Predicting Shock in Pediatric Patients Through Thermal Gradients and Machine Learning: A Multi-Model Approach

Juan D. Espinoza, and Carlos A. Fajardo

***Abstract*—Early detection of hemodynamic compromise in pediatric patients is critical for timely and effective intervention in intensive care. This study evaluates the use of thermal gradients, specifically the temperature difference between the abdomen and foot, as non-invasive physiological markers to improve prediction of shock. The dataset included thermal gradients, pulse rate, age, and four time-stamped measurements, enabling models to anticipate circulatory deterioration across different prediction horizons. These forecasting windows were examined to assess how far in advance the onset of shock could be reliably predicted. Several machine learning models were compared, and the best approach achieved an AUC of 0.8371 (95% CI: 0.7794 - 0.8818), with sensitivity of 0.710 (95% CI: 0.615-0.790) and specificity of 0.888 (95% CI: 0.826-0.930). Although methodological differences make direct comparison with previous studies challenging, this performance surpasses that reported in the existing literature. These findings highlight the potential of combining thermal gradients with conventional vital signs to enhance early and reliable risk stratification and support clinical decision-making in pediatric intensive care.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/10224

***Index Terms*—Shock Index, hemodynamic monitoring, pediatric critical care, machine learning, circulatory compromise, non-invasive assessment.**

I. Introduction

THE Shock Index (SI), calculated as the ratio between heart rate and systolic blood pressure, has become a valuable tool to assess circulatory status [1], [2]. This ratio reflects how the heart and blood vessels work together to keep blood flowing even when blood pressure drops, the heart often beats faster to maintain circulation, making SI a useful early sign of problems with blood flow. Unlike individual vital signs, SI gives a broader picture of how well the cardiovascular system functions, which makes it particularly useful to spot patients who may be at risk for hemodynamic instability. In pediatrics, normal heart rate and blood pressure values change significantly with age. For this reason, the Shock Index Pediatric Age-adjusted (SIPA) was developed to improve clinical utility. SIPA applies age-specific thresholds to better identify children at risk of hemodynamic compromise [3], [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Samuel Ortega (*Corresponding author: Carlos Fajardo*).

J. D. Espinoza, and Carlos Fajardo are with the Department of Electrical, Electronics and Telecommunications, Universidad Industrial de Santander, Bucaramanga, Colombia (e-mails: juan2248438@correo.uis.edu.co, and cafajar@uis.edu.co).

Detecting early signs of circulatory compromise is crucial, especially in acute care or after surgery. Performing quickly can prevent the situation from escalating [5]. However, many current monitoring methods are either invasive or provide only occasional snapshots of a patient's condition. That's where SI stands out: it is simple, non-invasive, and easy to calculate, offering real-time insight into the cardiovascular health of a patient.

In pediatric critical care, this becomes even more important. Children often compensate for a failing circulatory system for a relatively long time, which means that by the time symptoms such as low blood pressure appear, serious organ damage may already be underway [6]. Since SI captures both heart rate and systolic blood pressure, it can help detect early warning signs , giving healthcare teams the opportunity to intervene sooner [7], [8].

On the other hand, researchers have started looking into how SI relates to peripheral perfusion, and thermal gradients - differences in temperature between central and peripheral body regions - are gaining attention as another non-invasive biomarker to circulatory health. Some studies [9]–[13] show that when blood flow is impaired, less heat reaches the extremities, creating measurable temperature differences. These variations may reflect the body's efforts to redirect blood to essential organs during times of stress [12].

This work aims to identify the most effective machine learning models and the most reliable prediction horizon to forecast shock in pediatric patients, using the age-adjusted Shock Index as the diagnostic criterion, across multiple time windows. To achieve this, we use thermal gradients as non-invasive biomarkers and evaluate the performance of different modeling strategies, examining their potential to support predictive monitoring in intensive care or post-surgical settings.

II. Related Works

Infrared thermography has gained increasing attention as a non-invasive technique for assessing peripheral perfusion and detecting hemodynamic instability. Over the past decade, the field has evolved from broad application reviews to experimental validation in animal models, and more recently to predictive approaches leveraging machine learning and deep learning.

Among these advances a key experimental validation of the relationship between thermal gradients and perfusion emerged in 2020 through controlled animal studies [14]. The authors investigated how pharmacologically induced hemodynamic

variations in pigs influenced the thermal gradients between the central and peripheral body regions. Their findings demonstrated consistent changes in the temperature distribution in perfusion states, reinforcing the physiological rationale for using thermal imaging in perfusion assessment.

These experimental findings are based on a well-established foundation of Infrared Thermography (IRT) research, as comprehensively documented in [15]. This foundational review examines the biomedical applications of infrared thermal imaging between 2003 and 2019, including clinical uses in inflammation detection, vascular disorders, and perfusion monitoring. Notably, the review also tracks the field's technological evolution, highlighting the increasing adoption of artificial intelligence algorithms - particularly artificial neural networks (ANN), K-Nearest Neighbors (k-NN), and Support Vector Machine (SVM) for classification tasks in thermal imaging analysis.

Building on this technological trajectory, Sudhi et. al. provides a focused examination of thermography's potential for shock detection and monitoring. The study expands the clinical perspective by specifically evaluating thermography as a non-invasive, fast-response tool for early shock detection and continuous monitoring. Through comparative analysis with traditional techniques, it demonstrates the advantages of thermal imaging in sensitivity, specificity, and early warning capability, benefits further amplified by recent advances in sensor technology and AI system integration [16].

Expanding methodological approaches, Nagori et. al demonstrated that even static thermal measurements contain valuable predictive information when combined with basic clinical parameters. Using a generalized linear mixed model, the study successfully identified patients at risk of hemodynamic shock by integrating thermal gradient patterns with routinely available physiological data. This approach proved particularly valuable in clinical scenarios where continuous monitoring may not be feasible, showing that snapshot thermography combined with standard clinical markers can provide actionable risk stratification. The model achieved an AUC of 0.79, however, the study did not report confidence intervals, limiting a more precise assessment of its robustness and generalization [17].

The clinical applicability of such snapshot thermography approaches was further validated in critically ill populations by Amson et. al in their 2020 study of septic patients. Their protocol involved capturing a single thermal image within 24 hours after initial vasopressor administration, then applying logistic regression to assess the risk of mortality at eight-day. Notably, the study confirmed that specific temperature thresholds and core-to-skin thermal gradients were significant predictors, reinforcing the utility of thermal monitoring in intensive care settings where rapid risk assessment is critical [11].

Building upon these foundational findings with static measurements, the field has progressively shifted toward dynamic temporal analysis. This evolution is exemplified by [9], where researchers employed long- and short-term memory (LSTM) networks to analyze thermal video sequences of approximately 4.26 minutes. By extracting thermal gradients and corresponding heart rate data at one-second intervals, their model successfully predicted a binarized shock index (SIPA), demonstrating deep learning's capacity to unlock clinically relevant patterns from sequential thermal data. The model achieved an AUC of 0.81, with a reported standard error of 0.06, reflecting encouraging results while also highlighting the need for further refinement to enhance consistency and reliability.

However, while such advanced analytical approaches show great promise, several fundamental challenges persist in translating these findings into clinical practice. Despite the growing adoption of thermography for continuous monitoring and promising results from infrared thermography combined with machine learning (IRT-ML) in pediatric intensive care, the field remains limited. Although simpler linear models have been shown to provide some predictive insights, they struggle to capture the complex relationships between thermal data and clinical conditions. However, more complex models have been explored, but these often yield results with high variability, limiting their clinical applicability.

To address these limitations, our study leverages publicly available data to evaluate a range of model architectures. We then identify those that achieve the best performance across different prediction horizons, after which we fine-tune decision thresholds to optimize the trade-off between sensitivity and specificity using Youden's J statistic (Youden Index), a standard diagnostic method for determining the threshold that maximizes their combined values. This process represents an essential step toward integrating these algorithms into clinical decision support systems [18], [19].

III. Methods

A. Dataset

The dataset, collected and described by Nagori et al. [17], is publicly available on the Open Science Framework [20] under the CC0 1.0 Universal license. Data collection involved obtaining abdominal and foot temperature readings via thermography, in addition to recording each patient's pulse rate and age in months. Subsequently, hemodynamic parameters were recorded at 30-minute, 3-hour, 6-hour, and 12-hour intervals.

For data preprocessing, all records with missing values in any predictor variable were removed. Specifically, observations were excluded if they lacked foot temperature, abdominal temperature, age in months, or pulse rate. In addition, records were required to include systolic blood pressure (SBP) and heart rate (HR) to compute the shock index.

Subsequently, patients were classified as being in shock or not based on the Shock Index, Pediatric Age-Adjusted (SIPA) score, which applies age-specific critical thresholds. A binary label was assigned accordingly: a value of 1 (indicating shock) if a patient's shock index exceeded their age-specific threshold, and 0 (indicating no shock) otherwise.

B. Machine Learning Models

We conducted a systematic evaluation of ten machine learning models (six tree-based and four non-tree-based) to benchmark their performance in the early prediction of shock across

multiple time horizons. The tree-based group encompassed Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, and CatBoost, while the non-tree based models included Logistic Regression, SVM, K-NN and Multilayer Perceptron (MLP).This comprehensive evaluation allowed us to analyze how different model architectures respond to varying temporal windows, offering insights into their robustness and predictive capacity in the context of critical care. For the MLP model included in this comparison, we used the implementation provided by scikit-learn, configuring fully connected layers with ReLU activations, the Adam optimizer (learning rate = 0.001), Xavier uniform initialization, and a mini-batch size of $\min(200, n_{\text{samples}})$. Training proceeded for a maximum of 200 iterations under scikit-learn's internal convergence criterion, which halts optimization when the loss fails to improve for 10 consecutive iterations. In addition to the standard optimization settings, the model incorporated scikit-learn's built-in L2 regularization (strength = 0.0001), whose contribution to the objective function is scaled by the sample size. No other regularization strategies—such as dropout, weight decay beyond the L2 term, or normalization layers—were applied. Parameter counts varied according to the evaluated architecture.

The specific MLP architectures used for each prediction horizon corresponded to the configurations that achieved the highest average AUC computed over the test partitions of a 10-fold cross-validation procedure.Each tuple representing an architecture denotes the sequence of hidden layers, where each element specifies the number of neurons in that layer. The selected architectures were: (16, 8) for the 30-minute horizon, comprising 209 trainable parameters; (512, 256, 128, 64, 32, 16) for the 3-hour horizon, comprising 177,153 trainable parameters; and (16, 8) for both the 6-hour and 12-hour horizons, similarly totaling 209 parameters. Using these horizon-adapted configurations ensured that the MLP model reflected the most suitable network structure for each temporal window in the comparative evaluation, allowing the network to achieve a reasonable and consistent performance across the different prediction horizons.

To improve the discriminative power of the input features, the squared gradients of temperature differences between the abdominal and distal (foot) regions were employed during model training. This transformation was applied because the squared term amplifies the contribution of larger thermal variations, which are of greater clinical relevance as they more likely represent significant hemodynamic alterations indicative of shock. This enhancement allows the models to focus more effectively on detecting these substantial thermal variation patterns.

Furthermore, a stratified grouped splitting strategy was employed for data partitioning. The grouped approach ensured that all samples from a single patient were assigned exclusively to one fold, thereby preventing data leakage. Concurrently, the stratified component maintained a proportional distribution of the target variable classes across all folds, mitigating potential bias from class imbalance during model evaluation. Building on this careful data partitioning, a 10-fold cross-validation strategy was employed to maximize the use of the limited dataset and enhance statistical robustness. The grouping was carried out using the patient identifier, ensuring that all observations belonging to the same individual were consistently allocated to a single fold. All model training and evaluation steps were implemented using StratifiedGroupKFold (scikit-learn), guaranteeing fold integrity by patient. Within every fold, preprocessing—particularly z-score standardization (StandardScaler)—was fitted exclusively on the training split and subsequently applied to the corresponding test split. This fold-specific workflow kept all preprocessing and model-fitting operations isolated per fold and grouped by patient, effectively preventing information leakage.

Given the limited sample size (an average of approximately 253 samples per horizon from 51 patients) and the unequal number of observations per individual, introducing an additional validation split would have further reduced fold representativeness and stability. Consequently, no validation-based early stopping policy was applied. This decision was also aligned with the methodological constraints of the study: such validation-driven early stopping is not supported within the scikit-learn environment for some of the evaluated models—particularly support vector machines, decision trees, random forests, and k-nearest neighbors—which do not implement early stopping at all. Applying it only to a subset of models would have resulted in an inconsistent and non-comparable training procedure across model families. To maintain consistency and reproducibility across all experiments, the cross-validation procedure employed StratifiedGroupKFold with shuffling and a fixed random seed of 42, ensuring stable and comparable fold generation across model families and prediction horizons.

Model selection was then performed using the mean AUC obtained on the test partitions across the 10 folds, ensuring that the chosen configuration for each model corresponded to the one that demonstrated the most stable and consistent generalization under the patient-grouped evaluation scheme. To complement these estimates with an assessment of statistical uncertainty, 95% confidence intervals were computed using a nonparametric bootstrap procedure with 5,000 resamples, implemented using a non-parametric bootstrap procedure available within the SciPy statistical framework, with a fixed random seed of 40 to ensure reproducibility.

For each time interval, we selected the best-performing model for Bayesian hyperparameter optimization using the Optuna framework. The optimal decision threshold was determined by maximizing the Youden index (J = sensitivity + specificity $-$ 1). This index provides a balanced criterion that equally weights sensitivity and specificity, ensuring reproducible model comparisons without bias toward either metric, consistent with established practices in hemodynamic prediction [9], [17]. The Youden index is particularly recommended when no explicit cost function or prevalence-adjusted objective is specified [18], [21], making it appropriate for this exploratory study aimed at identifying broadly discriminative models. For future clinical implementation, the optimization criterion could be tailored to specific operational requirements—such as emphasizing sensitivity in critical care settings or specificity in screening scenarios.

After identifying the optimal hyperparameters for each horizon, we conducted additional tests to characterize the inference time of the selected models. Following Bayesian optimization, a cross-validation procedure was performed, and the best-performing fold—identified by its highest AUC—was retained to provide a stable and representative basis for inference-time evaluation. All measurements were executed on a CPU-only environment (Intel Core i9-13900H, 16 GB RAM, Windows 10), without GPU acceleration, reflecting a realistic deployment scenario for point-of-care or resource-constrained clinical settings.

IV. Results

A. Performance of Non-Tree-Based Models

Fig. 1 shows the performance of non–tree-based models in predicting shock within the next 30 minutes during the initial monitoring period. Logistic Regression achieved the highest discriminative ability among these models, with an AUC of 0.7658 (95% CI: 0.6991–0.8255). The Multilayer Perceptron (MLP) demonstrated comparable performance, with an AUC of 0.7490 (95% CI: 0.6785–0.8055), suggesting that both models are capable of effectively distinguishing between patients at risk of shock.

In comparison, the Support Vector Machine (SVM) exhibited a lower AUC of 0.7253 (95% CI: 0.6496–0.7892), indicating moderate predictive capacity and some sensitivity to variability in the training data. The K-Nearest Neighbors (KNN) classifier showed the lowest performance, with an AUC of 0.6995 (95% CI: 0.6301–0.7657), highlighting its limited discriminative power in this setting and potential susceptibility to noise or less informative features.

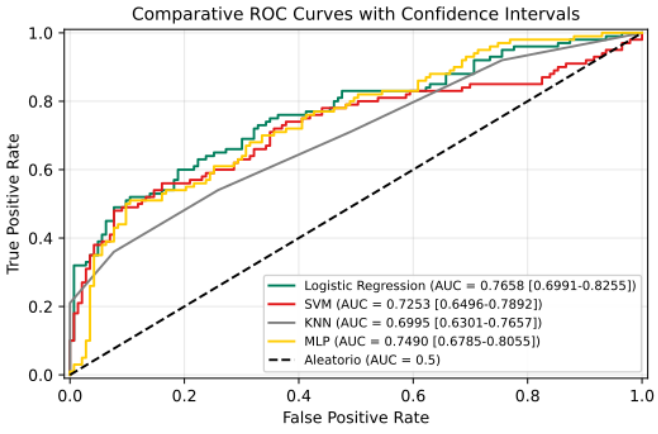


Fig. 1. Comparative ROC Curves with 95% confidence intervals for Non-Tree-Based Models

B. Performance of Tree-Based Models

Fig. 2 summarizes the performance of tree-based models during the first 30 minutes of observation, encompassing both individual algorithms and ensemble learning methods. Among these models, CatBoost demonstrated the highest performance, achieving an AUC of 0.8141 (95% CI: 0.7482–0.8652), indicating strong discriminative ability coupled with stable performance across data partitions.

This superior performance can be attributed to inherent features of the algorithm, including the use of symmetric trees, ordered encoding of categorical variables, and robust strategies to prevent target leakage, all of which enhance generalization in complex datasets.

Conversely, the Decision Tree model exhibited the lowest performance, with an AUC of 0.6846 (95% CI: 0.6238-0.7418), reflecting its limited capacity to model complex relationships without ensemble or regularization techniques. The remaining boosting-based methods—Gradient Boosting, XGBoost, and LightGBM—achieved AUCs ranging from 0.77 to 0.78, with relatively narrow confidence intervals, underscoring their ability to capture predictive patterns even in high-dimensional conditions. Overall, these results support the use of ensemble methods as more reliable and accurate alternatives in complex clinical scenarios from the earliest stages of evaluation.

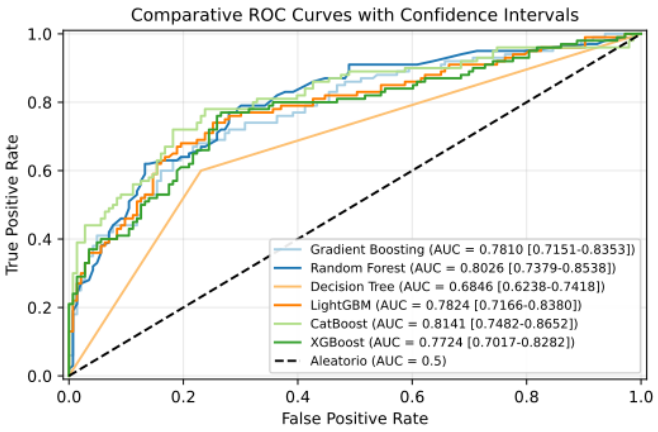


Fig. 2. Comparative ROC Curves with 95% confidence intervals for Tree-Based Models

C. Comparison by Prediction Time Horizon

Table I summarizes the performance of the evaluated models across different prediction time horizons. CatBoost without hyperparameter tuning consistently outperformed the other algorithms at the 30-minute and 3-hour horizons, achieving the highest AUC values.At the 6-hour horizon, MLP emerges as the superior model, surpassing CatBoost in predictive performance with a higher AUC (0.6636 (95% CI: 0.5926–0.7330) vs. 0.7052 (95% CI: 0.6352–0.7709)).

Conversely, at the 12-hour horizon, Logistic Regression achieved the best performance, with results closely matching those of catboost. Both models showed similar AUC values, confidence intervals, and predictive power, suggesting that Logistic Regression, despite its simplicity, performs comparably to more complex models in this scenario.

These findings reveal a decreasing trend in overall performance as the prediction horizon increases, reflecting growing clinical uncertainty and a decline in the discriminative power of the available variables when attempting to anticipate events further in advance.

TABLE I
AUC WITH 95% CONFIDENCE INTERVALS FOR EACH MODEL, ORGANIZED BY PREDICTION HORIZON

Model	30-min	3-hour	6-hour	12-hour
Logistic Regression	0.7658 [0.6991–0.8255]	0.7662 [0.6946–0.8242]	0.6473 [0.5737–0.7179]	**0.6840 [0.6125–0.7487]**
MLP	0.7490 [0.6785–0.8055]	0.7770 [0.7083–0.8323]	**0.7052 [0.6352–0.7709]**	0.6774 [0.6074–0.7440]
SVM	0.7253 [0.6496–0.7892]	0.7554 [0.6850–0.8154]	0.6468 [0.5708–0.7199]	0.6457 [0.5737–0.7144]
KNN	0.6995 [0.6301–0.7657]	0.7111 [0.6403–0.7709]	0.6319 [0.5589–0.7032]	0.6265 [0.5564–0.6963]
Gradient Boosting	0.7810 [0.7151–0.8353]	0.7691 [0.7021–0.8245]	0.6452 [0.5703–0.7156]	0.6255 [0.5528–0.6930]
XGBoost	0.7724 [0.7017–0.8282]	0.7531 [0.6877–0.8114]	0.5982 [0.5232–0.6707]	0.6276 [0.5522–0.6948]
CatBoost	**0.8141 [0.7482–0.8652]**	**0.7934 [0.7279–0.8452]**	0.6636 [0.5926–0.7330]	0.6809 [0.6101–0.7444]
Decision Tree	0.6846 [0.6238–0.7418]	0.7087 [0.6483–0.7650]	0.5572 [0.4923–0.6213]	0.5710 [0.5072–0.6338]
Random Forest	0.8026 [0.7379–0.8538]	0.7639 [0.6952–0.8202]	0.6548 [0.5832–0.7227]	0.6358 [0.5627–0.7036]
LightGBM	0.7824 [0.7166–0.8380]	0.7645 [0.6961–0.8224]	0.6654 [0.5924–0.7340]	0.6724 [0.6012–0.7368]

D. Fine-Tuning Process

TABLE II
OPTIMIZED CATBOOST HYPERPARAMETERS FOR EACH PREDICTION HORIZON

Hyperparameter	30 min	3 hr
depth	6	10
learning_rate	0.00372	0.0104
l2_leaf_reg	10	9
subsample	0.201	0.945
colsample_bylevel	0.229	0.262
iterations	2403	2132
border_count	158	234
random_strength	9.237	4.962
bagging_temperature	0.195	0.943
grow_policy	Lossguide	SymmetricTree
min_data_in_leaf	19	8
one_hot_max_size	2	4
leaf_estimation_iterations	8	5
scale_pos_weight	2.863	2.016

TABLE III
SELECTED HYPERPARAMETERS FOR THE MULTILAYER PERCEPTRON (MLP)

Hyperparameter	Value
activation	relu
solver	adam
learning_rate_init	0.00117
max_iter	598
batch_size	64
beta_1	0.964
beta_2	0.902
epsilon	1.710e-09

Model selection was driven by superior AUC performance at each respective horizon: CatBoost for the 30-minute and 3-hour intervals, and MLP for the 6-hour interval. To optimize generalization capacity, we conducted separate Bayesian hyperparameter optimizations using Optuna with 10-fold stratified cross-validation, targeting AUC maximization. The search spaces encompassed 14 parameters for CatBoost and 11 for MLP .Tables II and III present the optimal configurations obtained for CatBoost and MLP, respectively, across all prediction horizons.

For the 12-hour horizon, we directed the optimization process toward Logistic Regression, as it demonstrated the most competitive baseline performance at this interval. The tuning considered seven key hyperparameters associated with regularization, convergence criteria, and model specification. The resulting configurations for each parameter are summarized in Table IV, which reports the values obtained after optimization.

TABLE IV
OPTIMIZED LOGISTIC REGRESSION HYPERPARAMETERS FOR THE 12-HOUR PREDICTION HORIZON

Parameter	12 hr
C	0.0435
penalty	l2
l1_ratio	None
max_iter	32494
class_weight	balanced
tol	5.0267e-05
fit_intercept	True

Table V presents the results obtained after tuning the selected models. A notable improvement in the discriminative capacity of the CatBoost model is observed for the 30-minute interval, with the AUC increasing from 0.8141 (95% CI: 0.7482–0.8652) to 0.8371 (95% CI: 0.7794 - 0.8818). Substantial gains were also observed at the 3-hour horizon, where the AUC rose from 0.7934 (95% CI: 0.7279–0.8452) to 0.8214 (95% CI: 0.7606 - 0.8712), and at the 6-hour horizon, where it increased from 0.7052 (95% CI: 0.6352–0.7709) to 0.7259 (95% CI: 0.6546 - 0.7875). For the adjusted logistic regression at the 12-hour horizon, hyperparameter optimization yielded a measurable improvement in AUC from 0.6840 (95% CI: 0.6125–0.7487) to 0.7108 (95% CI: 0.6413–0.7738), accompanied by a modest reduction in confidence interval width.

TABLE V
AUC RESULTS WITH 95% CONFIDENCE INTERVALS AND STANDARD ERRORS (SE) FOR THE TUNED MODELS AT EACH PREDICTION INTERVAL

Interval	Model	AUC [95% CI]	SE
30 min	CatBoost	0.8371 [0.7794 – 0.8818]	0.0261
3 h	CatBoost	0.8214 [0.7606 – 0.8712]	0.0282
6 h	MLP	0.7259 [0.6546 – 0.7875]	0.0339
12 h	Logistic Regression	0.7108 [0.6413 – 0.7738]	0.0338

E. Optimal Decision Threshold Selection

Considering the critical stakes of clinical decision-making, where false positives can precipitate unnecessary tests or

TABLE VI
PERFORMANCE METRICS FOR EACH MODEL AND PREDICTION HORIZON. THE TWO CATBOOST MODELS CORRESPOND TO THE 30-MINUTE AND 3-HOUR HORIZONS (WITH THE FIRST CATBOOST ROW REPRESENTING THE 30-MINUTE MODEL), WHILE THE MLP AND LOGISTIC REGRESSION MODELS CORRESPOND TO THE 6-HOUR AND 12-HOUR HORIZONS, RESPECTIVELY

Model	AUC [95% CI]	Sensitivity [95% CI]	Specificity [95% CI]	PPV [95% CI]	NPV [95% CI]	Threshold
CatBoost	**0.837 [0.780 - 0.882]**	**0.660 [0.563 - 0.745]**	**0.874 [0.810 - 0.919]**	**0.786 [0.687 - 0.860]**	**0.786 [0.716 - 0.843]**	0.684
CatBoost	0.821 [0.760 - 0.869]	0.730 [0.636 - 0.807]	0.775 [0.699 - 0.837]	0.702 [0.608 - 0.781]	0.799 [0.723 - 0.858]	0.579
MLP	0.726 [0.655 - 0.789]	0.683 [0.588 - 0.764]	0.686 [0.599 - 0.761]	0.646 [0.553 - 0.729]	0.720 [0.633 - 0.793]	0.413
Logistic Regression	0.711 [0.643 - 0.771]	0.795 [0.711 - 0.859]	0.567 [0.480 - 0.650]	0.618 [0.537 - 0.693]	0.758 [0.663 - 0.833]	0.449

treatment and false negatives can postpone life-saving interventions, rigorous threshold optimization is imperative. Once the hyperparameters of the best performing models for each prediction horizon were finalized, we proceeded to define the decision threshold that would yield the greatest clinical benefit. To achieve this, we evaluated a range of potential cutoff values across every model's output probability and selected the one that maximized the Youden Index (sensitivity plus specificity minus one). Optimizing this index ensures that both true positive and true negative rates are maximized. To ensure robust threshold selection, we aggregated predicted probabilities across all cross-validation folds and computed a single optimal threshold using the Youden index, thereby enhancing generalizability beyond individual data splits.

Table VI summarizes the comprehensive performance evaluation for each model across all prediction horizons. We report AUC, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) with their respective 95% confidence intervals, along with the Youden index.

For short-term horizons of 30 minutes, 3 hours, the CatBoost model exhibited strong performance with relatively low variability. At the 30-minute horizon, the model demonstrated strong discriminatory performance with a sensitivity of 0.660 (95% CI: 0.563-0.745) and specificity of 0.874 (95% CI: 0.810-0.919). This performance profile persisted at the 3-hour horizon, maintaining sensitivity above 0.6 while showing a decrease in specificity to 0.775 (95% CI: 0.699-0.837). These consistent metrics underscore the model's reliability for early warning applications across multiple time points.

At the 6-hour prediction horizon, the MLP model demonstrated balanced diagnostic performance with moderate sensitivity (0.683; 95% CI: 0.588–0.764) and specificity (0.686; 95% CI: 0.599–0.761), indicating consistent discriminative capability for medium-term forecasting

For the 12-hour horizon, the tuned logistic regression model exhibited high sensitivity (0.795; 95% CI: 0.711–0.859) but diminished specificity (0.567; 95% CI: 0.480–0.650). This pattern aligns with the observed trend across increasing prediction windows: specificity demonstrates a consistent decline, while sensitivity fluctuates without a clear directional trend. Notwithstanding the reduction in specificity, the model preserves an acceptable trade-off between precision and recall, as indicated by a PPV of 0.618 (95% CI: 0.537–0.693) and an NPV of 0.758 (95% CI: 0.663–0.833). These metrics underscore the model's potential utility for longer-term prediction applications, despite the evolving performance characteristics.

F. Inference Time and Operational Feasibility

Inference time evaluation followed a standardized procedure in which the best-performing fold (selected according to its test AUC) was loaded together with its corresponding scaler, and inference was repeatedly measured on randomly sampled instances from the held-out test partition. For each model and prediction horizon, 100 independent inference calls were executed, ensuring that timing estimates reflected realistic deployment conditions, including preprocessing overhead associated with feature scaling. Across all scenarios, inference times remained well within operational limits for real-time clinical triage. Logistic Regression consistently yielded the lowest latency, followed by MLP and CatBoost, with all models achieving mean inference times comfortably below thresholds required for rapid decision-support or continuous monitoring workflows. Despite architectural differences and varying computational footprints, all models demonstrated inference speeds compatible with near–real-time deployment in resource-constrained clinical environments.

TABLE VII
INFERENCE TIME (MS) ACROSS MODELS AND HORIZONS

Model	Horizon	Inference Time (ms)	Std (ms)
CatBoost	30 min	0.39481	0.2655
CatBoost	3 h	0.5094	0.9006
MLP	6 h	0.2031	0.2157
Logistic Regression	12 h	0.0564	0.0364

V. DISCUSSION

This study evaluated temperature gradients as a biomarker for early shock assessment and advances the field in four main ways. First, it introduces a multi-horizon benchmarking framework that evaluates ten machine learning models across four forecasting intervals. Second, it develops horizon-specific models through Bayesian hyperparameter optimization to capture distinct temporal dynamics. Third, it implements a rigorous SIPA-based evaluation protocol using stratified group cross-validation and threshold optimization via the Youden Index. Finally, it shows that performance varies across horizons—no single architecture dominates—highlighting the necessity of tailored models for each clinical timeframe.

Within this framework, boosting algorithms excel at short-term forecasts, whereas MLPs and logistic regression perform better as the prediction window expands. This temporal dependence supports the development of four independently optimized models. The strongest performance occurred at the

30-minute horizon (AUC = 0.8371; 95% CI: 0.7794–0.8818), underscoring the value of high-resolution early prediction.

Each interval also requires careful tuning of hyperparameters and thresholds, as small AUC variations can affect critical clinical metrics. High specificity is essential to prevent alert fatigue [22], while high sensitivity is vital to avoid missing true shock events [23]. Thus, model configuration must align with the intended clinical objective, reinforcing the importance of horizon-aware modeling.

Despite these promising results, our temperature gradient-based assessment of cutaneous perfusion is subject to methodological constraints.The gradient approach may potentially mitigate uniform thermal noise through its focus on spatial differentials rather than absolute values, thereby offering advantages in handling systematic thermal variations. However, this methodology remains susceptible to non-uniform environmental artifacts that may compromise performance in heterogeneous thermal conditions. Factors such as uneven radiative sources or spatial inconsistencies in camera sensitivity can introduce systematic biases that are not fully addressed by our current methodology [24].

Regarding the study cohort, while recent evidence suggests constitutive skin pigmentation (phototype) does not significantly alter temperature measurements [25], potential residual effects in more diverse populations cannot be entirely ruled out. Furthermore, the use of the Shock Index Pediatric Age-Adjusted (SIPA) as an outcome, while age-appropriate, depends on the availability and accuracy of ancillary clinical data [26]. Although our analysis did not stratify by gender—justified by the absence of clear evidence for sex-based differences in thermal patterns [27]—this could obscure subtle, yet unidentified, variations.

VI. Conclusions

We used core-to-peripheral gradients to predict shock in pediatric patients by using a range of ML models. Various time horizons for prediction were evaluated, using one model for each time horizon.

Our results suggest that CatBoost was the best performer during the first two intervals—especially in the 30-minute window, which achieved the highest precision and lowest variability (AUC 0.8371 (95% CI: 0.7794 - 0.8818))—While the MLP demonstrated the best performance at the 6-hour horizon, Logistic Regression led at the 12-hour mark. Building on these findings, a combined strategy—employing ensemble techniques for immediate alerts, the MLP for medium-term predictions, and linear models for longer-term forecasts—could reduce clinicians' workload without sacrificing accuracy.

Together, these outcomes demonstrate a strong performance profile for the proposed approach. The model achieved higher AUC values compared to previous linear modeling strategies, indicating improved predictive capability across different horizons. Although these results exceed those reported in earlier studies, direct comparisons are limited by methodological differences, particularly in how relationships between observations were addressed. Even with these differences, the findings indicate that, when applied within an appropriate methodological framework, temperature gradients can provide valuable information for clinical decision support systems and strengthen predictive approaches in critical care.

Acknowledgments

The authors gratefully acknowledge the support received from Universidad Industrial de Santander (UIS) throughout the execution of this work. The institution's provision of academic, computational, and physical infrastructure was essential for carrying out the research. This project benefited from an academic environment that actively promotes scientific inquiry and the holistic education of its students.

Appendix
Hyperparameter Optimization Details

Tables VIII-X detail the hyperparameter search spaces defined for the models evaluated at each prediction horizon. A consistent search space was maintained across the first two horizons to ensure a controlled comparison of temporal generalizability. All optimizations were conducted via Bayesian optimization using the Optuna framework. Complementary software dependencies and their respective versions, essential for reproducibility, are documented in Table XI.

TABLE VIII
Hyperparameter Search Ranges for CatBoost Classifier

Hyperparameter	Search Range
depth	[3, 12]
learning_rate	[0.001, 0.3]
l2_leaf_reg	[1, 20]
subsample	[0.2, 1.0]
colsample_bylevel	[0.2, 1.0]
iterations	[500, 3000]
border_count	[32, 255]
random_strength	[0.0, 10.0]
bagging_temperature	[0.0, 1.0]
grow_policy	{SymmetricTree, Depthwise, Lossguide}
min_data_in_leaf	[1, 20]
one_hot_max_size	[2, 10]
leaf_estimation_iterations	[1, 10]
scale_pos_weight	[0.5, 3.0]

TABLE IX
Hyperparameter Search Ranges for the Multilayer Perceptron (MLP)

Hyperparameter	Search Range
activation	{relu, tanh, logistic}
solver	{adam, sgd}
learning_rate_init	[1e-4, 1e-1] (log-uniform)
max_iter	[200, 1000] (integer)
Parameters specific to Adam:	
batch_size	{16, 32, 64, auto}
beta_1	[0.8, 0.99]
beta_2	[0.9, 0.999]
epsilon	[1e-9, 1e-6] (log-uniform)
Parameters specific to SGD:	
momentum	[0.0, 0.9]
learning_rate	{constant, adaptive}
power_t	[0.1, 0.9]

TABLE X
HYPERPARAMETER SEARCH RANGES FOR LOGISTIC REGRESSION

Hyperparameter	Search Range
C	[0.001, 100]
penalty	{l1, l2, elasticnet}
l1_ratio	[0, 1] (only used if penalty = elasticnet)
max_iter	[500, 100000]
class_weight	{None, balanced}
tol	[1e-5, 1e-1]
fit_intercept	{True, False}

TABLE XI
LIBRARY VERSIONS USED IN THE EXPERIMENTS

Library	Version
pandas	2.2.2
numpy	2.0.2
scikit-learn	1.6.1
matplotlib	3.10.0
xgboost	3.1.1
catboost	1.2.8
lightgbm	4.6.0
optuna	4.5.0
scipy	1.16.3
statsmodels	0.14.5

REFERENCES

[1] S. L. Erica Koch, "Shock index in the emergency department: utility and limitations," *Open Access Emergency Medicine*, vol. 11, pp. 179–199, 2019. doi: 10.2147/OAEM.S178358.

[2] T. Berger, J. Green, T. Horeczko, Y. Hagar, N. Garg, A. Suarez, E. Panacek, and N. Shapiro, "Shock index and early recognition of sepsis in the emergency department: pilot study," *The Western Journal of Emergency Medicine*, vol. 14, pp. 168–174, Mar. 2013. doi: 10.5811/westjem.2012.8.11546.

[3] S. N. Acker, B. Bredbeck, D. A. Partrick, A. M. Kulungowski, C. C. Barnett, and D. D. Bensard, "Shock index, pediatric age-adjusted (sipa) is more accurate than age-adjusted hypotension for trauma team activation," *Surgery*, vol. 161, pp. 803–807, Mar. 2017. doi: 10.1016/j.surg.2016.08.050.

[4] R. Phillips, S. Acker, N. Shahi, G. Shirek, M. Meier, A. Goldsmith, J. Recicar, S. Moulton, and D. Bensard, "The shock index, pediatric age-adjusted (sipa) enhanced: Prehospital and emergency department sipa values forecast transfusion needs for blunt solid organ injured children," *Surgery*, vol. 168, no. 4, pp. 690–694, 2020. doi: 10.1016/j.surg.2020.04.061.

[5] A. Ochagavía, F. Baigorri, J. Mesquida, J. Ayuela, A. Ferrándiz, X. García, M. Monge, L. Mateu, C. Sabatier, F. Clau-Terré, R. Vicho, L. Zapata, J. Maynar, and A. Gil, "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," *Medicina Intensiva (English Edition)*, vol. 38, no. 3, pp. 154–169, 2014. doi: 10.1016/j.medine.2013.10.002.

[6] P. C. R. Garcia, C. T. Tonial, and J. P. Piva, "Septic shock in pediatrics: the state-of-the-art," *Jornal de Pediatria*, vol. 96 Suppl 1, pp. 87–98, Mar-Apr 2020. doi: 10.1016/j.jped.2019.10.007.

[7] M. Y. Rady, P. Nightingale, R. A. Little, and J. Edwards, "Shock index: a re-evaluation in acute circulatory failure," *Resuscitation*, vol. 23, no. 3, pp. 227–234, 1992. doi: 10.1016/0300-9572(92)90006-X.

[8] G. A. Ospina-Tascón, J.-L. Teboul, G. Hernandez, I. Alvarez, A. I. Sánchez-Ortiz, L. E. Calderón-Tapia, R. Manzano-Nunez, E. Quiñones, H. J. Madriñan-Navia, J. E. Ruiz, J. L. Aldana, and J. Bakker, "Diastolic shock index and clinical outcomes in patients with septic shock," *Annals of Intensive Care*, vol. 10, no. 1, p. 41, 2020. doi: 10.1186/s13613-020-00658-8.

[9] V. Vats, A. Nagori, P. Singh, R. Dutt, H. Bandhey, M. Wason, R. Lodha, and T. Sethi, "Early prediction of hemodynamic shock in pediatric intensive care units with deep learning on thermal videos," *Frontiers in Physiology*, vol. 13, 2022. doi: 10.3389/fphys.2022.862411.

[10] S. Bourcier, C. Pichereau, P.-Y. Boelle, S. Nemlaghi, V. Dubée, G. Lejour, J.-L. Baudel, A. Galbois, J.-R. Lavillegrand, N. Bigé, J. Tahiri, G. Leblanc, E. Maury, B. Guidet, and H. Ait-Oufella, "Toe-to-room temperature gradient correlates with tissue perfusion and predicts outcome in selected critically ill patients with severe infections," *Annals of Intensive Care*, vol. 6, no. 1, p. 63, 2016. doi: 10.1186/s13613-016-0164-2.

[11] H. Amson, C.-H. Vacheron, F. Thiolliere, V. Piriou, M. Magnin, and B. Allaouchiche, "Core-to-skin temperature gradient measured by thermography predicts day-8 mortality in septic shock: A prospective observational study," *Journal of Critical Care*, vol. 60, pp. 294–299, 2020. doi: 10.1016/j.jcrc.2020.08.022.

[12] A. Bridier, M. Shcherbakova, A. Kawaguchi, N. Poirier, C. Said, R. Noumeir, and P. Jouvet, "Hemodynamic assessment in children after cardiac surgery: A pilot study on the value of infrared thermography," *Frontiers in pediatrics*, vol. 11, p. 1083962, 04 2023. doi: 10.3389/fped.2023.1083962.

[13] A. Ortiz-Dosal, E. S. Kolosovas-Machuca, R. Rivera-Vega, J. Simón, and F. J. González, "Use of infrared thermography in children with shock: A case series," *SAGE Open Medical Case Reports*, vol. 2, p. 2050313X14561779, 2014. doi: 10.1177/2050313X14561779.

[14] M. Magnin, S. Junot, M. Cardinali, J. Y. Ayoub, C. Paquet, V. Louzier, J. M. B. Garin, and B. Allaouchiche, "Use of infrared thermography to detect early alterations of peripheral perfusion: evaluation in a porcine model," *Biomed Opt Express*, vol. 11, no. 5, pp. 2431–2446, 2020. doi: 10.1364/BOE.387481.

[15] R. Vardasca, C. Magalhaes, and J. Mendes, "Biomedical applications of infrared thermal imaging: Current state of machine learning classification," *Proceedings*, vol. 27, no. 1, 2019. doi: 10.3390/proceedings2019027046.

[16] M. Sudhi, D. K. Shetty, J. M. Balakrishnan, D. R. Prabhu, R. Kamath, and S. Girisha, "Thermal imaging applications in shock detection: Technological advancements and clinical implications," *Engineered Science*, vol. 33, p. 1367, 2025. doi: 10.30919/es1367.

[17] A. Nagori, L. Dhingra, A. Bhatnagar, R. Lodha, and T. Sethi, "Predicting hemodynamic shock from thermal images using machine learning," *Scientific Reports*, vol. 9, 01 2019. doi: 10.1038/s41598-018-36586-8.

[18] M. D. Ruopp, N. J. Perkins, B. W. Whitcomb, and E. F. Schisterman, "Youden index and optimal cut-point estimated from observations affected by a lower limit of detection," *Biometrical Journal. Biometrische Zeitschrift*, vol. 50, no. 3, pp. 419–430, 2008. doi: 10.1002/bimj.200710415.

[19] M. Hassanzad and K. Hajian-Tilaki, "Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in roc analysis: an update review," *BMC Medical Research Methodology*, vol. 24, no. 1, p. 84, 2024. doi: 10.1186/s12874-024-02198-2.

[20] T. Sethi, "Data for manuscript on automated shock detection in the icu using thermal imaging," 2018. doi: 10.17605/OSF.IO/VP86J.

[21] R. Fluss, D. Faraggi, and B. Reiser, "Estimation of the youden index and its associated cutoff point," *Biometrical Journal*, vol. 47, pp. 458–472, Aug. 2005. doi: 10.1002/bimj.200410135.

[22] E. C. Alberto, E. McKenna, M. J. Amberson, J. Tashiro, K. Donnelly, A. A. Thenappan, P. E. Tempel, A. S. Ranganna, S. Keller, I. Marsic, A. Sarcevic, K. J. O'Connell, and R. S. Burd, "Metrics of shock in pediatric trauma patients: A systematic search and review," *Injury*, vol. 52, pp. 3166–3172, Oct. 2021. doi:10.1016/j.injury.2021.06.014.

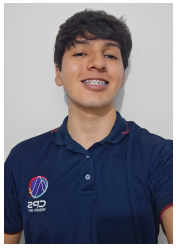
[23] M. Y. Alsagaff, R. B. Kurniawan, D. Purwati, A. U. D. U. Haq, P. Saputra, C. Milla, L. F. Kusumawardhani, C. P. Budianto, H. Susilo, and Y. Oktaviono, "Shock index in the emergency department as a predictor for mortality in covid-19 patients: A systematic review and meta-analysis," *Heliyon*, vol. 9, 2023. doi:10.1016/j.heliyon.2023.e18553.

[24] A. Nowakowski and M. Kaczmarek, "Artificial intelligence in ir thermal imaging and sensing for medical applications," *Sensors (Basel, Switzerland)*, vol. 25, 2025. doi:10.3390/s25030891.

[25] M. Charlton, S. A. Stanley, Z. Whitman, V. Wenn, T. J. Coats, M. Sims, and J. P. Thompson, "The effect of constitutive pigmentation on the measured emissivity of human skin," *PLoS ONE*, vol. 15, 2020. doi:10.1371/journal.pone.0241843.

[26] D. Marlor, J. Flint, J. R. Noel-MacDonnell, N. Cruz-Centeno, S. Stewart, M. Elman, and D. Juang, "Establishing pediatric age-adjusted shock index cut points in trauma patients younger than 1 year," *J Trauma Acute Care Surg*, vol. 97, pp. 386–392, Sept. 2024. doi:10.1097/TA.0000000000004251.

[27] I. Fernández-Cuevas, J. C. Bouzas Marins, J. Arnáiz Lastras, P. M. Gómez Carmona, S. Piñonosa Cano, M. Ángel García-Concepción, and M. Sillero-Quintana, "Classification of factors influencing the use of infrared thermography in humans: A review," *Infrared Physics & Technology*, vol. 71, pp. 28–55, 2015. doi:10.1016/j.infrared.2015.02.007.



Juan D. Espinoza received his B.Sc. degree in Electronic Engineering from the Universidad Industrial de Santander (UIS), Colombia. He is currently awaiting the award of his M.Sc. degree in Telecommunications Engineering from UIS. He is a member of the Connectivity and Signal Processing (CPS) Research Group at UIS. His research interests include the application of artificial intelligence to pediatric intensive care unit environments, with a particular focus on interpretable machine learning models for clinical decision support and non-invasive technologies. During his master's program, he was awarded a full scholarship covering tuition and living expenses.

Carlos A. Fajardo holds a Ph.D. in Engineering with a focus on High-Performance Computing, an M.Sc. in Electronic Engineering with a specialization in Advanced Digital Design, and a postgraduate certificate in University Teaching, all from UIS. He completed a postdoctoral fellowship at the Center for Brain-Inspired Computing (C-BRIC) at Purdue University, where he specialized in edge AI through hardware–software co-design, and also served as a visiting researcher at Purdue's Integration Lab. His research focuses on artificial intelligence applied to medical problems, with additional expertise in advanced digital systems and hardware–software co-design.