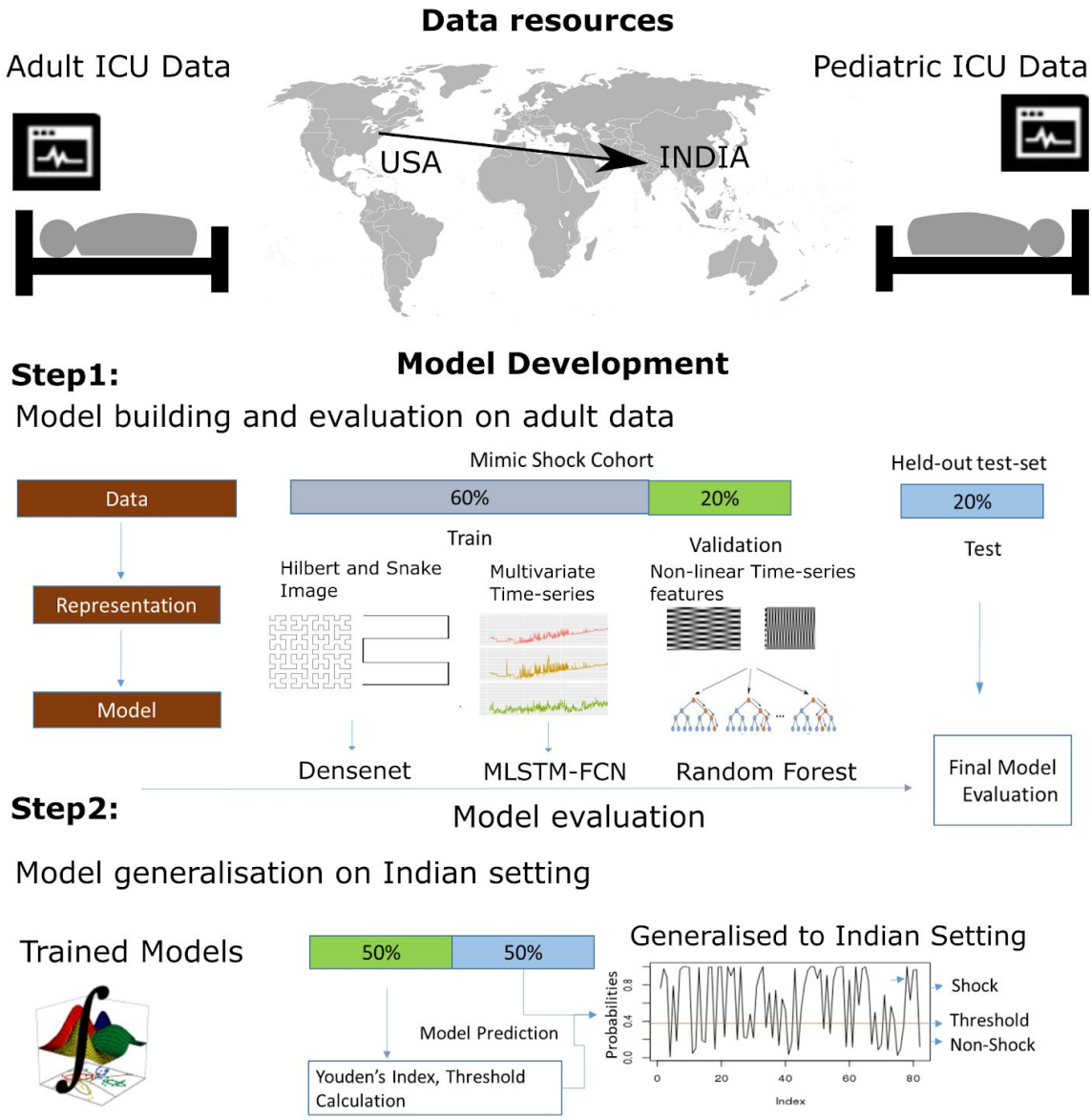

Figure. 1: Summary of the pipeline for generalizing the prediction of hemodynamic shock. Step 1 shows data resources from the USA and India. Step 2 shows the model building for the prediction of shock at 3, 4.5, 6, 7.5, 9 and 12 hours and evaluation involving image, sequence and non-linear time-series feature based models. Step 3 shows the generalization of learned models on Indian settings data at 3, 4.5, 6, 7.5, 9 and 12 hours.

Results:

Preprocessing, data characteristics and cohort building

Upon pre-processing the MIMIC-III matched subset of 22247 numeric data files for 1 minute resolution summarization we obtained 10269 subject ids with vital time-series data. Further the merged files were separated for ICU Stays. We obtained 17,294 ICU stays. We then applied the exclusion criteria and case-control cohort extraction as shown in Figure 5. We excluded the patients which do not have length of stay equal to or less than the sum of the observational period of 256 minutes, lead time according to cohort (3-12 hours) and a 30 minute outcome epoch window. 572 ICU stays were removed due to missing Shock Index. Since we are predicting new shocks, only the first instance of the shock event was considered. While extracting cases and controls we removed the cases in which observational windows or lead time have the Shock, later ICU-stays with observation windows having more than 10% imputation were removed from the Shock and Non-shock patients. The list of Shock and Non-shock trainable samples as per the exclusion and inclusion criteria are listed in the Cohort Characteristics table1. The mean and standard values of the features used in the modeling are listed in table1. SafeICU resource contains data at 15 seconds resolution which was brought to 1 minute resolution. SafeICU data resource AIIMS, New Delhi had 619 patients till July 2019, out of which we extracted the patients in case and control with different lead time values as done for MIMIC data. Cohort characteristics for the SafeICU data resource are listed in the table2.

Machine learning on hand engineered features found heart-rate as the most important predictor

With feature selection, we found heart-rate signals as the most important source of features selected by Boruta. Arterial blood pressure (ABP) Systolic is the second best source of important predictors. We also found 51 classes of features as important. We found a number of features in the frequency domain using continuous wavelet transform (Cwt) coefficients, Fast Fourier Transform (fft). We also found features such as absolute energy, quantile, mean, autocorrelation, Sum of recurring values as important, complete list and cohort wise and vitals wise important feature list presented in the Supplementary Table S1 and Supplementary Table S2. We then performed a random-forest model which achieved AUC of 84% (95%CI, 80-88) at 3 hour on the test set which declined to 78% (95%CI, 73-83) at 12 hour as shown in Figure 4. The hyperparameters such as number of features and number of

trees were optimized for AUC and OOB (out-of-bag error) as shown in supplementary Figure 2. The model generalized to SafeICU cohorts for 3 to 12 hour prediction time which achieved an AUC of maximum 59% (95%CI, 45-73) 3 hours. These models could not cross the 60% generalization mark at any lead time.

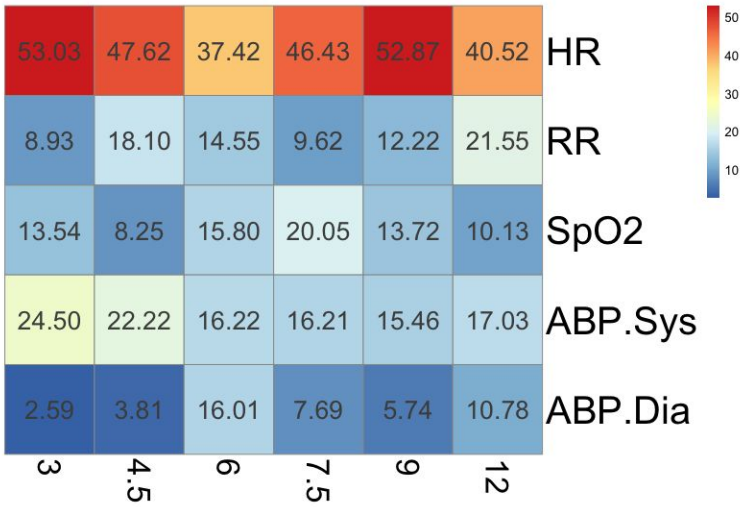

Figure 2: Heat map of the proportion of features found to be important using the Boruta algorithm at different time-points.

Performance of multivariate LSTM fully connected network and fractal deep-learning approach

Densenet models with Hilbert time-series representation achieved a maximum AUC of 83% (95%CI, 77-85) at 4.5 hours. These models when generalized to the across-site pediatric validation data, achieved an AUC of 70% (95%CI, 53-79) on Hilbert image representation when age-gender was not included. Sequence model MLSTM-FCN trained on mimic data-set for predicting shock at 3 hours achieved AUC of 83% (95%CI, 79-87) on mimic-cohort. MLSTM-FCN model when generalized to SafeICU data cohort achieved AUC of 69% (95%CI, 57-81) best among all the other models when age, gender included as predictor. More model evaluation parameters i.e. Positive predictive value (PPV) and sensitivity along with NLTS-RF model comparison are shown in Figure 3. All models performed consistently for PPV with exception to the snake-densenet model for sensitivity. Our models were able to predict at the maximum PPV of 94% (95%CI, 91-97) at 9hr before the event.

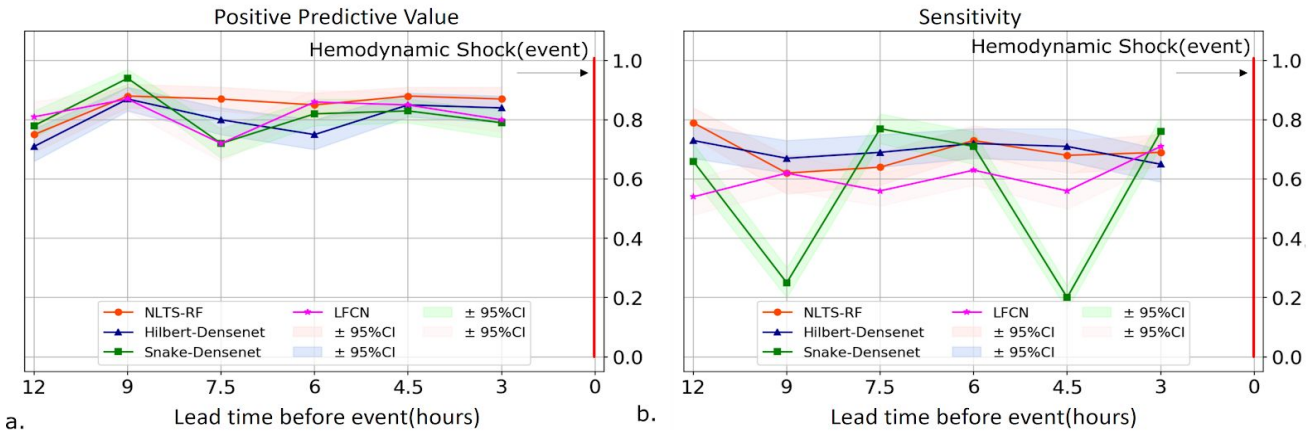

Figure 3: a.) Positive predictive value (PPV) and **b.)** Sensitivity at 3-12 hours lead time before the onset of the hemodynamic shock. The Youden-Index threshold was chosen by optimizing the trade off between PPV and sensitivity as depicted in the figure above.

Comparison of different models used, deep learning models achieved better generalization over hand engineered feature based model

Deep learning models as well nonlinear time-series feature based models were generalized to pediatric cohorts. The average AUC of generalizing deep-learning models is 63.2% and hand-engineered feature based models is 52.8%. Hand engineered features based models were found to be poorly generalizable when compared to deep-learning models as shown in (Figure 4, Table 3). This might be due to higher complexity of the hand engineered features compared to features generated from the deep-learning models. High complexity can result in poor generalization.

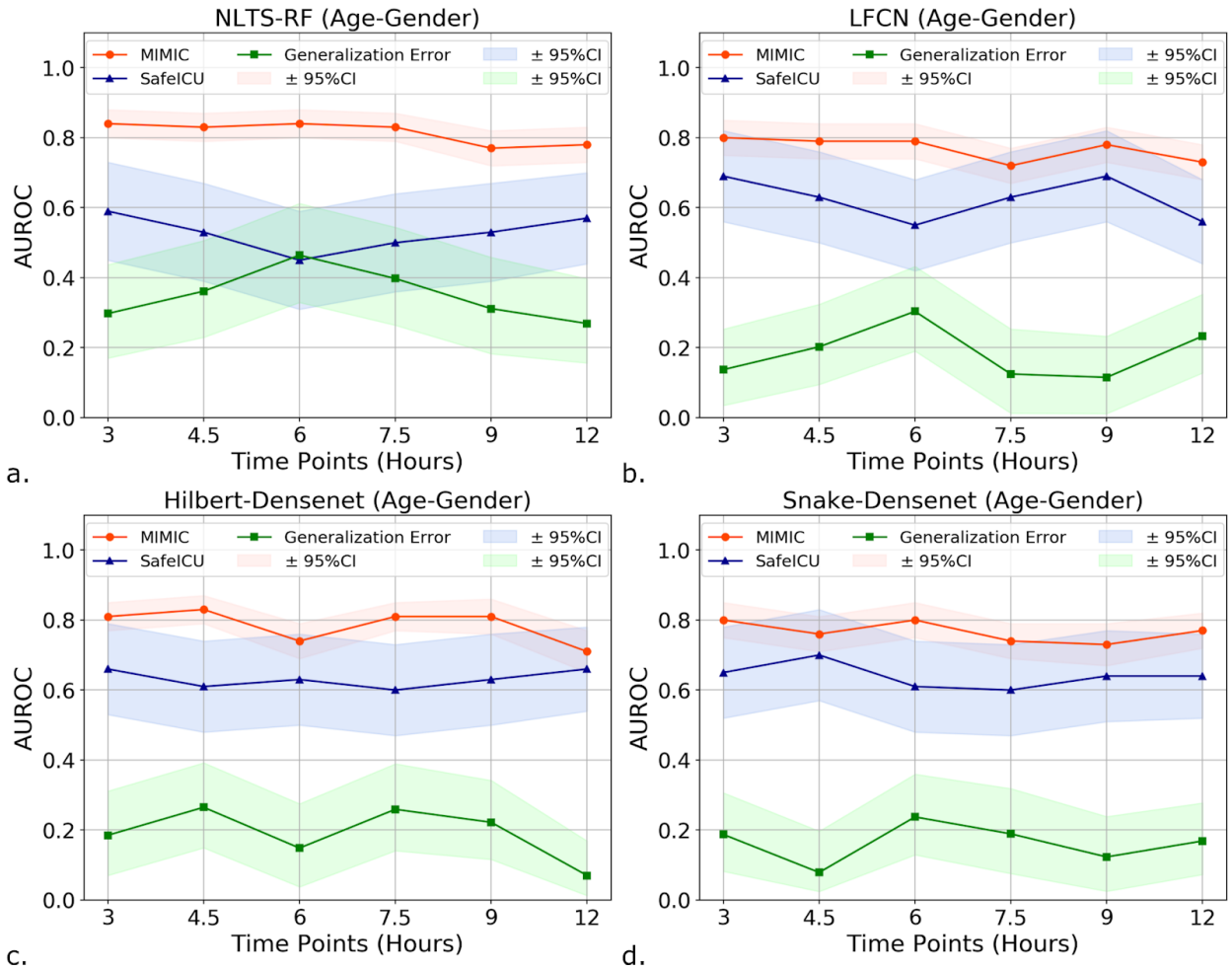

Figure 4: Time point-wise Area under the receiver operating characteristic curve on Mimic and SafeICU data and the percentage generalization error for SafeICU, for different models i.e. **a.** Hand engineered features-Randomforest model (NLTS-RF) b**.** LFCN model c**.** Snake-Densenet model d**.** Hilbert Densenet model

Discussion:

In this work, we predicted hemodynamic shock using deep-learning and hand-engineered features. Every hour of delays in the detection of shock contributes to the increased risk of mortality in patients admitted to the ICU[6,8] . This delay gets compounded by the amount of the predictors required. To look for the model performance over vitals time-series data based we added features such as Lactate and Anion gap from the EHR data. The model AUC didn't improve from AUC 77% when lactate was used as a feature in the random-forest model. The sample size also reduced due to the addition of features, we compared the performance on the reduced set only. Similarly adding anion gap as a feature, results in no change in AUC from 84% at 3hr. The use of as many as 36 variables to predict

the onset of shock with an AUC of 82% have been previously reported[12]. For us adding new features didn't help. 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 also made use of longitudinal time-series of the patient, which ensure to supply patterns in the patients physiology over a window of time unlike a single time-point value. The Non-linear time-series features were computed to represent the dynamics of the patient physiology in terms of various hand-engineered time-series patterns described in Supplementary Table S2.

In recent years we have seen advancement in artificial intelligence algorithms most prominent of which is deep learning. In our work, we used two important deep-learning approaches: Image classification and sequence classification. We used the State of art deep-learning model Densenet which was shown to produce the state-of art performance on the object recognition task [16]. Secondly in deep-learning models we used the multivariate time-series classification model MLSTM-FCN[17]. This model uses an attention mechanism to look after the variables responsible for prediction. The idea to use these approaches was to explore the deep learning methods for the prediction of hemodynamic shock. Non-linear time-series models performed comparably better than MLSTM-FCN and Densenet models for the adult ICU. But the MLSTM-FCN model found to generalize well over hand-engineered time-series features based models with an average 10% higher AUC. Also the Densenet model generalizes with an average 8% higher AUC than hand engineer features model.

The generalization of the models was tested on a cross-continent pediatric ICU. We used 50% of pediatric cohort data to capture the population based probability thresholds for the models called Youdon-Index[18.] These threshold computation was done to account the population prevalence. Our model can be generalized by learning new population based probability thresholds. However the performance on adult ICU could be better given our models are trained on an adult population, the performance can be further increased by using our models and re-train them on the new settings.

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. All the data was 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 at least after total length of observation and prediction time. Also the training window or observational and lead-time data are free from any incidence of shock.

Methods:

1. Dataset description & preprocessing

We used "Medical Information Mart for Intensive Care (MIMIC)" data[19] and SafeICU data resource[11]. 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 ICU STAYS table of the clinical data of MIMIC-III database.

SafeICU (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[11], ethical permission were approved by the Institutional review board (IEC/NP-211/08.05.2015) was used for constructing a final validation across continent cohorts. SafeICU data collected between February 2016 to July 2019 were used to construct the Shock prediction cohort.

1.1 Imputation - Preprocessed matched subset ICU stays recordings and SafeICU ICU stays data were imputed using univariate "singular spectrum analysis (SSA)" using R package "Rssa[20]. Time-series are firstly embedded into a trajectory matrix, which is then decomposed into components. The components were grouped and the final time-series was reconstructed. The reconstruction process fills the gaps in the time-series[21]. Try and Catch Error codes were written in R programming language to construct a trajectory matrix at different window length so as to facilitate the decomposition. Often in the absence of arterial blood pressure (ABP). Non-invasive blood pressure (NIBP) is present. We used Unmatched dataset given in the MIMIC database to construct a linear mixed effects model to predict ABP from NIBP. The Matched subset data were imputed for ABP (Systolic and Diastolic) with the

corresponding predicted value from NIBP in case of missing values. The SafeICU ICU-stays were also imputed using SSA. Finally NIBP data was used to impute the missing ABP data with predicted values along with SSA.

1.2 Epoch and Cohort Generation:

We created 30 minute epochs for dense labeling 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 256 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 3, 4.5, 6, 7.5, 9 and 12 hour prior to the first occurrence of a 30-minute shock window was taken (Fig.5). 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.

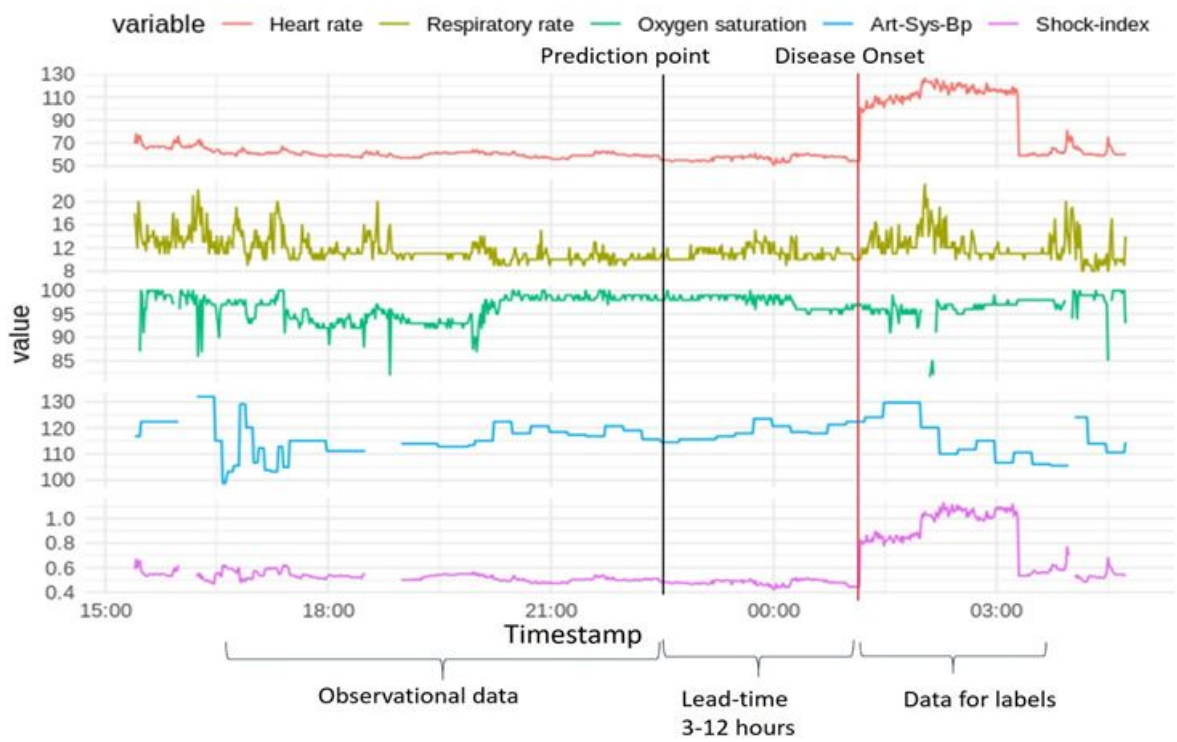

Figure 5: Time-series data is labeled for each 30 minute time window as Shock positive or negative, a 256 minutes of observational length window is used to train models to predict the shock status for next 3-12 hours of lead time (Red line)

Nonlinear times-series (NLTS) feature extraction and selection

Non-linear time-series features were extracted using the tsfresh python package[22]. This includes Wavelet transform coefficients, Fourier transform coefficients, discriminative power 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"[23]. Hyperparameters such as Number of trees and number of features were optimized using Out-of-bag error, validation AUPRC and validation AUROC as shown in the Supplementary Figure S2. Boruta was run for each n-features and n-tree combination. OOB, validation set's AUPRC and AUROC were recorded for each run. Finally a n-tree and n-features combination was selected. Boruta selected features are listed in Supplementary Table S1. The table of the description and variables found to be important were added to the Supplementary Table S2.

Modeling on Nonlinear times-series (NLTS) features (hand-engineered features):

Random-Forest models[24] were trained on the non-linear time-series features selected after running feature selection algorithm Boruta. Hyperparameters tuning was performed using grid search over number of trees and number of features for optimizing AUROC, AUPRC and out-of bag error.

Sequence models based approach:

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[25]. 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 used a state of the art MLSTM-FCN architecture[17] consisting of two branches first a LSTM/Attention block and fully connected CNN block of 128, 256 and 128 filters. The output of the two branches were concatenated which inputs a Dense layer of 2 neurons which returns probability upon application of Softmax activation function. Age and gender were added to the last layer feature extract to produce the final model using a logistic regression with L1 penalty to predict the future shock status.

Image based prediction model approach:

The times-series data can be converted to Hilbert[26] and Snake image representations. Since our data is in a form of multivariate time-series, individual time-series representations were aligned sample-wise into a volume to be directly fed to CNN. We used state-of-the-art Densenet architecture to build the Shock prediction model. Dense

Convolutional Network works in feed forward fashion to connect each layer to every other layer in the network. A L number of layer networks has L(L+1)/2 connections. Each layer takes feature maps of all the former layers as its inputs and it uses its own feature maps as input for all succeeding layers[16]. Reuse of features is another characteristic of Densenets, in addition it significantly reduces the number of parameters in usage and it can achieve state-of-the-art performance with less computation. DenseNets can scale naturally upto hundreds of layers, without causing any difficulties of optimization.

Densenet output features were extracted from the last layer and concatenated with age and gender. Concatenated features were transferred to a logistic regression model with L1 penalty to predict the next 3 to 12 hour shock status.

Model Development and Evaluation:

Mimic data cohort was used to train models to predict the next 3, 4.5, 6, 7.5, 9 and 12 hour shock status. From the Mimic cohort, raw signal data and non-linear time-series features were zero centered using global mean. The cohort was split into 60% training, 20% cross-validation, and 20% test sets. Same splits were used for all the models. To overcome the class-imbalance, only the training set was oversampled for the minority class. All the hyperparameters such as epoch numbers, batch size, number of trees and Youden-index were optimized on the validation set. Final results were reported on the test set.

Model generalization evaluation:

The SafeICU data was split into 50% training and 50% test sets. The SafeICU train set was used to learn Youden-index[18] specific to Pediatric Setting. The learned Youden-Index was used to compute model evaluation parameters on the SafeICU test-set.

Acknowledgements

This work was supported by the Wellcome Trust/DBT India Alliance Fellowship IA/CPHE/14/1/501504 awarded to Tavpritesh Sethi. 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

TS, AN designed the study. AN, PS, TS, RL were involved in acquiring data. AN developed the cohorts, AN, AK, AS were responsible for the Modeling and statistical analysis. AN, TS interpreted the data and wrote the first draft

of the report. TS, AN, PS, RL revised the report critically for important intellectual content. HB, PA, SY, AN, TS involved in the Web App development. All authors approved the final version of the manuscript. The corresponding author confirms to have had full access to the data in the study and final responsibility for the decision to submit for publication.

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 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.

Abbreviations: SI: Shock Index, CNN: Convolutional neural network, MLSTM-FCN: Multivariate Long short term memory - Fully connected network. AUROC: Area under the receiver operating characteristics, LFCN: MLSTM-FCN, NIBP- Non-invasive blood pressure.

Conflicts of interest

The authors declare that they have no competing interests.

Research involving human participants and/or animals

Ethics approval and consent for each institution to participate in the study was approved by the institutional research ethics committee in accordance with local ethical regulations.

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. 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).

6. Province, E. G.-D. T. C. G. of 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).

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

9. 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).

10. 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).

11. 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).

12. 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).

13. 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).

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

15. 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).

16. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* **2017-January**, 2261–2269 (Institute of Electrical and Electronics Engineers Inc., 2017).

17. Karim, F., Majumdar, S., Darabi, H. & Harford, S. Multivariate LSTM-FCNs for time series classification. *Neural Networks* **116**, 237–245 (2019).

18. Ruopp, M. D., Perkins, N. J., Whitcomb, B. W. & Schisterman, E. F. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical J.* **50**, 419–430 (2008).

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

20. Golyandina, N., Korobeynikov, A. & Zhigljavsky, A. SSA Analysis of One-Dimensional Time Series. in 31–120 (2018). doi:10.1007/978-3-662-57380-8_2

21. Ghodsi, M., Hassani, H., Rahmani, D. & Silva, E. S. Vector and recurrent singular spectrum analysis: which is better at forecasting? *J. Appl. Stat.* **45**, 1872–1899 (2018).

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

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

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

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

26. Yin, B., Balvert, M., Zambrano, D., Schönhuth, A. & Bohte, S. M. *AN IMAGE REPRESENTATION BASED CONVOLUTIONAL NETWORK FOR DNA CLASSIFICATION*. (2018).

Table1 | Mimic-III cohort characteristics

Variables	**3 Hours**		**4.5 Hours**		**6 Hours**		**7.5 Hours**		**9 Hours**		**12 Hours**	
	Non-Shock (n = 814)	**Shock (n = 1258)**	**Non-Shock (n =797)**	**Shock (n =1218)**	**Non-Shock (n =789)**	**Shock (n =1202)**	**Non-Shock (n =773)**	**Shock (n =1159)**	**Non-Shock (n =776)**	**Shock (n =1123)**	**Non-Shock (n =730)**	**Shock (n =1033)**
Arterial Systolic Blood pressure, mm Hg	**127.39(7.94)**	**122.18 (8.8)**	**127.37(8.2)**	**122.37 (8.98)**	**127.37(8.06)**	**122.28 (8.95)**	**127.2(8.12)**	**122.67 (9.12)**	**127.73(8.22)**	**122.74 (9.27)**	**127.6(8.43)**	**122.71(9.39)**
Arterial Diastolic Blood pressure, mm Hg	**62.77(3.57)**	**62.25 (4.61)**	**62.6(3.7)**	**62.1 (4.69)**	**62.55(3.49)**	**62.04 (4.69)**	**62.09(3.62)**	**62.11(4.76)**	**62.22(3.59)**	**62.06(4.81)**	**62.05(3.78)**	**61.9(4.81)**
Heart rate, per min	**67.53(4.03)**	**82.97 (4.67)**	**67.65(4.11)**	**82.53 (4.6)**	**67.69(4.07)**	**82.46(4.6)**	**67.66(4.06)**	**82.44(4.59)**	**68.11(4.15)**	**82.76(4.64)**	**68.63(4.16)**	**82.77(4.58)**
Respiratory rate, per min	**17.63(2.61)**	**18.77(2.82)**	**17.59(2.61)**	**18.63(2.8)**	**17.53(2.59)**	**18.6(2.81)**	**17.39(2.62)**	**18.62(2.82)**	**17.42(2.63)**	**18.67(2.82)**	**17.56(2.56)**	**18.67(2.84)**
Shock Index	**0.54(0.07)**	**0.7(0.08)**	**0.54(0.06)**	**0.69(0.08)**	**0.54(0.06)**	**0.69(0.07)**	**0.54(0.05)**	**0.69(0.07)**	**0.54(0.06)**	**0.69(0.07)**	**0.55(0.05)**	**0.69(0.07)**
Oxygen Saturation	**96.46 (1.48)**	**96.75(1.46)**	**96.64(1.45)**	**96.75 (1.46)**	**96.68 (1.37)**	**96.73(1.5)**	**96.69(1.35)**	**96.77(1.48)**	**96.74(1.38)**	**96.76(1.47)**	**96.74(1.42)**	**96.76(1.47)**

Gender (Female %)	43%	40%	41%	43%	40%	43%	41%	43%	40%	43%	40%	42%

Table2 | Safe-ICU cohort characteristics

Variables	3 Hours		4.5 Hours		6 Hours		7.5 Hours		9 Hours		12 Hours	
	Non-Shock (n = 117)	Shock (n = 47)	Non-Shock (n =48)	Shock (n =123)	Non-Shock (n =45)	Shock (n =127)	Non-Shock (n =43)	Shock (n =115)	Non-Shock (n =115)	Shock (n =43)	Non-Shock (n =116)	Shock (n =50)
Age (months)	63.59(60.17)	60.13(56.35)	64.39(59.41)	60.31(54.75)	59.56(57.6)	59.84(54.02)	66.26(60.3)	50.12(50.67)	71.32(61.62)	48.73(49.16)	67.05(58.82)	53.09(50.78)
Arterial Systolic Blood pressure, mm Hg	99.34(8.47)	93.1 (8.28)	54.8(6.89)	55.66(7.81)	101.9(8.02)	93.07(9.15)	101.39(8.49)	90(9.31)	101.5(8.18)	89.13(8.89)	101.66(8.48)	91.01(8.74)
Arterial Diastolic Blood pressure, mm Hg	56.37(6)	54.2 (5.57)	99.62(8.74)	92.95(9.91)	57.54(6.18)	56.03(7.27)	57.8(6.53)	53.45(7.07)	58.25(6.76)	52.52(6.63)	57.17(5.96)	52.77(6.19)
Heart rate, per min	125.51(8.14)	136.79(8.51)	128.29(7.77)	136.47(8.37)	123.03(7.96)	135.5(8.28)	126.33(6.98)	136.52(7.76)	126.37(8.29)	137.03(7.95)	126.58(7.99)	135.65(7.81)
Respiratory rate, per min	30.82(5.39)	32.69(5.37)	30.18(4.63)	31.54(4.89)	30.6(5.36)	32.48(6.07)	30.69(4.75)	33.02(5.49)	29.5(4.84)	33.75(5.58)	29.93(4.55)	33.09(5.46)
Shock Index	2.52(0.28)	1.63(0.22)	1.37(0.15)	1.62(0.32)	1.41(0.62)	1.58(0.31)	1.34(0.15)	1.64(0.32)	1.36(0.17)	1.65(0.28)	1.34(0.17)	1.65(0.21)

Oxygen Saturation	**94.29(2.7)**	**93.59(3.6)**	**94.87(2.83)**	**94.2(3.45)**	**95.19(2.45)**	**93.69(3.53)**	**94.75(2.65)**	**94.08(3.07)**	**94.8fs2(2.66)**	**93.24(3.2)**	**94.36(2.34)**	**93.1(2.84)**
Gender (Female %)	**32%**	**54%**	**38%**	**52%**	**33%**	**53%**	**35%**	**51%**	**34%**	**49%**	**39%**	**44%**

Values are mean (SD) unless indicated

Model	Deep-learning						Non-Deep-Learning					
Prediction-Time(Hours)	12	9	7.5	6	4.5	3	12	9	7.5	6	4.5	3
MIMIC-AUC (%)	0.77	0.81	0.81	0.8	0.83	0.81	0.78	0.77	0.83	0.84	0.83	0.84
SAFEICU-AUC (%)	0.66	0.69	0.63	0.63	0.7	0.69	0.57	0.53	0.5	0.45	0.53	0.59
Generalisation-Drop (%)	-15.74	-15.36	-19.11	-23	-18.22	-17.01	-26.92	-31.17	-39.76	-46.43	-36.14	-29.76

Table 3- Comparison of Generalization drop in AUC between deep-learning and non-deep learning models, when models trained on MIMIC (USA) data were tested on Safe-ICU (India) data.