Early Prediction of Hemodynamic Shock in the ICU with Deep Learning on Thermal Videos

Vanshika Vats[1§], Pradeep Singh[1§], Aditya Nagori[2,3§], Raman Dutt[4], Harsh Bandhey[1], Mahika Wason[1], Rakesh Lodha[5], Tavpritesh Sethi[1,4*]

[1]Indraprastha Institute of Information Technology, Delhi-110020, 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]Shiv Nadar University, Tehsil Dadri, Greater Noida, Uttar Pradesh-201314, India
[5]All India Institute of Medical Sciences, Department of Pediatrics, New Delhi-110029, India
[§]These authors contributed equally *Corresponding Author

Abstract

Shock is one of the major killers in ICUs and early interventions can potentially reverse it. In this study, we advance a non-contact thermal imaging modality to continuous monitoring of hemodynamic shock working on 406 patient videos of 256 seconds length for 22 patients longitudinally. Deep learning was performed upon these videos to extract Center-to-Peripheral Difference (CPD) in temperature values. CPD along with heart rate, was finally analysed to predict the shock status up to next 12 hours using Long-Short Term Memory models. Our models achieved best area under the receiver-operating-characteristics curve of 0.81 ± 0.06 and area under precision-recall curve of 0.78 ± 0.05 at 5 hours, providing sufficient time to stabilize the patient. Our approach, thus, provides a reliable prediction using an automated decision pipeline, that can save lives and provide better care.

Introduction

Hemodynamic shock is characterized by inadequate supply of oxygen and nutrients to the cells than what is essentially required. Such a condition could cause tissue malfunction leading to a rapid organ failure which can eventually result in death. Although reversible in the initial stages, the mortality rate due to delay in detection and treatment of such shocks are as high as 34% in the ICU patients admitted in the developing countries[1]. Moreover, the under-privileged areas have a higher risk of shock ignorance mainly due to a lack of technological advancement and a low doctor-to-patient ratio. An early detection with effective management can prove to reverse the effects, saving the patient from rapid organ failure[2-4]. Most of the methods dealing with the shock in today's date are invasive or require repeated contact, making the patient prone to hospital-acquired infections. The non-invasive ways such as Non-Invasive Blood Pressure monitoring and Ultrasonography are non-continuous but require contact with the delicate skin of infants which are infection-prone too[5]. Thus, there is a need for having a non-contact and non-invasive modality that can prove to be effective for shock management.

The most recent minimally invasive way to get sufficient data is to work on the thermal images[6]. The studies have found out that the possibility of shock can be determined using the temperature difference observed between the abdomen and foot of the patient (Centre-to-Peripheral Difference)[7-8]. The feature has been exploited along with some vitals and the use of machine learning methods such as Histogram of Oriented Gradients features with Random Forest classifier for detection and prediction of shock with a single snapshot of an image at a time point[9]. But a single instance might not be able to give robust results because of limited information. Plus, using the handcrafted features for machine learning algorithms might cost us time. To expand upon the work, we aim to reduce the manual tasks performed to calculate the features using deep learning methods and work on the continuous time series data, improving the efficacy of detection and prediction. This will be helpful in increasing the time window of action and

treatment rather than spending the time on diagnosis. The impact this problem creates and the scope of extending the solution makes this study worthwhile in saving the lives of many.

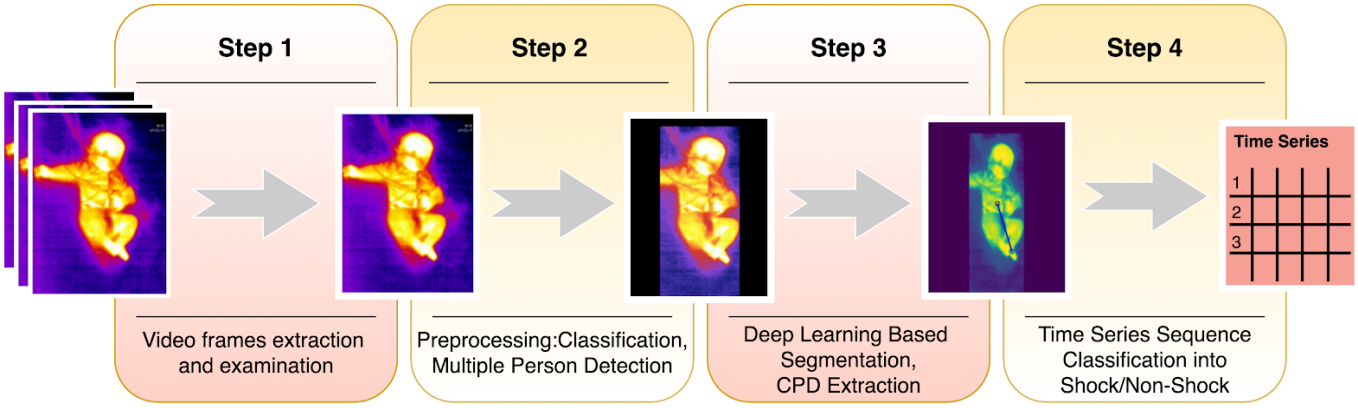


Figure 1: Shock Prediction Steps. The summary of the shock pipeline shows the steps from video frame extraction to shock prediction. Step 1 comprises sampling videos to extract frames. Step 2 classifies frames into covered or uncovered, while also finding the presence of multiple people in the frame and mask them, so as to avoid confusion. The masked images are then input to the ResUNet based segmentation model and CPD is hence extracted. The series of CPDs are then passed through a time series sequence classifier and finally the predictions are made for shock for the next 12 hrs.

Results

Patient Characteristics. Statistical Inferences of the cohort characteristics between the shock and non-shock groups are depicted in Table 1. Shock and Non-shock conditions were decided using Shock Index Pediatric Age-adjusted calculation. It can be observed that the most significant difference between the two categories is in the heart rate, as expected, along with the respiratory rate.

Table 1: Cohort characteristics and statistical significance of control (non-shock) vs affected (shock) classes. The p-values were calculated using either Wilcoxon rank sum test (W) or Student's t-test (t) after testing for normality by D'Agostino-Pearson normality test. (n - Number of sequences, IQR - Interquartile Range)

Variable	Non-shock Seq n = 274	Shock Seq n = 132	Statistical Tests
	Median (IQR)	Median (IQR)	p-value (W/t)
Age (months)	54.44 (59.43)	75.89 (107.02)	0.6087 (W)
Arterial Systolic Blood pressure, mm Hg	130.75 (11.95)	128.09 (25.21)	0.0014 (t)
Systolic Blood pressure, mm Hg	106.00 (12.00)	102.00 (5.00)	0.002 (W)
Heart rate, per min	111.52 (20.27)	143.17 (63.81)	< 0.0001 (t)
Respiratory rate, per min	25.85 (9.66)	22.87 (13.20)	< 0.0001 (W)
Oxygen saturation (SpO2)%	97.86 (2.62)	98.32 (3.45)	0.9932 (W)

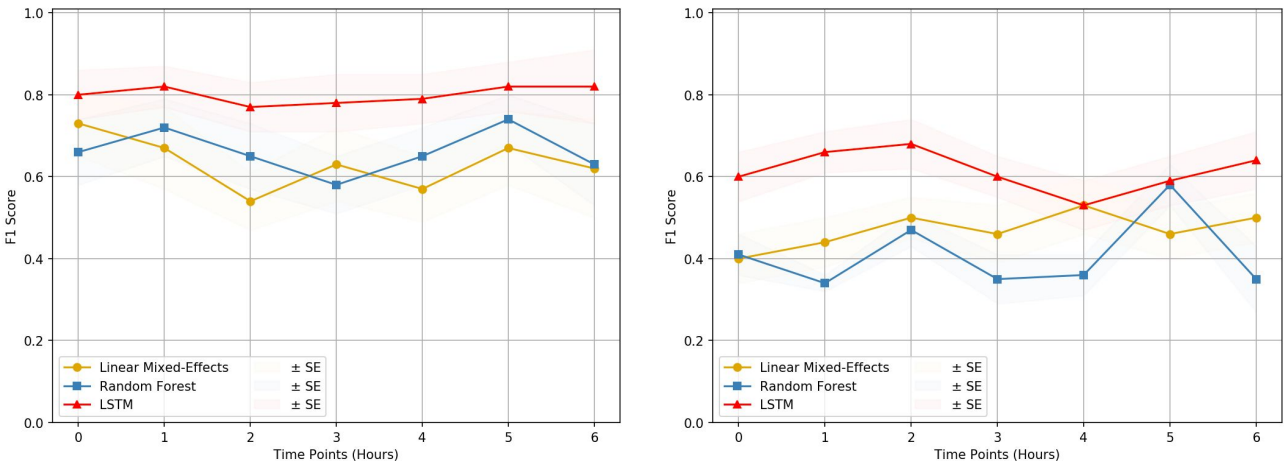
Segmentation of abdomen and feet achieved a total dice loss of 0.0391 using ResUNet

Since thermal images lack texture, it is important for the model to recognise the structural shape features. ResUNet was specifically used to make the task possible on relatively less available thermal data, and to capture both the local and global information from the image. A dice loss of 0.0391 (Dice coefficient=0.9609) with a Binary Cross Entropy loss of 0.0692 was achieved for segmenting out abdomen, feet and the background. The mode intensities of the segmented areas were then used to find out the Central-to-Peripheral Difference and hence build the longitudinal models from continuous long-duration videos.

LSTM model was found to be the best performing

We compared three models to finally arrive at the best performing one. Linear-Mixed Effects[10], Random Forest[11] and LSTM[12] were tested at various time points from the observation taken. Based on various metric evaluations, LSTM was found out to be best performing on our given time series data, and hence

was chosen as the primary choice for our study. The F1 score comparison of the three models is shown in Figure 2, and the rest of the evaluations are shown in the Supplementary Extended data Figure 4.



(a) **(b)**

Figure 2: Comparison of Model F1 Scores for CPD and combination. The F1 score comparison for the three models (i.e. Linear Mixed-Effects model, Random Forest, and LSTM) tested on **(a)** CPD+HR and **(b)** CPD parameter only. It can be observed that the LSTM model outperforms the other two models in both the cases, making it the primary choice of this research. Rest of the comparison plots are depicted in supplementary material.

Shock detection at 0 hr using LSTM model

The CPD was extracted from every uncovered window possible using the segmented out abdomen and feet regions in a continuous way with videos sampled at 1fps. Keeping in consideration the Unique Health IDs of patients and then performing SMOTE[13] upsampling, the observation windows of 256s (i.e. 4.26 min) were passed into LSTM networks, along with heart rate as an additional covariate, for sequence classification. 10-fold stratified cross-validation gave a mean AUPRC of 0.796 and mean AUROC of 0.788.

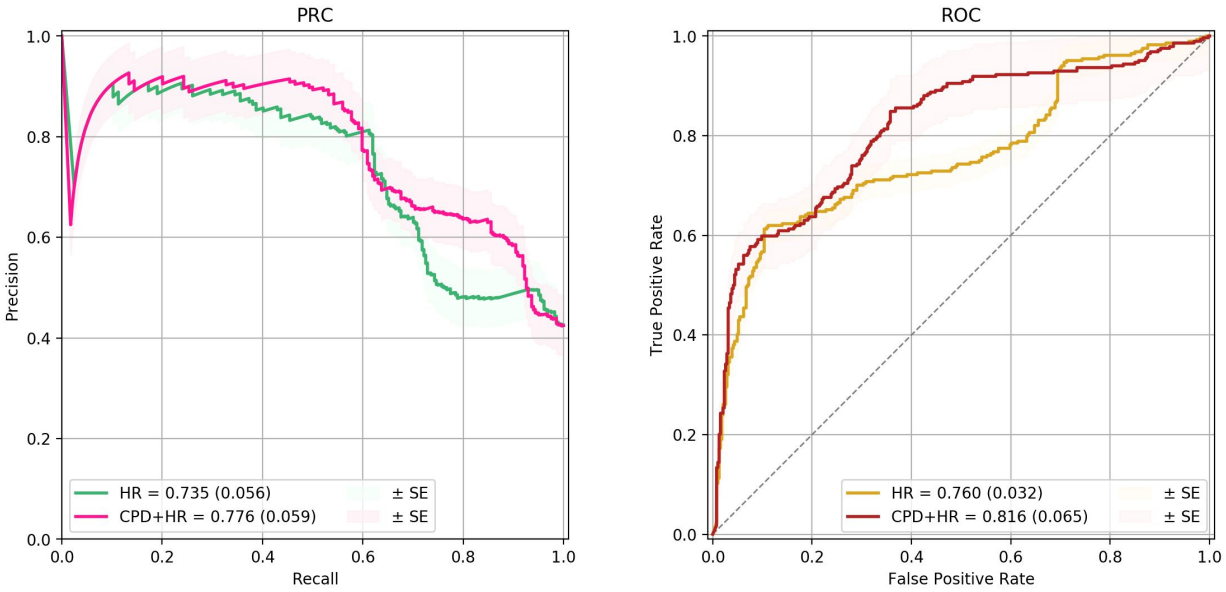
Evaluation of LSTM for the prediction of hemodynamic shock up to 12hr reveals that CPD increases the model performance

The SAFE-ICU resources allowed us to match the time sequences with their corresponding states of shock/non-shock for the next 12hr since the observation was taken. We tested the LSTM model performance for 1 to 12 hour of lead time on heart-rate and CPD combination as shown in Table 2. The evaluations with only heart-rate feature is shown in Supplementary Extended data Table 2. The results showed a continuous trend of precedence of using CPD with heart rate over using the heart-rate alone, on multiple metrics such as AUPRC, AUROC etc till 6 hours, which gives a good time window to alert the medical practitioners for the advent of shock.

Time Pt.	S/NS	AUPRC	AUROC	Accuracy	Sensitivity	Specificity	PPV	NPV	Youden
		Mean (SE)							Mean
0hr (D)	132, 274	0.79 (0.06)	0.78 (0.05)	0.85 (0.03)	0.74 (0.06)	0.92 (0.02)	0.86 (0.05)	0.84 (0.03)	0.50
1hr (P)	115, 271	0.71 (0.06)	0.76 (0.04)	0.83 (0.04)	0.83 (0.04)	0.82 (0.06)	0.80 (0.06)	0.88 (0.02)	0.42
2hr (P)	125, 253	0.56 (0.05)	0.69 (0.04)	0.83 (0.02)	0.72 (0.06)	0.90 (0.03)	0.82 (0.05)	0.84 (0.04)	0.56
3hr (P)	133, 232	0.67 (0.06)	0.74 (0.06)	0.83 (0.04)	0.69 (0.08)	0.93 (0.04)	0.88 (0.06)	0.82 (0.04)	0.57
4hr (P)	123, 242	0.75 (0.05)	0.75 (0.05)	0.81 (0.04)	0.70 (0.06)	0.94 (0.03)	0.90 (0.05)	0.78 (0.06)	0.64
5hr (P)	120, 247	0.78 (0.05)	0.81 (0.06)	0.84 (0.04)	0.76 (0.07)	0.94 (0.02)	0.88 (0.05)	0.84 (0.05)	0.62
6hr (P)	124, 228	0.66 (0.10)	0.73 (0.06)	0.89 (0.03)	0.82 (0.08)	0.92 (0.04)	0.81 (0.09)	0.95 (0.02)	0.62
7hr (P)	116, 211	0.66 (0.06)	0.78 (0.04)	0.79 (0.04)	0.81 (0.05)	0.77 (0.05)	0.73 (0.06)	0.87 (0.04)	0.58
8hr (P)	85, 216	0.52 (0.06)	0.72 (0.06)	0.80 (0.05)	0.82 (0.06)	0.80 (0.09)	0.79 (0.06)	0.89 (0.04)	0.69
9hr (P)	90, 211	0.65 (0.06)	0.77 (0.05)	0.82 (0.03)	0.81 (0.08)	0.84 (0.04)	0.77 (0.06)	0.87 (0.04)	0.51
10hr (P)	77, 215	0.64 (0.07)	0.79 (0.04)	0.82 (0.04)	0.89 (0.05)	0.78 (0.06)	0.73 (0.07)	0.95 (0.01)	0.49
11hr (P)	62, 226	0.55 (0.05)	0.79 (0.03)	0.88 (0.03)	0.84 (0.05)	0.88 (0.03)	0.74 (0.05)	0.94 (0.01)	0.58
12hr (P)	78, 205	0.54 (0.08)	0.73 (0.05)	0.82 (0.04)	0.75 (0.08)	0.83 (0.07)	0.72 (0.06)	0.91 (0.03)	0.50

Table 2: Performance of the proposed model predicting the presence of Shock/Non-shock using automated CPD and HR. The Time Pt. column depicts the subsequent hours from the time of taking the observation, at which the results were recorded. The unequal number of shock and non-shock sequences is due to the absence of patient data with the increasing number of hours. (S/NS - Number of Shock/Non-Shock sequences present, AUPRC - Area

Under Precision Recall Curve, AUROC - Area Under Receiver Operating Characteristics, PPV - Positive predictive value, NPV - Negative predictive value, D - Detection, P - Prediction)



(a) **(b)**

Figure 2: Quality Assessment for 5 hr Prediction. Quality evaluation of the LSTM time series classification models by the **(a)** AUPRC and **(b)** AUROC for 5 hour prediction. The rest of the intermediate time stamps results till 12 hr are shown in the supplementary material. The results of all are shown in Table 2. The Standard Error (SE) for each is calculated from cross validation, by taking k=10.

Discussion:

The study presents a deep learning based continuous and non-contact shock detection and prediction model which leverages Center-to-Peripheral Difference (CPD) as one of its main parameters. Since hemodynamic shock can lead to organ failure and eventually to death in the ICU, its prediction on time can save lives. Care needs to be taken to evaluate the parameters in a non-invasive way such as not to cause the infections by the contact. Since, the field of thermal image inspection has only started to be explored, it can thus be leveraged, along with some non-invasively monitored vital parameters, for a prediction of shock. The use of deep learning methods prove to be really beneficial in reducing the manual preprocessing and increase the accuracy of the methods.

In this study, we have extended our previous work on hemodynamic shock prediction with longitudinal continuous monitoring of body thermal patterns which are previously found to be predictive for future shock prediction[9]. The longitudinal monitoring of temperature gradient opens up a rich source of information about patients physiology. The underlying stochastic patterns can have a discriminative value for future hemodynamic shock risk. We leveraged these reasonings to extract the center to peripheral intensity difference from the thermal images in time-series fashion. We do so by applying our data specific trained filters for multiple person detection, cover and uncovered patient detection models which is followed by segmenting the body parts into Abdomen and foot using artificial intelligence based models called UNet. Percentage difference of segmented body parts, identified as abdomen and foot were taken, termed as Centre to Peripheral Difference (CPD). These segmentation models were trained on the images collected during May-September 2016 and February-April 2017[6,9]. The extracted CPD time-series along with vitals time-series were used to predict the future (1-12 hr) hemodynamic shock using sequence models called Long-Short Term Memory (LSTM). This model was also compared with Random Forest and Linear Mixed-Effects models, known for working with longitudinal data. It was observed that the LSTM model outperformed both the other two. We also trained and checked the performance of models which can classify the length of video and images for future risk of hemodynamic shock. The class of algorithms in deep learning called convolutional neural networks (CNNs) which have

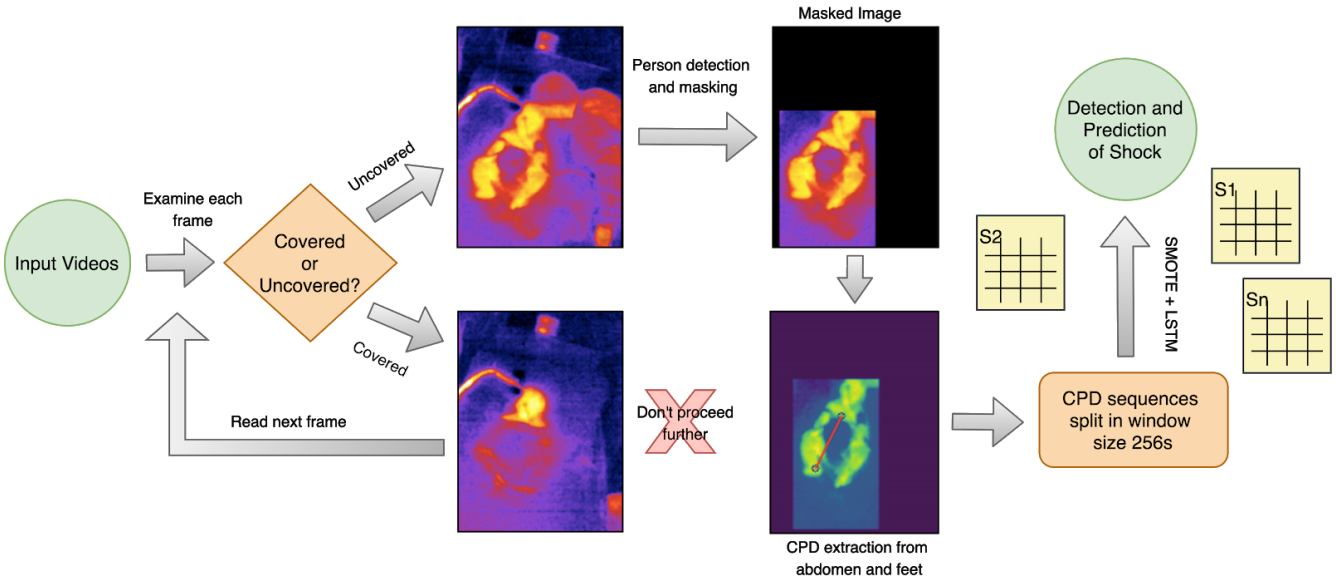


Figure 4: Shock Prediction Pipeline. Illustration of the pipeline followed for the detection and prediction of shock and no-shock. Each frame of the video was examined for uncovered/covered. The uncovered frames were then filtered off the presence of people other than the patient present in the frame. The frame was finally passed from the segmentation and CPD extraction model, further collecting the sequences used for LSTM time series classification. The time sequences had CPD and HR as features and an appropriate window length of 256s was chosen. Since the data was highly imbalanced, SMOTE upsampling method was used in training the LSTM model. The detection and prediction of shock was done at 0hr and for the next 12 hours respectively.

shown state-of-the-art performance for image classification were used for the task. However, we found the domain features such CPD, performed better than the classification based direct image and video classification for future risk of shock. Even after many experiments performed for a direct classification of images using the concepts of TV Chambolle[14] denoising, data augmentation and undersampling/ oversampling, we were able to get the best AUROC of only 0.60 for shock detection using ResNet-50[15]. This can be because of the cluttered background with diapers and tubes, and an increased region of interest for the information extraction from a limited variety of images. This also tells the importance of the domain specific features over automated CNN based approaches. We built the models for multiple time-points while comparing only Vitals models (as baseline) versus models with CPD and vitals as predictors. The CPD and vitals models out-performed the only vitals model upto 6 hours of prediction. Later than this, their performance came out to be lower than the vitals based models. This might be because of the less extensive dataset as the time point increases as the number of patients having their vitals for long hours present with the hospital, reduce. This might be because of patient discharge from the ICU or the failure of data procurement by the vital instruments. The metric AUPRC and F1 score are the most significant for an imbalanced dataset such as ours as it doesn't get biased by the presence of true negatives and thus gives a clearer perspective of a classifier's utility. The results till 6 hours show a

promising window with a good prediction rate which can prove to be really helpful for the doctors to help find a buffer time prior to the shock event to start the treatment.

However, a few limitations can come up as monitoring thermal patterns could be hard at times when the patient is being covered or being operated; caregivers can also block the recordings. But, these issues can be resolved by keeping ~4.3 minutes of uncovered slots for the monitoring. Our models only require as minimal as ~4.3 minutes of thermal recordings which does not recognise or identify an individual's identity, hence is safe for patient privacy and does not take much time away from the time assigned for care. The results can be made more robust by expanding the dataset by including more number of patients in the study. Also, the dataset can be made varied and generalised by including observations from multiple other clinical sites. But the study performed illustrates a great non-invasive and a minimal feature architecture which promises to be a life-saver by informing the clinicians about shock well in advance.

Methods

Cohorts and study design:

- **Safe ICU Framework:** All the data reported in this research was collected from Paediatric ICU of AIIMS, a tertiary care hospital of India. Ethics committee of the medical institute granted a consent (Ref. No. IEC/NP-211/08.05.2015, AA-2/09.02.2017), as there was no alteration required in patient care nor the patient's personal details required during the process of data warehousing. Segregation was not done during the data analytics and thus no data was discarded in the process. There were 8 beds including neonatal beds in the Paediatric ICU. The four underlying principles constitute the Sepsis Advanced Forecasting Engine for the ICUs framework. Principles comprising Capture Reliably, Approach Systemically, Phenotype deeply and enable decisions.

All the admitted (enrolled) patients' vital data collection started from February 2016, was limited to comprise patients who had Arterial Blood Pressure recordings(analyzed).

- **Vital records from multi-parameter monitoring:** For monitoring the patients, Paediatric ICUs are equipped with Mindray TM monitors. In-house software was written for Health Level 7 (HL7) Standards[16] based querying of the Central Monitoring Station (CMS). 1 Terabyte portable hard disk embedded with Raspberry Pi was used to store the streaming data. In order to query and to receive vitals data, Client socket programming with respect to the device protocols were used. These vital data were received at the resolutions of 15 second for unsolicited data and 1 second for the real time data. To receive streaming data, 64*1024 bytes of character array was used as a buffer. Pipe delimited text file was generated every day at 0000 hr. Software code was written in such a way that it automatically logs the data into a text file on daily base. In order to identify the lossy data, alarm messages were sent to the android mobile phones. That is whenever there is an interruption in data streaming, Pushbullet™ and RPushbullet generate the alarm messages. These high-resolution vital data have been warehoused for all the ICU patients starting from February 2016 to May 2020.
- **Treatment Charts :** Proper notes of the treatment were entered by the in-house doctors in word documents and the backup of these word files were scheduled at fixed hours every day, say 1700 hours and 1900 hours. Docx2txt python module is used to parse these text files and then it is converted to tabular format for text mining.
- **Cohort based on Binary shock index:** The SAFE-ICU described earlier has warehoused over 3,00,000 patient-hours of monitoring data from the PICU. It is used to extract time-stamped vital data for the patient's at 0-12hr heart rate and blood pressure recordings. Shock index was calculated as the ratio of median heart rate and median non-invasive blood pressure or the arterial systolic blood pressure. This was calculated over the median of moving sequential windows of 30

data points at a resolution of 15s. Shock-index Paediatric Age-Adjusted (SIPA) is used to compute shock/no-shock age specific binarized outcome for each patient [17].

- **Laboratory Investigations:** PostgreSQL database (2013) is the AIIMS hospital database, used to store all the laboratory investigations. Queries are written to retrieve the required data of the children admitted to the ICU.

Thermal Imaging: A standard thermal video capturing and operating procedures were followed, in order to ensure that there is less effect by extraneous factor, say, patient positioning, device handling etc (Supplementary Methods S1).

Thermal cameras only capture infrared radiation, so as to make sure that study does not reveal the patient's identity. Camera was placed properly and at a good distance from the patient, so that there was no direct contact involved nor any change in patient routine care. The thermal videos were captured in a standard color-scale guaranteeing that the full body of infants were visible.

A total of 130 infants and children (male, female), aged between 0.2-204 months, who were admitted and had the arterial line recordings were considered in this study. Total of 450 thermal videos were recorded using Android Smartphone attachment (Seek Thermal®). 76 (52 patients) out of 450 were considered were filtered out from the rest noisy videos with thermal noise and a few other external factors. Out of these, 44 thermal videos from 30 children (male, female) were chosen based on the availability of their abdomen and feet uncovered. 16 out of 44 videos were of a duration 1 min or less. Since the project looks at long windows of longitudinal data, only 26 long videos (2+ hours; 7200s or more), with their vitals present, remain consisting of 22 unique patients.

Thermal videos of every single patient were collected at different time points on different days. Thus each patient possibly has different values for shock-status, which in turn eliminates bias due to the patient's propensity characteristics, say gender, age etc. Vital data with respect to the time-stamp of video were

extracted from the data warehouse at 15s intervals (SAFE ICU)[6]. A comparison was made between the shock and non-shock groups using either Wilcoxon rank sum test or two-tailed Student's t-test, after testing for normality by D'Agostino-Pearson normality test using GraphPad Prism version 6.00, GraphPad Software, La Jolla California USA, www.graphpad.com.

Classification into Covered and Uncovered: The patients in ICU are kept under observation for a long duration. Since it is a very critical area, the patients are kept covered by a blanket most of the time. The blankets are removed for a short period of time generally only when a nurse or a doctor comes to examine the patient or to change the vitals and medicines. Since the main hypothesis of this research is to work with the Center-to-Peripheral-Difference parameter, there are only a few short windows in a few hours per video which can be useful. To train the data, images were augmented and normalized by their mean and variance especially extracted out to suit the thermal data. A ResNet-152 architecture was trained using PyTorch[18] framework in Python3[19] to classify each frame into covered and uncovered, i.e. abdomen and feet are visible. The model was finally implemented on videos sampled at 1fps.

Multiple Person Detection: Since the patients are in intensive care, they are barely left alone, especially when uncovered, because that is when the caretaker tends to their needs. For the CPD extraction task, there is a need to filter out the presence of this additional person, so as not to confuse the algorithm between the person and the patient. A variety of images of the patient alone and along with the person/caretaker was taken and augmented. Instead of forming two separate classes of a child and a person, the results were found to be better on using only one class as a person, which could detect multiple instances of the person in an image. This might be because the algorithm was not able to grasp any particular form/structure of the person, as different body parts (e.g. a hand in some, head in another) were present in different images. The frames could just have been discarded but a few videos in the dataset

contained the presence of the caretaker throughout the duration for which they were captured. Thus, to avoid the loss of such crucial data, a tradeoff was done. Assuming that in most of the cases, the area covered by the additional person in the frame is comparatively less than the area covered by the patient, the bounding box with the larger area, presumably the patient, was kept and the rest was masked. The now visible area could be further used for CPD extraction. This masking was done using YOLOv3[20] in PyTorch having DarkNet-53 as its base architecture. The thermal images were manually annotated and trained to get the best checkpoint weights suitable for multiple person detection.

Segmentation and CPD Extraction from Abdomen and Feet: Nagori A. et al [9] proved that the probability of shock depends directly on Center-to-Peripheral Distance (CPD). For this study, the abdomen has been taken as the center and peripheral is taken to be the foot. The images were annotated manually and pixelwise using js-annotator-tool. The target maps contained 3 one-hot encoded layers corresponding to abdomen, feet and background. The input images were normalized with the mean and variance especially extracted from the distribution of the dataset in use. Appropriate image padding was done to ensure the aspect ratio of the images remains the same in case of any change in input dimension. To account for a low dataset of pixel wise segmented images for training, a ResUNet with ResNet-18, pretrained on ImageNet, was used as an encoder. UNet[21], being specifically introduced to segment the less abundantly found medical data, helps to gather more local and global information even in the dearth of data, and thus efficiently segmenting out the images. The skip connections from the encoder to the decoder helps the model to keep the original pixels at that particular scale in consideration while recreating at the decoder and thus learning finer details efficiently. Smaller skip connections in ResNet-18 encoder helps to deal with the problem of vanishing gradients and thus makes the learning more efficient. A cutoff threshold was set on the predicted outputs to remove any weakly predicted pixels. The area detected was used as a region of interest in the original image and the mode of the detected probabilities

was taken as the point of temperature extraction from the segmented out abdomen and feet. The difference was divided by the abdomen value to keep CPD robust from the thermal noise.

$$\text{Difference percent} = \frac{(Abdomen\ Intensity - Foot\ Intensity)}{Abdomen\ Intensity} \times 100$$

LSTM Time Series Sequence Classification: The videos were sampled at 1 fps to extract the CPD data from every uncovered window possible. Windows of 256 data points corresponding to 256s (4.26 min, padded, if necessary) were taken as an input to the LSTM based classifier. The windows less than 256 are padded with 0s and the windows greater than 256 are split in an overlapping fashion, when necessary. Each CPD, along with the heart rate at its corresponding time point was taken to finally label it with the shock index, and hence the presence of shock/non-shock. The missing heart rate data at certain points was imputed with linear interpolation if the missing data was less than 10% of the time series length. Since the data is highly imbalanced with more non-shock sequences, the training data is augmented with the SMOTE oversampling method. The LSTM sequence classifier was followed by a series of dense layers with a dropout of 0.2, which then passed through a sigmoid layer to output the binary shock index, and hence, the occurrence of shock/no-shock.

Linear Mixed-Effects and Random Forest sequence classification on tsfresh features: The tsfresh[22] features were extracted from 256-length sequences and trained on the same train and validation distributions as the previous LSTM model. 'Boruta'[23] package from R-language was used for this purpose. Variation Inflation Factor (VIF) is used to reduce multicollinearity in data. If the VIF value exceeds 10, then the collinearity is considered problematic and hence that particular variable causing it should be removed. The remaining features were used to train the linear mixed-effects and random forest models.

Direct Classification of thermal images/videos for future risk of shock: Apart from CPD extraction, an attempt was made to classify into shock/no-shock by directly giving the whole images/videos as the input. In one direction, we tried to classify each video frame read at a time and conducted experiments with several modern architectures based on convolutional neural networks (CNN). The concepts of TV Chambolle denoising, data augmentation and undersampling/oversampling, were used to get the best shock detection AUROC of 0.60 using ResNet-50. Also, the information extracted from a single image frame can be very limited. So instead, we tried to use direct and continuous video samples of length 256s as an input to a conjunction of various CNN and LSTM models, trained in a time distributed manner. Being a fundamental extension of the direct image classification problem, it suffered from similar limitations.

Outcome variable - Binary shock index: The SAFE-ICU initiative has enabled this research to gather the PICU data and extract the vitals and the corresponding time stamps at the 0th hour (time of video capturing) and at the next 12 hours. Shock index was taken as the median heart rate and median non-invasive blood pressure or the arterial systolic blood pressure, for moving sequential windows of 30 data points at a resolution of 15s. Shock-index Paediatric Age-Adjusted (SIPA) was then used to compute the age-specific binary outcome for each patient.

Time Points: The time at which the video was captured was taken as the 0th hour, and the predictions of shock/no-shock were performed for the next 12 hours.

Model Evaluation: The video data was first partitioned patient-wise such as to keep train, validation and test sets unseen from each other. For the 10-fold cross-validation, the data was partitioned with the ratio of 60:20:20 into these three sets in a stratified manner, i.e. keeping the distribution of low-percentage shock class comparable in all three sets. The training data was augmented for the low-found shock class

using SMOTE oversampling method; the validation and test sets remain unchanged of their size in each respective fold. The model analysis was mostly done on the Area Under Precision Recall Curve (AUPRC) and Area under Receiver Operating Characteristic (AUROC) curve. Other standard metrics like F1-score, PPV, NPV, Specificity and Sensitivity, were evaluated at the Youden's Index (J)[24]. Since, there is a high significance of prevalence in the medical domain, calculating the metrics at Youden's Index becomes important.

$$J = max\ (\ \text{Sensitivity}(c) + \text{Specificity}(c) - 1\ )$$

Data Availability: The data is available on a reasonable request from the corresponding author.

Conflicts of interest: The authors declare that they have no competing interests.

Acknowledgements

This work was supported by the Wellcome Trust/DBT India Alliance Fellowship IA/CPHE/14/1/501504 awarded to Tavpritesh Sethi. Tavpritesh Sethi and Pradeep Singh also acknowledge the support received from the Department of Science and Technology vide project DST/INT/ISR/P-21/2017. We also thank Mr. Varun Prakash and Mr. Anil Sharma for the technical support provided at PICU, AIIMS, New Delhi.

Author contributions

Concept & Design: TS, RL, AN, PS. Standard thermal video data acquisition: PS, TS, RL. Vital data acquisition: AN, PS, TS, RL. Cohort construction using vital data and thermal videos: PS AN. Segmentation Models & Extraction of CPD: VV. Multiple person detection & covered-uncovered detection: VV. LSTM, Models & Evaluation: VV, AN, HB. Mixed-effects and Random forest, Models & Evaluation: PS, AN, VV. Statistical analysis: VV AN. Direct classification models: RD, HB, MW. Interpretation of Results: VV, AN, PS, TS. Manuscript & Revision: VV, PS, AN, RL, TS. Final version of the manuscript was approved by all the authors involved in this research.

References:

1. Divatia, J. V. et al. Intensive Care in India: The Indian Intensive Care Case Mix and Practice Patterns Study. *Indian Journal of Critical Care Medicine* **20**, 216–25 (2016).
2. Vincent, J.-L. & De Backer, D. Circulatory shock. *The New England Journal of Medicine* **369**, 1726–34 (2013).
3. Rivers, E. et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *The New England Journal of Medicine* **345**, 1368–1377 (2001).
4. Early Goal-Directed Therapy Collaborative Group of Zhejiang Province et al.. 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 (2010).
5. Oranges, T., Dini, V. & Romanelli, M. Skin Physiology of the Neonate and Infant: Clinical Implications. *Advances in Wound Care* **4**, 587–595 (2015).
6. Sethi, T. et al. Validating the Tele-diagnostic Potential of Affordable Thermography in a Big-data Data-enabled ICU. In *Proceedings of the Special Collection on eGovernment Innovations in India*, 64–69 (ACM, 2017).
7. Houwink, A. P. I., Rijkenberg, S., Bosman, R. J. & van der Voort, P. H. J. The association between lactate, mean arterial pressure, central venous oxygen saturation and peripheral temperature and mortality in severe sepsis: a retrospective cohort analysis. *Critical care (London, England)* **20**, 56 (2016).
8. Bourcier, S. et al. Toe-to-room temperature gradient correlates with tissue perfusion and predicts outcome in selected critically ill patients with severe infections. *Annals of Intensive Care* **6** (2016).

9. Nagori, A. et al. Predicting Hemodynamic Shock from Thermal Images using Machine Learning. *Sci Rep* **9**, 91 (2019). https://doi.org/10.1038/s41598-018-36586-8
10. Bates, D. M., Mächler, M., Bolker, B. M., and Walker, S. C. lme4: Linear Mixed-Effects Models Using Eigen and S4. R Package Version 1.**1-7**. (2014a) Available online at: http://CRAN.R-project.org/package=lme4.
11. Cutler, A., & Cutler, D., & Stevens, J., Random Forests. 10.1007/978-1-4419-9326-7_5 (2011).
12. Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural Computation* **9**, 1735–1780 (1997).
13. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res*. **16**, 321–357 (2002).
14. L. Condat A Direct Algorithm for 1-D Total Variation Denoising. In *IEEE Signal Processing Letters* **20**, 1054-1057 (2013), doi: 10.1109/LSP.2013.2278339.
15. He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV (2016), 770-778, doi: 10.1109/CVPR.2016.90
16. Benson, T. and Grieve, G. Principles of health interoperability: SNOMED CT, HL7 and FHIR (2016) 10.1007/978-3-319-30370-3.
17. 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).
18. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* **32**, 8024–8035, (Curran Associates, Inc. 2019).
19. Python Software Foundation. Python Programming Language-Official Website. http://www.python.org/ (2018).

20. Redmon, J., and Farhadi, A. YOLOv3: An Incremental Improvement. *CoRR* (2018)

21. Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR* (2015).

22. Christ, M., Kempa-Liehr, A.W. and Feindt, M. Distributed and parallel time series feature extraction for industrial big data applications. *ArXiv* e-prints: 1610.07717 (2016).

23. Kursa, M & Rudnicki, W. Feature Selection with Boruta Package. *Journal of Statistical Software* **36**. 1-13. 10.18637/jss.v036.i11 (2010).

24. 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 Journal* **50**, 419–430 (2008).

25. Hanley, J. A. and McNeil, B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**, 839–843 (1983).

26. Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* **10**, 1895–1923 (1998)

27. Bradski, G., The OpenCV Library. *Dr. Dobb's Journal of Software Tools* **3** (2000).

28. Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous systems. *CoRR* (2015).

29. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

30. Kovac, Z. and Belina, D. The pathophysiology of hemodynamic shock syndrome (part one). *Lijec Vjesn* **120**, 379–392, (1998).

31. Forman, G. and Scholz, M. Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement. *Association for Computing Machinery* **12**, 49–57 (2010).