

PREDICTING SHOCK INDEX IN PEDIATRIC INTENSIVE CARE UNITS BY
THERMAL GRADIENTS

JUAN DANIEL ESPINOZA CARO

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE
TELECOMUNICACIONES
MAESTRÍA EN INGENIERÍA DE TELECOMUNICACIONES
BUCARAMANGA

2026

PREDICTING SHOCK INDEX IN PEDIATRIC INTENSIVE CARE UNITS BY
THERMAL GRADIENTS

JUAN DANIEL ESPINOZA CARO

Thesis submitted in partial fulfillment of the requirements for the degree of Master in
Telecommunications Engineering.

Advisor

Carlos A. Fajardo, Ph.D.

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE
TELECOMUNICACIONES
MAESTRÍA EN INGENIERÍA DE TELECOMUNICACIONES
BUCARAMANGA

2026

CONTENT

	p.
1. INTRODUCTION	10
1.1. CLINICAL CONTEXT AND RELEVANCE OF SHOCK	10
1.2. ASSESSMENT AND MONITORING OF SHOCK	10
1.3. PEDIATRIC-SPECIFIC CONSIDERATIONS IN SHOCK EVALUATION	12
1.4. NON-INVASIVE METHODS FOR PERFUSION ESTIMATION	13
1.5. THERMOGRAPHY-BASED MACHINE LEARNING FOR SHOCK PRE- DICTION AND RELATED WORKS	16
2. OBJECTIVES	18
2.1. GENERAL OBJECTIVE	18
2.2. SPECIFIC OBJECTIVES	18
3. METHODOLOGY	19
3.1. DATABASE CHARACTERIZATION AND PREPROCESSING	19
3.2. COMPARATIVE ANALYSIS OF THERMAL GRADIENT BY SHOCK STA- TUS	22
3.3. MACHINE LEARNING MODEL EVALUATION FRAMEWORK	24
3.4. SHAP-BASED MODEL INTERPRETABILITY	28
4. RESULTS	30
4.1. COMPARATIVE ANALYSIS OF CLINICAL FACTORS BETWEEN GROUPS	30
4.2. MODEL PERFORMANCE AND OPTIMIZATION	31
4.2.1. Baseline Performance Across All Models and Horizons	31
4.2.2. Ablation Study: Impact of Thermal Gradient Features	33

4.3. HYPERPARAMETER OPTIMIZATION AND FEATURE IMPORTANCE	34
4.3.1. Hyperparameter Optimization	35
4.3.2. SHAP Analysis	36
5. DISCUSSION	38
5.1. MAIN FINDINGS AND INTERPRETATION	38
5.2. COMPARISON WITH OTHER STUDIES	39
5.3. METHODOLOGICAL CONSIDERATIONS	40
5.4. LIMITATIONS	41
6. CONCLUSIONS	42
BIBLIOGRAPHY.	44
ANNEXES	50

LIST OF FIGURES

	p.
Figure 1. Timeline representation of the temporally structured dataset	20

LIST OF TABLES

	p.
Table 1. Post-preprocessing dataset summary by temporal horizon	21
Table 2. Thermal gradient differences between shock groups	30
Table 3. Baseline AUC performance across prediction horizons	31
Table 4. Baseline Youden index by prediction horizon	32
Table 5. Thermal gradient ablation study	33
Table 6. Optimized CatBoost performance by horizon	35
Table 7. SHAP feature importance in optimized CatBoost models	36
Table 8. Optimized CatBoost hyperparameters by horizon	50

ANNEXES

p.

Annex A. OPTIMIZED CATBOOST HYPERPARAMETERS

50

RESUMEN

TÍTULO PREDICCIÓN DEL ÍNDICE DE CHOQUE EN UNIDADES DE CUIDADOS INTENSIVOS PEDIÁTRICOS MEDIANTE GRADIENTES TÉRMICOS *

AUTOR: JUAN DANIEL ESPINOZA CARO **

PALABRAS CLAVE: Aprendizaje automático, compromiso circulatorio, cuidados intensivos pediátricos, evaluación no invasiva, índice de choque, monitoreo hemodinámico.

DESCRIPCIÓN: El shock causa morbilidad y mortalidad en cuidados intensivos pediátricos, donde su reconocimiento tardío puede causar deterioro hemodinámico y falla orgánica. Las alteraciones de la perfusión periférica se reflejan en la temperatura cutánea, por lo que la imagen térmica puede facilitar su detección.

Este estudio evalúa los gradientes térmicos periféricos, específicamente la diferencia abdomen–pie, como marcadores no invasivos para predecir el shock. Los gradientes y variables clínicas se utilizaron en modelos de aprendizaje automático.

La predicción se definió como un desenlace binario derivado del índice de shock (SI), calculado como la frecuencia cardíaca dividida por la presión arterial sistólica. Al inicio ($t = 0$), se registraron las intensidades térmicas del abdomen y pie, la frecuencia del pulso y la edad. La presión arterial sistólica y la frecuencia cardíaca se obtuvieron a los 30 minutos, 3, 6 y 12 horas para calcular el SI y generar objetivos binarios de predicción.

El desempeño de los modelos se comparó entre ventanas y la importancia de las variables se analizó mediante valores SHAP. Se realizó un estudio de ablación para evaluar el impacto de características derivadas de la temperatura.

En conjunto, los hallazgos respaldan combinar gradientes térmicos con signos vitales convencionales para identificar pacientes pediátricos en riesgo de shock y resaltan el potencial del monitoreo térmico no invasivo con aprendizaje automático.

* Trabajo de grado

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería Eléctrica, Electrónica y de Telecomunicaciones. Director: Carlos A. Fajardo, Ph.D.

ABSTRACT

TITLE: PREDICTING SHOCK INDEX IN PEDIATRIC INTENSIVE CARE UNITS BY THERMAL GRADIENTS *

AUTOR: JUAN DANIEL ESPINOZA CARO **

Keywords: Circulatory compromise, hemodynamic monitoring, machine learning, non-invasive assessment, pediatric critical care, Shock Index.

Description: Shock is a cause of morbidity and mortality in pediatric intensive care units, where delayed recognition can lead to hemodynamic deterioration and organ failure. Because alterations in peripheral perfusion are reflected in skin temperature distribution, thermal imaging may provide a noninvasive approach for early shock detection.

This study evaluates peripheral thermal gradients, specifically the temperature difference between the abdomen and foot, as noninvasive markers for shock prediction. Abdomen-foot thermal gradients and clinical variables were used as inputs for machine learning models trained to estimate shock status.

The prediction target was a binary outcome derived from the shock index (SI), calculated as heart rate divided by systolic blood pressure. At baseline ($t = 0$), thermal intensities from the abdomen and foot were recorded with pulse rate and patient age. Systolic blood pressure and heart rate were obtained at 30 minutes, 3 hours, 6 hours, and 12 hours to calculate SI and generate binary prediction targets.

Model performance was compared across prediction windows, and feature importance was analyzed using SHAP values. An ablation study assessed the impact of temperature-derived features.

Overall, the findings support combining thermal gradients with conventional vital signs for identification of pediatric patients at risk of shock and highlight the potential of integrating noninvasive thermal monitoring with machine learning.

* Master's Thesis

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería Eléctrica, Electrónica y de Telecomunicaciones. Advisor: Carlos A. Fajardo, Ph.D.

1. INTRODUCTION

1.1. CLINICAL CONTEXT AND RELEVANCE OF SHOCK

Shock is a critical condition characterized by inadequate tissue perfusion and oxygen delivery, leading to cellular dysfunction and potential organ failure. In intensive care units (ICUs), the incidence of shock is estimated to be around 22 %, with mortality rates ranging from 37 % to 45 % depending on the underlying etiology ^{1 2 3}. Early recognition and management of hemodynamic instability are therefore essential to improve clinical outcomes, particularly in pediatric populations where physiological compensation may mask early signs of deterioration.

1.2. ASSESSMENT AND MONITORING OF SHOCK

Traditional methods for assessing shock often rely on invasive procedures or intermittent measurements, which limit their usefulness for continuous hemodynamic monitoring ⁴. To address these limitations, alternative physiological indices have been

¹ David D Berg et al. «Epidemiology of Shock in Contemporary Cardiac Intensive Care Units». En: *Circ Cardiovasc Qual Outcomes* 12.3 (2019). PMC11032172, e005618. DOI: 10 . 1161 / CIRCOUTCOMES . 119 . 005618.

² Tavazzi Guido et al. «Cardiogenic shock in general intensive care unit: a nationwide prospective analysis of epidemiology and outcome». En: *Eur Heart J Acute Cardiovasc Care* 13.11 (2024), págs. 768-778. DOI: 10.1093/ehjacc/zuae108.

³ Jean-Louis Vincent et al. «Frequency and mortality of septic shock in Europe and North America: a systematic review and meta-analysis». En: *Critical Care* 23.1 (2019), pág. 196. DOI: 10 . 1186 / s13054-019-2478-6.

⁴ A. Ochagavía et al. «Hemodynamic monitoring in the critically patient. Recommendations of the Cardiological Intensive Care and CPR Working Group of the Spanish Society of Intensive Care and Coronary Units». En: *Medicina Intensiva (English Edition)* 38.3 (2014). doi: 10.1016/j.medine.2013.10.002, págs. 154-169.

proposed, such as the peripheral perfusion index and the rate-pressure product^{5 6 7}. Unlike these indices, which reflect specific aspects of circulatory function, the Shock Index (SI) has emerged as a simple yet powerful indicator of circulatory status by capturing the balance between cardiac output and peripheral resistance. Unlike individual vital signs, SI provides a more integrated view of cardiovascular performance, allowing the identification of patients at risk of developing hemodynamic compromise. The SI is defined as the ratio of heart rate to systolic blood pressure^{8 9 10}. In clinical practice, early identification of shock typically relies on bedside findings such as tachycardia, tachypnea, altered peripheral perfusion, delayed capillary refill, cool skin, oliguria, and laboratory indicators like elevated lactate or base deficit¹¹. These observable signs constitute the primary screening tools used by clinicians when evaluating undifferentiated patients for possible circulatory failure. However, traditio-

-
- ⁵ Ertuğ Dinçer et al. «Evaluation of peripheral perfusion index in predicting blood product need for trauma patients». En: *Ulus Travma Acil Cerrahi Derg* 31.6 (2025). PMID: 40511758, págs. 525-530. DOI: 10.14744/tjtes.2025.65957. eprint: <https://dx.doi.org/10.14744/tjtes.2025.65957>.
 - ⁶ Ahmed Hasanin et al. «The ability of pulse oximetry-derived peripheral perfusion index to detect fluid responsiveness in patients with septic shock». En: *Journal of Anesthesia* 35.2 (abr. de 2021), págs. 254-261. DOI: 10.1007/s00540-021-02908-w.
 - ⁷ Félix D. Martos-Benítez y Bárbara I. Guzmán-Breff. «Volume expansion and variation in hemodynamic parameters». En: *Emergencias* 30.3 (jun. de 2018). English, Spanish, págs. 177-181. PMID: 29687672.
 - ⁸ Erica Koch et al. «Shock index in the emergency department: utility and limitations». En: *Open Access Emergency Medicine* 11 (2019). doi: 10.2147/OAEM.S178358, págs. 179-199.
 - ⁹ Tony Berger et al. «Shock index and early recognition of sepsis in the emergency department: pilot study». En: *The Western Journal of Emergency Medicine* 14.2 (mar. de 2013). doi: 10.5811/westjem.2012.8.11546, págs. 168-174.
 - ¹⁰ Matthias Noitz et al. «The Predictive Value of Clinical Signs to Identify Shock in Critically Ill Patients». En: *Diagnostics* 15.17 (2025). DOI: 10.3390/diagnostics15172252.
 - ¹¹ Matthew C. Strehlow. «Early identification of shock in critically ill patients». En: *Emergency Medicine Clinics of North America* 28.1 (feb. de 2010), págs. 57-66. DOI: 10.1016/j.emc.2009.09.006.

nal diagnostic thresholds, particularly systolic blood pressure below 90 mmHg are now recognized as late manifestations rather than early markers of shock¹². This limitation underscores the need for simple screening tools capable of signaling hemodynamic instability before over hypotension develops, thereby supporting earlier recognition and intervention.

1.3. PEDIATRIC-SPECIFIC CONSIDERATIONS IN SHOCK EVALUATION

In pediatric critical care, the evaluation of shock presents unique challenges due to the dynamic nature of normal physiological ranges across different age groups. Children can maintain near-normal blood pressure for prolonged periods despite significant circulatory compromise, making hypotension a late and unreliable indicator of shock¹³. For this reason, age-adjusted SI cutoffs have been introduced to improve diagnostic accuracy and facilitate earlier detection of hemodynamic instability^{14 15}. Moreover, studies have shown that SI can identify subtle alterations in perfusion that

¹² Brock Bruns et al. «Prehospital hypotension redefined». En: *Journal of Trauma* 65.6 (dic. de 2008), págs. 1217-1221. DOI: 10.1097/TA.0b013e318184ee63.

¹³ Pedro Celiny Ramos Garcia, Cristian Tedesco Tonial y Jefferson Pedro Piva. «Septic shock in pediatrics: the state-of-the-art». En: *Jornal de Pediatria* 96 Suppl 1.Suppl 1 (2020). doi: 10.1016/j.jpmed.2019.10.007, págs. 87-98.

¹⁴ Shannon N. Acker et al. «Shock index, pediatric age-adjusted (SIPA) is more accurate than age-adjusted hypotension for trauma team activation». En: *Surgery* 161.3 (mar. de 2017). doi: 10.1016/j.surg.2016.08.050, págs. 803-807.

¹⁵ Ryan Phillips et al. «The shock index, pediatric age-adjusted (SIPA) enhanced: Prehospital and emergency department SIPA values forecast transfusion needs for blunt solid organ injured children». En: *Surgery* 168.4 (2020). doi: 10.1016/j.surg.2020.04.061, págs. 690-694.

precede overt clinical deterioration, enabling earlier therapeutic interventions ^{16 17}.

1.4. NON-INVASIVE METHODS FOR PERFUSION ESTIMATION

Tissue hypoperfusion in shock is commonly assessed using serum lactate and invasive hemodynamic measurements, such as continuous arterial blood pressure monitoring through an arterial catheter or central venous parameters obtained through central venous catheterization¹⁸. Although clinically valuable, these approaches are invasive because they require blood sampling or the insertion of vascular catheters, which may demand trained personnel, specialized equipment, and additional time, while also exposing patients to complications such as bleeding, hematoma, vascular injury, thrombosis, and catheter-related infection. In this context, growing evidence indicates that non-invasive bedside indicators of peripheral perfusion can provide clinically relevant information. As a widely used bedside indicator of peripheral perfusion in critical care, capillary refill time (CRT) has been evaluated not only for routine hemodynamic assessment but also as a guide for resuscitation strategies. CRT approaches have demonstrated similar outcomes to lactate-guided approaches in terms of microcirculation, tissue oxygenation, and organ function, while achieving therapeutic targets more rapidly when CRT normalization (≤ 3 s) is used to support fluid administration decisions ¹⁹. Beyond its role in guiding resuscitation, CRT

¹⁶ Mohamed Y. Rady et al. «Shock index: a re-evaluation in acute circulatory failure». En: *Resuscitation* 23.3 (1992). doi: 10.1016/0300-9572(92)90006-X, págs. 227-234.

¹⁷ Gustavo A. Ospina-Tascón et al. «Diastolic shock index and clinical outcomes in patients with septic shock». En: *Annals of Intensive Care* 10.1 (2020). doi: 10.1186/s13613-020-00658-8, pág. 41.

¹⁸ Maurizio Cecconi et al. «Consensus on circulatory shock and hemodynamic monitoring. Task force of the European Society of Intensive Care Medicine». En: *Intensive Care Medicine* 40.12 (2014), págs. 1795-1815. DOI: 10.1007/s00134-014-3525-z.

¹⁹ R. Castro et al. «Effects of capillary refill time-vs. lactate-targeted fluid resuscitation on regional, microcirculatory and hypoxia-related perfusion parameters in septic shock: a randomized contro-

has also shown strong predictive performance for elevated lactate levels, with high sensitivity and satisfactory discrimination, supporting its role as a feasible surrogate marker, particularly in resource-limited settings ¹⁰. Even in high-risk populations such as pediatric patients undergoing congenital cardiac surgery, CRT has been identified as an independent predictor of mortality and need for extracorporeal support, with performance approaching that of lactate measurements despite its simplicity and non-invasive nature ²⁰. Nevertheless, CRT is a semi-quantitative and observer-dependent measure, commonly assessed visually with a temporal resolution limited to seconds and influenced by factors such as applied pressure, measurement site, ambient conditions, and examiner interpretation. These limitations motivate the use of complementary non-invasive methods that provide more objective information on peripheral perfusion, such as infrared thermography, which enables contactless assessment of skin temperature distribution as an indirect marker of peripheral blood flow.

Infrared thermography (IRT) provides a useful non-invasive approach for hemodynamic assessment because it captures objective, contactless, and spatially resolved information on skin temperature distribution related to peripheral perfusion and microcirculatory status. As a physiological proof of concept, controlled studies have shown that pharmacologically induced hemodynamic changes produce consistent variations in central-to-peripheral thermal gradients, supporting thermal imaging as a proxy for perfusion status ²¹. This relationship is clinically relevant in shock, where

lled trial». En: *Ann. Intensive Care* 10.1 (2020), pág. 150. DOI: 10.1186/s13613-020-00767-4.

²⁰ Gustavo Cruz et al. «Capillary Refill Time and Serum Lactate as Predictors of Mortality and Post-operative Extracorporeal Membrane Oxygenation Requirement in Congenital Heart Surgery». En: *Children* 10.5 (2023). DOI: 10.3390/children10050875.

²¹ Mathieu Magnin et al. «Use of infrared thermography to detect early alterations of peripheral perfusion: evaluation in a porcine model». En: *Biomed Opt Express* 11.5 (8 de abr. de 2020). doi: 10.1364/BOE.387481, págs. 2431-2446.

decreased cardiac output and compensatory redistribution of blood flow may reduce heat transfer to the extremities, generating measurable temperature differences between central and peripheral regions ^{22 23}. Consistent with this mechanism, clinical studies have reported that central-to-peripheral thermal gradients correlate with circulatory alterations and perfusion-related outcomes in critically ill patients ^{24 25 22}. In pediatric critical care, thermal imaging has also been explored as a non-invasive strategy for monitoring hemodynamic status and identifying early signs of cardiovascular compromise ^{26 22}. Building upon this physiological and clinical evidence, IRT-derived temperature features have increasingly been investigated as non-invasive thermal biomarkers for assessing hemodynamic instability and supporting early shock prediction.

-
- ²² Armelle Bridier et al. «Hemodynamic assessment in children after cardiac surgery: A pilot study on the value of infrared thermography». En: *Frontiers in pediatrics* 11 (abr. de 2023). doi: 10.3389/fped.2023.1083962, pág. 1083962.
 - ²³ Vrinda Lath et al. «Utility of core to peripheral temperature gradient using infrared thermography in the assessment of patients with sepsis and septic shock in the emergency medicine department». En: *International Journal of Emergency Medicine* 18 (2025). DOI: 10.1186/s12245-025-00890-8.
 - ²⁴ Simon Bourcier et al. «Toe-to-room temperature gradient correlates with tissue perfusion and predicts outcome in selected critically ill patients with severe infections». En: *Annals of Intensive Care* 6.1 (2016). doi: 10.1186/s13613-016-0164-2, pág. 63.
 - ²⁵ Harry Amson et al. «Core-to-skin temperature gradient measured by thermography predicts day-8 mortality in septic shock: A prospective observational study». En: *Journal of Critical Care* 60 (2020). doi: 10.1016/j.jcrc.2020.08.022, págs. 294-299.
 - ²⁶ Vanshika Vats et al. «Early Prediction of Hemodynamic Shock in Pediatric Intensive Care Units With Deep Learning on Thermal Videos». En: *Frontiers in Physiology* 13 (2022). doi: 10.3389/fphys.2022.862411.

1.5. THERMOGRAPHY-BASED MACHINE LEARNING FOR SHOCK PREDICTION AND RELATED WORKS

Methodological studies have further shown that even single time-point thermal measurements, when combined with routinely collected physiological parameters, can yield clinically actionable predictive insights. Notably, previous work using a linear mixed-effects model achieved competitive discrimination ($AUC \approx 0.79$), indicating that meaningful predictive information can be extracted from thermographic data acquired at a single time point. However, the study did not report confidence intervals or detailed uncertainty estimates, limiting a rigorous assessment of robustness and generalizability ²⁷. This single time-point paradigm has also been supported in critically ill septic populations, where thermal images obtained after vasopressor initiation showed that predefined temperature thresholds and core-to-skin gradients were significantly associated with short-term mortality risk ²⁵.

More recently, research has shifted from static measurements toward temporal modeling of thermal dynamics. This transition is exemplified by studies employing long short-term memory (LSTM) networks to analyze thermal video sequences and synchronized physiological signals, successfully predicting a binarized shock index with an AUC of 0.81 (standard error = 0.06), illustrating the capacity of deep learning methods to extract clinically meaningful temporal patterns from continuous thermal data ²⁶. Although advanced analytical methods show great promise, several challenges remain before they can be used in clinical practice. Despite the increasing use of thermography in pediatric intensive care, the field still lacks standardized protocols for data collection and processing. Furthermore, there is limited evidence that these models work across different patient groups, and we still do not fully unders-

²⁷ Aditya Nagori et al. «Predicting Hemodynamic Shock from Thermal Images using Machine Learning». En: *Scientific Reports* 9 (ene. de 2019). doi: 10.1038/s41598-018-36586-8.

tand how thermal changes relate to clinical outcomes. While simple linear models may provide basic insights, they often fail to capture complex non-linear patterns. On the other hand, more complex models may capture richer non-linear patterns, but their reported performance can also be accompanied by greater uncertainty or reduced stability, which may limit confidence in their reliability for hospital use. These limitations highlight the need for more robust machine learning approaches that not only improve predictive accuracy, but also support consistent and clinically reliable performance, helping move thermal monitoring from the lab to real-world clinical practice.

2. OBJECTIVES

2.1. GENERAL OBJECTIVE

To develop a predictive machine learning model for shock index based on core-to-peripheral thermal gradients in pediatric intensive care units.

2.2. SPECIFIC OBJECTIVES

To characterize and curate the database of thermal gradients and phenotype information from children in Pediatric Intensive Care Units.

To evaluate and compare the performance of different machine learning model architectures for the prediction of shock based on thermal gradients.

To assess the impact of core-to-peripheral temperature gradients in the machine learning model for shock prediction.

3. METHODOLOGY

3.1. DATABASE CHARACTERIZATION AND PREPROCESSING

The dataset used in this study was collected by Nagori et al.²⁷ and is publicly available on the Open Science Framework²⁸ under the CC0 1.0 Universal license. The dataset contains thermographic measurements and physiological variables from pediatric patients, with repeated observations available for some individuals.

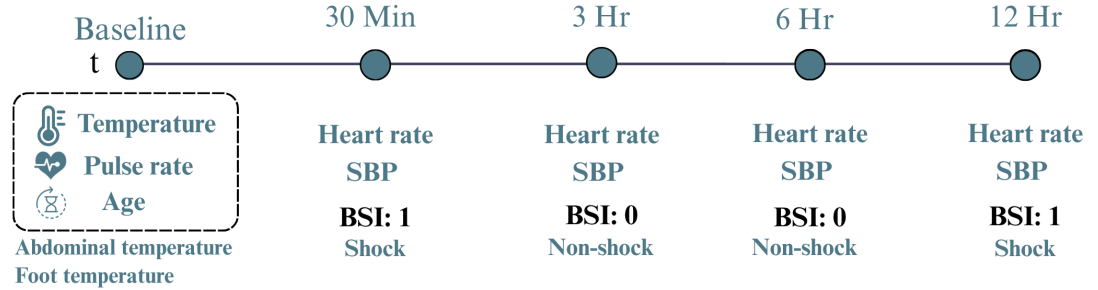
Data were organized according to a predefined temporal structure. At baseline ($t = 0$), thermographic measurements were obtained from the abdominal and foot regions, and the corresponding skin temperature values were extracted. At the same baseline time point, pulse rate and patient age in months were recorded. Hemodynamic variables, including systolic blood pressure (SBP) and heart rate (HR), were available at four subsequent prediction horizons: 30 minutes, 3 hours, 6 hours, and 12 hours after thermographic acquisition.

To clarify the temporal structure of the dataset, Figure 1 shows an illustrative fragment of how each thermographic acquisition at baseline was linked to future hemodynamic measurements. For each patient and baseline acquisition, thermal and clinical variables recorded at $t = 0$ were used as predictors, whereas SBP and HR measured at later time horizons were used to compute the future shock index and define the corresponding binary prediction target.

The baseline predictor variables considered in the initial dataset included abdominal temperature, foot temperature, pulse rate, age in months. SBP and HR were not used as baseline predictors; instead, they were used to derive the shock status outcome

²⁸ Tavpritesh Sethi. *Data for manuscript on automated shock detection in the ICU using Thermal Imaging*. doi: 10.17605/OSF.IO/VP86J. 2018.

Figure 1. Timeline representation of the temporally structured dataset used for shock prediction. Baseline refers to $t = 0$ measurements that were used as predictors, while future heart rate and systolic blood pressure (SBP) were obtained at predefined prediction horizons to derive the corresponding binary shock index (BSI).



at each prediction horizon.

The target variable, corresponding to the presence of hemodynamic shock, was derived using the Shock Index Pediatric Age-adjusted (SIPA) ¹⁴. For each observation and prediction horizon, the shock index was calculated as:

$$SI = \frac{HR}{SBP}$$

The computed shock index was compared with the corresponding age-specific SIPA threshold. Observations exceeding this threshold were labeled as shock (1), while the remaining observations were labeled as no shock (0). Because the outcome was derived separately at each prediction horizon, four horizon-specific analytical datasets were generated.

Patient ages ranged from 0 to 144 months, with a mean age of 43.41 months and a standard deviation of 49.33 months. The dataset included 35 male and 31 female patients.

Preprocessing was performed independently for each prediction horizon. Observations with missing values in the baseline predictor variables were excluded. Records

without abdominal temperature, foot temperature, pulse rate, or age in months were removed. In addition, observations lacking SBP or HR at the corresponding prediction horizon were excluded because these variables were required to compute the shock index and assign the binary outcome.

Table 1 summarizes the number of samples and unique patients after preprocessing for each prediction horizon, as well as the corresponding distribution of shock and no-shock observations.

Table 1. Number of samples, unique patients, and outcome distribution after preprocessing per temporal horizon.

Time Horizon	Samples	Patients	Shock	No Shock
30 min	243	51	100 (41.15 %)	143 (58.85 %)
3 h	238	51	100 (42.02 %)	138 (57.98 %)
6 h	228	49	104 (45.61 %)	124 (54.39 %)
12 h	239	50	112 (46.86 %)	127 (53.14 %)

In addition to the original thermographic measurements, a derived variable representing the abdomen–foot thermal gradient was computed following the approach described by Nagori et al.²⁷. The gradient was calculated as:

$$\Delta T_{\%} = \frac{T_{abdomen} - T_{foot}}{T_{abdomen}} \times 100$$

where $T_{abdomen}$ and T_{foot} represent the skin temperature values extracted from the abdominal and foot regions, respectively. This variable was included as a central-to-peripheral temperature contrast intended to characterize differences in skin temperature distribution between anatomical regions. The resulting preprocessed, horizon-specific datasets were then used for the exploratory mixed-effects analysis and subsequent predictive modeling stages.

3.2. COMPARATIVE ANALYSIS OF THERMAL GRADIENT BY SHOCK STATUS

Before developing the predictive models, an exploratory comparative analysis was conducted to evaluate whether thermal gradient values differed between patients who developed hemodynamic shock and those who did not. This analysis was performed independently for each predefined time interval: 30 minutes, 3 hours, 6 hours, and 12 hours.

For each time interval, a separate linear mixed-effects model was fitted using thermal gradient as the dependent variable and shock status as the fixed effect of interest. Patient identification was included as a random intercept to account for the correlation among repeated measurements obtained from the same patient. This approach allowed the model to estimate the expected thermal gradient for each group while accounting for patient-specific baseline variability.

Before fitting the mixed-effects model, shock status was encoded as a binary indicator. Let $S_{ij}^{(t)}$ denote the shock status of observation j from patient i at prediction horizon t , where $S_{ij}^{(t)} = 1$ indicates shock and $S_{ij}^{(t)} = 0$ indicates no shock. The following model was then fitted separately for each prediction horizon:

$$Y_{ij}^{(t)} = \beta_0^{(t)} + \beta_1^{(t)} S_{ij}^{(t)} + b_i^{(t)} + \epsilon_{ij}^{(t)} \quad (1)$$

where $Y_{ij}^{(t)}$ represents the abdomen–foot thermal gradient for observation j from patient i at prediction horizon t . The variable $S_{ij}^{(t)}$ is a binary indicator of future shock status at that horizon, with $S_{ij}^{(t)} = 1$ indicating shock and $S_{ij}^{(t)} = 0$ indicating no shock. The fixed intercept $\beta_0^{(t)}$ represents the estimated mean thermal gradient for the reference group, corresponding to observations classified as no shock. The coefficient $\beta_1^{(t)}$ represents the average difference in thermal gradient between shock and no-shock observations within the same prediction horizon. Therefore, the expected mean thermal gradient for the no-shock group is $\beta_0^{(t)}$, whereas the expected mean

thermal gradient for the shock group is $\beta_0^{(t)} + \beta_1^{(t)}$. The term $b_i^{(t)}$ represents the patient-specific random intercept, which accounts for repeated observations within the same patient, and $\epsilon_{ij}^{(t)}$ denotes the residual error term.

The term $b_i^{(t)}$ represents the patient-specific random intercept, which allows each patient to have an individual baseline deviation from the overall group mean. This term accounts for the fact that repeated measurements from the same patient are likely to be more similar to each other than measurements from different patients. Finally, $\epsilon_{ij}^{(t)}$ represents the residual error, corresponding to the unexplained variability in thermal gradient after considering shock status and patient-specific effects.

Estimated marginal means were obtained for the shock and no-shock groups within each time interval. The between-group contrast was calculated as the estimated mean thermal gradient in the shock group minus the estimated mean thermal gradient in the no-shock group. Thus, the primary objective of this analysis was to compare the expected thermal gradient between groups, rather than to predict the occurrence of shock.

To obtain robust measures of uncertainty, 95 % confidence intervals for the estimated marginal means and between-group differences were calculated using bootstrap resampling. Resampling was performed at the patient level to preserve the correlation structure of repeated measurements within individuals. Two-sided bootstrap p-values were derived from the empirical distribution of the between-group contrasts. A p-value below 0.05 was considered statistically significant.

Because a separate model was fitted for each time interval, time was not included as a fixed effect in the models. Therefore, this analysis was interpreted as an interval-specific comparison between shock and no-shock groups rather than as a formal longitudinal model of temporal change.

3.3. MACHINE LEARNING MODEL EVALUATION FRAMEWORK

The interval-specific mixed-effects analysis was used as an exploratory step to determine whether thermal gradient, the main physiological variable derived from thermographic measurements, showed measurable differences between patients who developed hemodynamic shock and those who did not. By accounting for repeated measurements within patients, this analysis provided a patient-aware assessment of the univariate association between thermal gradient and shock status at each prediction horizon. This preliminary analysis also provided a broader physiological and statistical context for interpreting the subsequent machine learning results.

After this preliminary characterization, a predictive modeling stage was conducted to evaluate whether the available thermal and clinical variables could be used jointly to discriminate between shock and no-shock cases. Unlike the mixed-effects analysis, which focused on estimating group-level differences in thermal gradient, the machine learning stage aimed to assess classification performance across the same predefined prediction horizons: 30 min, 3 h, 6 h, and 12 h.

A systematic evaluation of ten machine learning models was conducted. The evaluated models included six tree-based algorithms and four non-tree-based models were also evaluated: Logistic Regression (LR), Support Vector Machine (SVM), k -Nearest Neighbors (k -NN), and Multilayer Perceptron (MLP). This comparison allowed the assessment of how different model families performed across prediction horizons, providing insight into their discriminative capacity in the context of pediatric critical care.

All continuous features were standardized using z -score normalization. The scaler was fitted exclusively on the training partition of each cross-validation fold and subsequently applied to the corresponding held-out fold to prevent data leakage.

Given the clustered structure of the dataset, with multiple observations per patient, a standard random split could lead to information leakage and overly optimistic perfor-

mance estimates. Therefore, a 10-fold stratified grouped cross-validation strategy was implemented using `StratifiedGroupKFold`. Patient identifiers were used as grouping variables to ensure that all observations from the same patient were assigned exclusively to either the training or held-out fold. Stratification was used to preserve the proportion of shock and non-shock cases across folds. This strategy allowed model performance to be estimated while preserving patient-level independence.

Because the dataset included a limited number of patients and an unequal number of observations per patient, no additional validation split was created. Creating such a split would have reduced the amount of patient-level data available for training and could have increased the variability of the performance estimates. Therefore, model performance was estimated within the 10-fold patient-grouped cross-validation framework.

For the initial benchmarking stage, all algorithms were evaluated under the same preprocessing, grouping, and evaluation conditions. This design was intended to provide a homogeneous comparison across model families, allowing their relative performance to be assessed under a common experimental framework.

Model performance was evaluated using two complementary criteria: the area under the receiver operating characteristic curve (AUC) and the maximum Youden index²⁹. AUC was used to assess global discriminative ability independently of a fixed classification threshold, whereas the Youden index was used to evaluate threshold-dependent classification performance.

For each model and prediction horizon, predicted probabilities from the held-out folds were aggregated across the 10-fold patient-grouped cross-validation procedure. The

²⁹ Ronen Fluss, David Faraggi y Benjamin Reiser. «Estimation of the Youden Index and its associated cutoff point». En: *Biometrical Journal* 47.4 (ago. de 2005). doi: 10.1002/bimj.200410135, págs. 458-472. DOI: 10.1002/bimj.200410135.

AUC was then computed using these out-of-fold predictions, and 95 % confidence intervals were estimated using bootstrap resampling. In parallel, the optimal classification threshold was identified as the probability threshold that maximized the Youden index:

$$J = \text{sensitivity} + \text{specificity} - 1$$

The maximum Youden index obtained at this optimal threshold was reported for each model and prediction horizon. Therefore, model comparison considered both global discrimination, represented by AUC, and the best threshold-dependent operating performance, represented by the maximum Youden index.

After the baseline benchmarking stage, a single model family was selected for subsequent analyses based on its overall performance across the four prediction horizons. Model selection was primarily guided by cross-validated AUC, while the maximum Youden index was considered as complementary evidence of threshold-dependent performance. Rather than selecting a different model for each prediction horizon, priority was given to the model family that showed the most consistent overall performance across horizons.

This strategy was adopted to maintain a fair comparison in the subsequent analysis of the impact of thermal gradient features. By using the same model family across all feature-set comparisons, differences in performance could be more directly attributed to the inclusion, exclusion, or transformation of thermal gradient information rather than to changes in the underlying learning algorithm.

Before Bayesian hyperparameter optimization, an ablation study was conducted using the selected model family. The purpose of this analysis was to quantify the contribution of thermal gradient features and to determine which representation of the thermal gradient provided the greatest improvement in classification performance. During this ablation stage, the model family and hyperparameter configuration were

kept fixed, and only the thermal-gradient feature configuration was varied.

The ablation study compared five feature configurations: a baseline configuration without thermal gradient features, a configuration including the linear thermal gradient term ($\Delta T \%$), a configuration including the squared thermal gradient term $(\Delta T \%)^2$, a configuration including the cubic thermal gradient term $(\Delta T \%)^3$, and a full-gradient configuration including all thermal-gradient components. The squared term was included to evaluate whether the magnitude of the thermal difference, regardless of direction, improved classification performance. The cubic term was included to assess whether preserving the sign of the gradient while amplifying larger differences provided additional discriminative information.

For each prediction horizon, ablation performance was evaluated using the maximum Youden index. This metric was selected because the ablation analysis focused on threshold-dependent classification performance and on identifying whether thermal-gradient features improved the balance between sensitivity and specificity. The same 10-fold patient-grouped cross-validation framework was used, and comparisons were made while keeping all non-gradient variables unchanged.

The findings from the ablation study were then used to define the thermal-gradient feature configuration carried forward to the hyperparameter optimization stage. After this feature configuration was established, Bayesian hyperparameter optimization was performed using the Optuna framework. The optimization objective was defined as the mean cross-validated AUC within the 10-fold patient-grouped cross-validation framework, while the maximum Youden index was retained as a complementary threshold-dependent performance metric.

The hyperparameter search spaces, default configurations, ablation feature configurations, and final optimized hyperparameters are reported in Annex 1. Baseline benchmarking results, ablation results, and optimized model results were analyzed separately to distinguish the initial model comparison stage, the thermal-gradient

feature assessment, and the final optimization stage.

3.4. SHAP-BASED MODEL INTERPRETABILITY

After Bayesian hyperparameter optimization, an interpretability analysis was performed to examine how the optimized models used the available clinical and thermal information to generate shock predictions. For this purpose, SHAP (SHapley Additive exPlanations) values were computed for the optimized CatBoost models at each prediction horizon.

SHAP is a model interpretability method based on Shapley values from cooperative game theory. It assigns each input variable a contribution value that indicates how much that variable shifts the model output from the average prediction. A positive SHAP value indicates that the feature increases the predicted risk of shock relative to the average prediction, whereas a negative SHAP value indicates that the feature decreases the predicted risk. Values close to zero indicate little influence on a given prediction.

Thus, SHAP allowed us to evaluate the direction and magnitude of the contribution of each clinical and thermal feature to individual predictions. For global interpretation, feature importance was summarized using the mean absolute SHAP value, which represents the average magnitude of a feature's contribution across observations, regardless of whether the effect was positive or negative.

This analysis was conducted after model optimization in order to interpret the final predictive models under the same configuration used for performance evaluation. While the preceding stages focused on model selection, thermal-gradient feature assessment, and hyperparameter tuning, as the SHAP analysis aimed to identify the relative contribution of each predictor to the optimized model outputs.

For each prediction horizon, SHAP values were computed within the 10-fold patient-grouped cross-validation framework. Specifically, for each fold, SHAP values were

estimated on the corresponding held-out partition using the model trained on the remaining folds. The resulting SHAP values were then aggregated across folds to obtain global feature importance summaries for each horizon. This out-of-fold strategy allowed feature contributions to be assessed using predictions generated on data not used to train the corresponding fold-specific model.

Global feature importance was summarized using the mean absolute SHAP value for each predictor. This metric represents the average magnitude of a feature's contribution to the model output, regardless of whether the contribution increases or decreases the predicted risk. The analysis focused on the main clinical and thermal predictors, including pulse rate, age, and the selected thermal gradient representation, to determine whether thermal information contributed complementary predictive value beyond conventional baseline variables.

4. RESULTS

4.1. COMPARATIVE ANALYSIS OF CLINICAL FACTORS BETWEEN GROUPS

The first results examined whether the thermal gradient differed between patients classified as shock and no shock across the predefined prediction horizons. Table 2 summarizes the model-estimated thermal gradient for both groups, along with the estimated between-group difference and corresponding p-value for each horizon.

Table 2. Estimated thermal gradient for the no-shock group and between-group differences by shock status across prediction horizons.

Interval	No shock	Shock	Difference	p-value
30min	10.3 [7.13–13.47]	15.21 [11.67–18.74]	4.91 [1.09–8.74]	0.012
3hr	10.93 [7.95–13.9]	14.78 [11.39–18.16]	3.85 [0.10–7.59]	0.044
6hr	10.96 [7.67–14.25]	14.33 [10.82–17.84]	3.36 [-0.57–7.30]	0.094
12 hr	11.23 [7.94–14.51]	14.03 [10.56–17.51]	2.81 [-1.04–6.65]	0.153

Values are presented as model estimates with 95 % confidence intervals. The no-shock group was used as the reference category. The difference was calculated as shock minus no shock.

The estimated between-group contrasts were positive across all prediction horizons, indicating that patients classified as shock tended to present higher abdomen–foot thermal gradients than patients classified as no shock. The separation between groups was strongest at 30 minutes and 3 hours, where the differences reached statistical significance.

At 6 and 12 hours, the same directional pattern was observed, but the magnitude of the differences decreased and the confidence intervals included zero. Taken together, these results suggest a time-dependent pattern in the thermal gradient signal: the abdomen–foot thermal contrast was more clearly associated with shock status in the short-term windows, whereas its ability to distinguish between groups diminished

as the prediction horizon increased.

4.2. MODEL PERFORMANCE AND OPTIMIZATION

4.2.1. Baseline Performance Across All Models and Horizons Tables 3 and 4 summarize the baseline performance of the ten machine learning models across the four prediction horizons. Table 3 reports AUC values with 95 % confidence intervals, whereas Table 4 reports the maximum Youden index obtained for each model and horizon.

Table 3. AUC with 95 % confidence intervals for all models across prediction horizons. Bold values indicate the highest AUC achieved within each prediction horizon.

Model	30-min	3-hour	6-hour	12-hour
LR	0.7557 [0.6503–0.8466]	0.7631 [0.6306–0.8660]	0.6320 [0.5048–0.7563]	0.6861 [0.5854–0.7808]
MLP	0.7348 [0.6348–0.8214]	0.7297 [0.5999–0.8416]	0.6998 [0.5907–0.7947]	0.6745 [0.5844–0.7660]
SVM	0.7135 [0.5948–0.8199]	0.7346 [0.5966–0.8435]	0.6691 [0.5545–0.7730]	0.6309 [0.5305–0.7298]
k -NN	0.6599 [0.5641–0.7507]	0.6966 [0.5734–0.8038]	0.6351 [0.5401–0.7251]	0.6016 [0.5030–0.6999]
GB	0.7912 [0.7136–0.8639]	0.7561 [0.6632–0.8405]	0.6409 [0.5452–0.7313]	0.6425 [0.5407–0.7434]
XGB	0.7687 [0.6889–0.8437]	0.7476 [0.6601–0.8309]	0.6278 [0.5310–0.7223]	0.5953 [0.4993–0.6858]
CB	0.8148 [0.7401–0.8871]	0.7848 [0.6987–0.8653]	0.6862 [0.5911–0.7738]	0.6524 [0.5611–0.7429]
DT	0.6679 [0.5865–0.7505]	0.6728 [0.5895–0.7555]	0.5646 [0.4787–0.6533]	0.5821 [0.5084–0.6544]
RF	0.7994 [0.7238–0.8689]	0.7605 [0.6678–0.8473]	0.6374 [0.5275–0.7404]	0.6245 [0.5234–0.7202]
LGBM	0.7915 [0.7151–0.8652]	0.7514 [0.6619–0.8360]	0.6720 [0.5800–0.7563]	0.6549 [0.5677–0.7351]

The highest observed performance was concentrated in the shorter prediction horizons. At 30 minutes, CatBoost achieved the highest AUC and the highest maximum

Table 4. Youden index for all models across prediction horizons. Bold values indicate the highest Youden index achieved within each prediction horizon.

Model	30-min	3-hour	6-hour	12-hour
LR	0.4185	0.4258	0.2420	0.3394
MLP	0.3763	0.3955	0.3076	0.3068
SVM	0.3605	0.4189	0.3924	0.2667
k -NN	0.2449	0.2745	0.2721	0.1596
GB	0.4549	0.3982	0.2461	0.2924
XGB	0.4634	0.4163	0.2427	0.2121
CB	0.5230	0.4556	0.3017	0.2778
DT	0.3282	0.3436	0.1291	0.1635
RF	0.4774	0.4066	0.2244	0.2181
LGBM	0.5139	0.4443	0.3085	0.2294

Youden index. A similar pattern was observed at 3 hours, where CatBoost again obtained the highest value for both metrics. Other ensemble tree-based models, particularly Random Forest, LightGBM, and Gradient Boosting, also showed strong performance at these short-term horizons, although their threshold-dependent performance was slightly lower than that of CatBoost.

At the 6-hour horizon, the ranking of models differed according to the metric considered. MLP achieved the highest AUC, whereas SVM obtained the highest maximum Youden index. CatBoost and LightGBM remained among the higher-performing models in terms of AUC, but they did not achieve the highest Youden values at this horizon. This result suggests that, at intermediate horizons, the model with the strongest global discrimination was not necessarily the same model with the best threshold-dependent classification performance.

At 12 hours, Logistic Regression achieved the highest observed AUC and maximum Youden index. However, the overall values at this horizon were lower than those observed at 30 minutes and 3 hours. This pattern indicates that long-horizon prediction was more challenging for all model families, with a weaker separation between shock and no-shock cases compared with earlier prediction windows.

Across horizons, CatBoost showed the most consistent overall baseline performan-

ce. It achieved the highest AUC and Youden index at both 30 minutes and 3 hours and maintained competitive AUC values at 6 and 12 hours. Although other models achieved the highest values at specific later horizons, CatBoost provided the strongest cross-horizon profile in the baseline comparison. For this reason, CatBoost was selected as the model family for the subsequent thermal-gradient impact analyses, allowing later comparisons to be performed using a fixed learning architecture.

4.2.2. Ablation Study: Impact of Thermal Gradient Features Following the baseline model comparison, the ablation study was performed using CatBoost to assess how different thermal-gradient feature representations affected classification performance across prediction horizons. Table 5 summarizes the maximum Youden index obtained for each feature configuration.

Table 5. Ablation study: impact of thermal gradient components on classification performance measured by the Youden index. Results are shown for models including: no gradient features (Baseline), only the linear term ($\Delta T_{\%}$), only the squared term ($(\Delta T_{\%})^2$), only the cubic term ($(\Delta T_{\%})^3$), and the full model including all gradient components.

Model (Horizon)	<i>Baseline</i>	$\Delta T_{\%}$	$(\Delta T_{\%})^2$	$(\Delta T_{\%})^3$	<i>all gradients</i>
CB (30 min)	0.5068	0.5229	0.5402	0.5214	0.5256
CB (3 h)	0.4664	0.4614	0.4988	0.4710	0.4527
CB (6 h)	0.3280	0.3047	0.2711	0.3039	0.2837
CB (12 h)	0.3503	0.2869	0.2971	0.2807	0.2662

The ablation analysis revealed a horizon-dependent contribution of thermal-gradient features. Rather than producing a uniform improvement across all prediction windows, the added thermal information appeared to be more beneficial in the short-term horizons. This suggests that the abdomen–foot thermal gradient may provide

useful, although moderate, discriminative information closer to the shock event, where peripheral temperature alterations may be more directly reflected in the model output.

Among the evaluated gradient representations, the squared term showed the clearest contribution in the short-term windows. This pattern indicates that the magnitude of the thermal imbalance was more informative than the raw directional gradient. In other words, emphasizing larger central-to-peripheral temperature differences improved classification more consistently than preserving the sign of the gradient through the cubic transformation.

At longer horizons, the addition of thermal-gradient features did not improve the threshold-dependent performance of the CatBoost model. This suggests that the thermal signal became less informative as the prediction window moved farther from the event, and that the available non-gradient variables provided a stronger basis for classification at 6 and 12 hours. Overall, the ablation results supported the use of the squared thermal-gradient representation for the subsequent optimization stage, while also showing that its contribution was primarily relevant for short-term prediction.

4.3. HYPERPARAMETER OPTIMIZATION AND FEATURE IMPORTANCE

Following the ablation analysis, the modeling pipeline focused on the configuration that provided the most consistent contribution of thermal-gradient information, namely the squared abdomen–foot thermal-gradient representation. This section presents the optimization of the CatBoost models using this feature set and evaluates their performance across the four prediction horizons. In addition, feature-importance analysis was performed to examine how clinical and thermal variables contributed to the optimized model predictions over time.

4.3.1. Hyperparameter Optimization Table 6 summarizes the performance of the optimized CatBoost models across the four prediction horizons, including AUC, sensitivity, specificity, and Youden index.

Table 6. Clinical performance metrics for optimized models at selected operating thresholds

Model (Horizon)	AUC [95 % CI]	Sensitivity	Specificity	Youden index
CB (30 min)	0.838 [0.761–0.906]	0.660	0.874	0.534
CB (3 h)	0.820 [0.734–0.898]	0.730	0.775	0.505
CB (6 h)	0.737 [0.651–0.821]	0.548	0.823	0.371
CB (12 h)	0.669 [0.584–0.754]	0.518	0.740	0.258

After optimization, the CatBoost models preserved the temporal pattern observed during the baseline comparison, with stronger performance at shorter prediction horizons and lower performance as the horizon increased. The highest discriminative performance was observed at 30 minutes and 3 hours, where the optimized models achieved AUC values above 0.80 and the highest Youden indices. This indicates that the optimized models retained their strongest classification utility in the short-term prediction windows.

The operating profiles differed across horizons. At 30 minutes, the model showed high specificity with moderate sensitivity, suggesting a more conservative classification behavior. At 3 hours, the balance between sensitivity and specificity was more even, while maintaining a comparable level of global discrimination. This suggests that the 3-hour model provided the most balanced short-term operating profile among the optimized models.

At 6 and 12 hours, both discrimination and threshold-dependent performance decreased. The decline was more evident in sensitivity than in specificity, indicating that the optimized models became less effective at identifying future shock cases as the prediction horizon increased, while retaining a comparatively greater ability to identify no-shock cases.

Overall, Bayesian optimization refined the performance of the CatBoost models but did not change the main temporal pattern: prediction was strongest in the short-term horizons and progressively weakened at longer horizons. This result is consistent with the previous analyses, where thermal-gradient group separation and baseline model performance were also more pronounced closer to the shock event.

These optimized CatBoost models were then used for the subsequent interpretability analysis to examine the variables contributing to model predictions across prediction horizons.

4.3.2. SHAP Analysis Table 7 summarizes the mean absolute SHAP values for the main predictive variables in the optimized CatBoost models across prediction horizons. The reported variables include pulse rate (PR), age, and the squared thermal gradient feature $(\Delta T_{\%})^2$, which represented the selected thermal-gradient configuration after the ablation analysis.

Table 7. Mean absolute SHAP values for the main predictive features across optimized CatBoost models at each prediction horizon.

Model (Horizon)	PR	Age	$(\Delta T_{\%})^2$
CB (30 min)	0.3176	0.4534	0.1342
CB (3 h)	0.8573	1.1443	0.3061
CB (6 h)	0.3762	0.3824	0.1358
CB (12 h)	0.1909	0.2561	0.1222

Across the optimized CatBoost models, age and pulse rate showed the largest contributions to model output across all prediction horizons. This indicates that the final models relied primarily on baseline demographic and hemodynamic information when assigning shock risk. However, the squared thermal gradient also maintained a consistent contribution across horizons, indicating that thermal information added

a secondary but persistent signal to the prediction process.

The largest SHAP magnitudes for all three variables were observed at the 3-hour horizon. This pattern suggests that the 3-hour window represented the point at which the optimized model made the strongest use of both clinical variables and thermal-gradient information. In contrast, feature contributions were generally lower at 6 and 12 hours, consistent with the reduction in predictive performance observed at longer prediction horizons.

The squared thermal gradient did not exceed age or pulse rate in importance, but its contribution remained present in every optimized model. This finding is consistent with the ablation results, where the squared gradient representation provided the clearest benefit in the short-term windows. Together, these results indicate that thermal-gradient information acted as a complementary predictor rather than as the primary driver of model predictions.

Overall, the SHAP analysis showed that the optimized CatBoost models were mainly driven by age and pulse rate, while the squared abdomen–foot thermal gradient contributed additional information across time horizons. This supports the interpretation that thermal gradients enrich the predictive signal when combined with conventional clinical variables, particularly in short-term shock prediction.

5. DISCUSSION

5.1. MAIN FINDINGS AND INTERPRETATION

The results suggest a horizon-dependent relationship between abdomen–foot thermal gradients and future shock status. In the comparative analysis, patients classified as shock tended to present higher thermal gradients than patients classified as no shock, with clearer group separation at 30 minutes and 3 hours, where the confidence intervals for the estimated differences did not include zero and the p-values were below 0.05. At 6 and 12 hours, the estimated differences remained positive, but they were not statistically significant; therefore, no between-group difference can be concluded for these longer prediction horizons.

The baseline machine learning comparison showed a similar temporal structure. Model performance was generally higher at the shorter horizons and decreased as the prediction window increased. CatBoost showed the most consistent overall performance across horizons, particularly in the short-term windows, which motivated its use as the unified model family for the subsequent analyses. At the same time, the results showed that model ranking was not identical across metrics, especially at the 6-hour horizon, where AUC and Youden index favored different models. This highlights the distinction between global discrimination and threshold-dependent classification performance.

The ablation analysis provided a more specific view of the contribution of thermal-gradient features. The added value of thermal information was concentrated in the short-term horizons, where the squared thermal-gradient representation showed the clearest improvement. This pattern indicates that the magnitude of the abdomen–foot thermal imbalance was more informative than the raw directional gradient in these windows. At longer horizons, gradient-based configurations did not improve over the

baseline feature set, suggesting that the usefulness of thermal information varied with the prediction horizon.

After Bayesian optimization, the selected CatBoost models preserved the same temporal pattern observed in the previous analyses. Performance remained stronger at 30 minutes and 3 hours and decreased at 6 and 12 hours. Thus, the optimization stage refined model performance without changing the overall structure of the results across horizons.

The SHAP analysis added an interpretability layer to this pattern. Age and pulse rate were the dominant contributors to model predictions, while the squared thermal gradient showed a smaller but persistent contribution across horizons. This indicates that the thermal-gradient feature contributed additional information within the optimized models, although the main predictive structure remained driven by conventional clinical variables.

5.2. COMPARISON WITH OTHER STUDIES

Previous studies have explored thermographic information as a noninvasive marker of hemodynamic instability using different acquisition strategies. Static temperature measurements from single thermographic captures have been evaluated as potential indicators of circulatory alterations²⁷, while thermal video recordings have been used to characterize temporal thermal patterns associated with shock risk²⁶. These approaches support the general premise that alterations in skin temperature distribution may provide information related to circulatory status.

The present study differs from prior work by providing a more comprehensive evaluation of thermographic information across several analytical levels. First, it examined whether abdomen–foot thermal gradients differed between shock and no-shock groups across multiple prediction horizons, allowing the temporal behavior of the thermal signal to be characterized. Second, it evaluated the discriminative perfor-

mance of a broad set of machine learning models rather than relying on a single predictive approach. Third, it assessed the contribution of thermal-gradient features through ablation analysis and compared different representations of the gradient. Finally, SHAP analysis was used to place the thermal gradient in relation to conventional clinical variables, providing insight into whether thermographic information acted as a primary or complementary contributor to model predictions.

Whereas previous studies have commonly emphasized raw thermal measurements, peripheral-central temperature differences, or multimodal prediction strategies, the present analysis extends this perspective by examining how the thermal signal behaves across time, how it performs across model families, and how its contribution compares with established clinical predictors. In this sense, the value of thermographic information was not evaluated only as an isolated predictor, but as part of a broader predictive framework that considered temporal horizon, model architecture, feature representation, and feature importance.

5.3. METHODOLOGICAL CONSIDERATIONS

Several methodological decisions are important for interpreting the findings of this study. First, the analyses were designed to account for the repeated-measures structure of the data. This was relevant because multiple observations from the same patient cannot be considered statistically independent. Preserving patient-level structure in both the comparative and predictive analyses reduced the risk of overly optimistic estimates and provided a more appropriate framework for evaluating group differences and model performance.

Second, model performance was interpreted using both discrimination and threshold-dependent metrics. This distinction was important because a model with strong global ranking ability does not necessarily provide the best operating point for binary classification. The observed differences between AUC-based and Youden-based

rankings in some horizons illustrate why both perspectives are needed when evaluating prediction models intended for clinical use.

Third, the ablation analysis was structured to isolate the contribution of thermal-gradient features while keeping the remaining modeling conditions unchanged. This design strengthened the interpretation of the feature-impact analysis, since differences in performance could be more directly related to the inclusion or transformation of gradient information rather than to changes in the model family or experimental setup.

Finally, using CatBoost as a unified model family for subsequent optimization and interpretability analyses improved comparability across horizons and feature configurations. This choice was particularly relevant for SHAP interpretation, because feature-attribution values are model-dependent. Interpreting feature contributions within the same algorithmic framework reduced the risk of attributing differences in importance to the thermal variables when they might instead arise from differences in model architecture.

5.4. LIMITATIONS

This study was conducted using a single pediatric cohort with a relatively small sample size. This should be interpreted in the context of the limited availability of open pediatric thermographic datasets, particularly those including thermal-gradient information together with hemodynamic variables. To our knowledge, publicly available datasets with this specific combination of pediatric thermographic and shock-related measurements remain scarce. Future work should prioritize the development of larger, multicenter datasets with standardized thermographic acquisition protocols and longitudinal hemodynamic follow-up to enable external validation and broader evaluation of thermal-gradient-based prediction models.

6. CONCLUSIONS

This study aimed to evaluate whether peripheral thermal gradients, combined with baseline clinical variables, could support the early prediction of hemodynamic shock across multiple temporal horizons. Within the single-centre cohort analyzed, thermal gradient measurements showed a statistically significant difference between shock and non-shock patients only at the 30-minute horizon and 3 hours. At 6 and 12 hours, the observed differences did not reach statistical significance. Among the ten machine learning models compared, CatBoost achieved the highest discriminative performance at short horizons (30 min and 3 h) for both AUC and Youden index. At the 6-hour horizon, the choice of the best model depends on the metric: MLP for global discrimination (AUC) and SVM for the sensitivity–specificity trade-off (Youden). At 12 hours, logistic regression performed comparably to more complex models, suggesting a potentially more linear relationship at longer horizons.

The squared thermal gradient term was associated with a higher Youden index at 30 min and 3 h in our cohort, but this observation is preliminary and requires confirmation in larger, independent datasets. Beyond 6 hours, none of the gradient-containing configurations outperformed the baseline model. Hyperparameter optimisation of CatBoost consistently improved AUC across all horizons, while the Youden index increased at 30 min, 3 h and 6 h but decreased slightly at 12 h. SHAP analysis indicated that age and pulse rate were the dominant predictors across all horizons, with the squared thermal gradient showing a modest and consistent contribution, most notably at 3 hours.

Given the limited sample size and the current scarcity of paediatric thermographic datasets, these findings should be interpreted as exploratory. External validation in multicentre cohorts is essential before any clinical application. Future research should focus on optimising the Youden index directly for long-horizon predictions,

assessing clinical impact through implementation studies, and performing sensitivity analyses with thresholds adjusted to local prevalence.

BIBLIOGRAPHY.

Acker, Shannon N. et al. «Shock index, pediatric age-adjusted (SIPA) is more accurate than age-adjusted hypotension for trauma team activation». En: *Surgery* 161.3 (mar. de 2017). doi: 10.1016/j.surg.2016.08.050, págs. 803-807 (vid. págs. 12, 20).

Amson, Harry et al. «Core-to-skin temperature gradient measured by thermography predicts day-8 mortality in septic shock: A prospective observational study». En: *Journal of Critical Care* 60 (2020). doi: 10.1016/j.jcrc.2020.08.022, págs. 294-299 (vid. págs. 15, 16).

Berg, David D et al. «Epidemiology of Shock in Contemporary Cardiac Intensive Care Units». En: *Circ Cardiovasc Qual Outcomes* 12.3 (2019). PMC11032172, e005618. DOI: 10.1161/CIRCOUTCOMES.119.005618 (vid. pág. 10).

Berger, Tony et al. «Shock index and early recognition of sepsis in the emergency department: pilot study». En: *The Western Journal of Emergency Medicine* 14.2 (mar. de 2013). doi: 10.5811/westjem.2012.8.11546, págs. 168-174 (vid. pág. 11).

Bourcier, Simon et al. «Toe-to-room temperature gradient correlates with tissue perfusion and predicts outcome in selected critically ill patients with severe infections». En: *Annals of Intensive Care* 6.1 (2016). doi: 10.1186/s13613-016-0164-2, pág. 63 (vid. pág. 15).

Bridier, Armelle et al. «Hemodynamic assessment in children after cardiac surgery: A pilot study on the value of infrared thermography». En: *Frontiers in pediatrics* 11 (abr. de 2023). doi: 10.3389/fped.2023.1083962, pág. 1083962 (vid. pág. 15).

Bruns, Brock et al. «Prehospital hypotension redefined». En: *Journal of Trauma* 65.6 (dic. de 2008), págs. 1217-1221. DOI: 10.1097/TA.0b013e318184ee63 (vid. pág. 12).

Castro, R. et al. «Effects of capillary refill time-vs. lactate-targeted fluid resuscitation on regional, microcirculatory and hypoxia-related perfusion parameters in septic shock: a randomized controlled trial». En: *Ann. Intensive Care* 10.1 (2020), pág. 150. DOI: 10.1186/s13613-020-00767-4 (vid. pág. 13).

Cecconi, Maurizio et al. «Consensus on circulatory shock and hemodynamic monitoring. Task force of the European Society of Intensive Care Medicine». En: *Intensive Care Medicine* 40.12 (2014), págs. 1795-1815. DOI: 10.1007/s00134-014-3525-z (vid. pág. 13).

Cruz, Gustavo et al. «Capillary Refill Time and Serum Lactate as Predictors of Mortality and Postoperative Extracorporeal Membrane Oxygenation Requirement in Congenital Heart Surgery». En: *Children* 10.5 (2023). DOI: 10.3390/children10050875 (vid. pág. 14).

Dinçer, Ertuğ et al. «Evaluation of peripheral perfusion index in predicting blood product need for trauma patients». En: *Ulus Travma Acil Cerrahi Derg* 31.6 (2025). PMID: 40511758, págs. 525-530. DOI: 10.14744/tjtes.2025.65957. eprint: <https://dx.doi.org/10.14744/tjtes.2025.65957> (vid. pág. 11).

Fluss, Ronen, David Faraggi y Benjamin Reiser. «Estimation of the Youden Index and its associated cutoff point». En: *Biometrical Journal* 47.4 (ago. de 2005). doi: 10.1002/bimj.200410135, págs. 458-472. DOI: 10.1002/bimj.200410135 (vid. pág. 25).

Garcia, Pedro Celiny Ramos, Cristian Tedesco Tonial y Jefferson Pedro Piva. «Septic shock in pediatrics: the state-of-the-art». En: *Jornal de Pediatria* 96 Suppl 1.Suppl 1 (2020). doi: 10.1016/j.jpmed.2019.10.007, págs. 87-98 (vid. pág. 12).

Guido, Tavazzi et al. «Cardiogenic shock in general intensive care unit: a nationwide prospective analysis of epidemiology and outcome». En: *Eur Heart J Acute Cardiovasc Care* 13.11 (2024), págs. 768-778. DOI: 10.1093/ehjacc/zuae108 (vid. pág. 10).

Hasanin, Ahmed et al. «The ability of pulse oximetry-derived peripheral perfusion index to detect fluid responsiveness in patients with septic shock». En: *Journal of Anesthesia* 35.2 (abr. de 2021), págs. 254-261. DOI: 10.1007/s00540-021-02908-w (vid. pág. 11).

Hassanzad, Mojtaba y Karimollah Hajian-Tilaki. «Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review». En: *BMC Medical Research Methodology* 24.1 (2024). doi: 10.1186/s12874-024-02198-2, pág. 84.

Jacquet-Lagrèze, Matthias et al. «Prognostic value of capillary refill time in adult patients: a systematic review with meta-analysis». En: *Critical Care* 27.1 (2023), pág. 473. DOI: 10.1186/s13054-023-04751-9.

Koch, Erica et al. «Shock index in the emergency department: utility and limitations». En: *Open Access Emergency Medicine* 11 (2019). doi: 10.2147/OAEM.S178358, págs. 179-199 (vid. pág. 11).

Lath, Vrinda et al. «Utility of core to peripheral temperature gradient using infrared thermography in the assessment of patients with sepsis and septic shock in the emergency medicine department». En: *International Journal of Emergency Medicine* 18 (2025). DOI: 10.1186/s12245-025-00890-8 (vid. pág. 15).

Magnin, Mathieu et al. «Use of infrared thermography to detect early alterations of peripheral perfusion: evaluation in a porcine model». En: *Biomed Opt Express* 11.5 (8 de abr. de 2020). doi: 10.1364/BOE.387481, págs. 2431-2446 (vid. pág. 14).

Martos-Benítez, Félix D. y Bárbara I. Guzmán-Breff. «Volume expansion and variation in hemodynamic parameters». En: *Emergencias* 30.3 (jun. de 2018). English, Spanish, págs. 177-181. PMID: 29687672 (vid. pág. 11).

Nagori, Aditya et al. «Predicting Hemodynamic Shock from Thermal Images using Machine Learning». En: *Scientific Reports* 9 (ene. de 2019). doi: 10.1038/s41598-018-36586-8 (vid. págs. 16, 19, 21, 39).

Noitz, Matthias et al. «The Predictive Value of Clinical Signs to Identify Shock in Critically Ill Patients». En: *Diagnostics* 15.17 (2025). DOI: 10.3390/diagnostics15172252 (vid. págs. 11, 14).

Ochagavía, A. et al. «Hemodynamic monitoring in the critically patient. Recommendations of the Cardiological Intensive Care and CPR Working Group of the Spanish Society of Intensive Care and Coronary Units». En: *Medicina Intensiva (English Edition)* 38.3 (2014). doi: 10.1016/j.medine.2013.10.002, págs. 154-169 (vid. pág. 10).

Ortiz-Dosal, Alejandra et al. «Use of infrared thermography in children with shock: A case series». En: *SAGE Open Medical Case Reports* 2 (2014). doi: 10.1177/2050313X14561779. pág. 2050313X14561779.

Ospina-Tascón, Gustavo A. et al. «Diastolic shock index and clinical outcomes in patients with septic shock». En: *Annals of Intensive Care* 10.1 (2020). doi: 10.1186/s13613-020-00658-8, pág. 41 (vid. pág. 13).

Phillips, Ryan et al. «The shock index, pediatric age-adjusted (SIPA) enhanced: Pre-hospital and emergency department SIPA values forecast transfusion needs for blunt solid organ injured children». En: *Surgery* 168.4 (2020). doi: 10.1016/j.surg.2020.04.061, págs. 690-694 (vid. pág. 12).

Rady, Mohamed Y. et al. «Shock index: a re-evaluation in acute circulatory failure». En: *Resuscitation* 23.3 (1992). doi: 10.1016/0300-9572(92)90006-X, págs. 227-234 (vid. pág. 13).

Ruopp, Marcus D et al. «Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection». En: *Biometrical Journal. Biometrische Zeitschrift* 50.3 (2008). doi: 10.1002/bimj.200710415, págs. 419-430.

Sethi, Tavpritesh. *Data for manuscript on automated shock detection in the ICU using Thermal Imaging*. doi: 10.17605/OSF.IO/VP86J. 2018 (vid. pág. 19).

Strehlow, Matthew C. «Early identification of shock in critically ill patients». En: *Emergency Medicine Clinics of North America* 28.1 (feb. de 2010), págs. 57-66. DOI: 10.1016/j.emc.2009.09.006 (vid. pág. 11).

Sudhi, Manu et al. «Thermal Imaging Applications in Shock Detection: Technological Advancements and Clinical Implications». En: *Engineered Science* 33 (2025). doi: 10.30919/es1367, pág. 1367.

Vardasca, Ricardo, Carolina Magalhaes y Joaquim Mendes. «Biomedical Applications of Infrared Thermal Imaging: Current State of Machine Learning Classification». En: *Proceedings* 27.1 (2019). doi: 10.3390/proceedings2019027046.

Vats, Vanshika et al. «Early Prediction of Hemodynamic Shock in Pediatric Intensive Care Units With Deep Learning on Thermal Videos». En: *Frontiers in Physiology* 13 (2022). doi: 10.3389/fphys.2022.862411 (vid. págs. 15, 16, 39).

Vincent, Jean-Louis et al. «Frequency and mortality of septic shock in Europe and North America: a systematic review and meta-analysis». En: *Critical Care* 23.1 (2019), pág. 196. DOI: 10.1186/s13054-019-2478-6 (vid. pág. 10).

ANNEXES

Annex A. OPTIMIZED CATBOOST HYPERPARAMETERS

After the initial model comparison, CatBoost was selected as the model family for the optimization stage. Table 8 reports the final hyperparameter configuration obtained for each prediction horizon. These configurations correspond to the optimized CatBoost models used in the final performance evaluation.

Table 8. Optimized CatBoost hyperparameters for each prediction horizon.

Hyperparameter	30 min	3 h	6 h	12 h
depth	6	10	4	3
learning_rate	0.0037	0.0104	0.0016	0.0036
l2_leaf_reg	10	9	1	6
subsample	0.2005	0.9449	0.8911	0.2470
colsample_bylevel	0.2294	0.2617	0.9508	0.4355
iterations	2403	2132	670	537
border_count	158	234	54	46
random_strength	9.2365	4.9623	0.5295	5.6884
bagging_temperature	0.1946	0.9432	0.7047	0.5799
grow_policy	Lossguide	SymmetricTree	SymmetricTree	SymmetricTree
min_data_in_leaf	19	8	17	1
one_hot_max_size	2	4	5	8
leaf_estimation_iterations	8	5	10	6
scale_pos_weight	2.8631	2.0162	2.6985	2.7716