Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and internal validation of machine-learning models for predicting intraoperative hypothermia in pediatric surgical patients: a retrospective cohort study

Shifa Ma

894278l6@qq.com

The First People's Hospital of Nanning

Jicai Deng

The First People's Hospital of Nanning

Bin Liang

The First People's Hospital of Nanning

Chunxiu Lu

The First People's Hospital of Nanning

Research Article

Keywords: Intraoperative hypothermia, Pediatrics, Machine learning, Random forest, Explainable artificial intelligence

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11310864/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Development and internal validation of machine-learning models for predicting intraoperative hypothermia in pediatric surgical patients: a retrospective cohort study

Shifa Ma, Jicai Deng, Bin Liang, Chunxiu Lu

Operating Room, Department of Anesthesiology, The First People's Hospital of Nanning, Nanning, Guangxi 530022, China

Corresponding author

Shifa Ma, MSc, Associate Chief Nurse

Operating Room, Department of Anesthesiology, The First People's Hospital of Nanning, Guangxi 530022, China

E-mail: 89427816@qq.com

Funding

This work was supported by the Self-Funded Research Project of the Health Commission of Guangxi Zhuang Autonomous Region, China (No. Z-A20261049)

Abstract

Background Intraoperative hypothermia (core temperature < 36.0 °C) is a common and preventable complication of pediatric surgery, associated with coagulopathy, surgical-site infection, and prolonged hospitalization. Existing pediatric risk-assessment tools rely mainly on logistic regression and cannot capture nonlinear predictor–outcome relationships. We aimed to develop, compare, and internally validate machine-learning models for pediatric intraoperative hypothermia.

Methods This retrospective cohort study included 3,637 children aged 0–14 years who underwent surgery at a tertiary hospital in Nanning, China, between January 2020 and December 2025, with intraoperative hypothermia (core temperature < 36.0 °C) as the outcome. Two of the six final predictors—surgical duration and intraoperative blood loss—are recorded during surgery, so the models were specified for intraoperative risk stratification rather than preoperative screening. Fourteen candidate predictors were

screened by univariate analysis and least absolute shrinkage and selection operator regression. Five models were developed in a stratified 7:3 training–validation split: logistic regression, decision tree, random forest, AdaBoost, and XGBoost. Discrimination, calibration, clinical utility, and bootstrap internal validation were assessed. SHapley Additive exPlanations (SHAP) values interpreted the optimal model.

Results **Hypothermia occurred in 177 of 3,637 children (4.87%). Six predictors were retained: surgical duration, intraoperative blood loss, American Society of Anesthesiologists (ASA) classification, surgical grade, surgical site, and body mass index (BMI). In the validation set, the random forest model showed the best combined performance (area under the receiver operating characteristic curve [AUC], 0.882; Hosmer\u2013Lemeshow P = .58), whereas AdaBoost had the highest AUC (0.885) but poor calibration (P < .001). Bootstrap validation showed low optimism for the random forest model (0.048; optimism-corrected AUC 0.885; 95% CI 0.824\u20130.946), with stable performance across sex and age subgroups. Longer surgical duration, greater blood loss, higher ASA classification, and higher surgical grade were independent risk factors, and higher BMI protective (all P < .001). SHAP ranked surgical grade most influential.**

Conclusions **The random forest model offered the best discrimination–calibration trade-off among the algorithms. With SHAP-based interpretability, the models can stratify hypothermia risk during surgery to guide individualized perioperative warming. Because two predictors are recorded intraoperatively, they inform risk stratification at the point of surgery rather than preoperative risk prediction.**

Keywords: **Intraoperative hypothermia; Pediatrics; Machine learning; Random forest; Explainable artificial intelligence**

Background

Intraoperative hypothermia, commonly defined as a core body temperature below 36.0 °C at any point during surgery, is one of the most frequent complications of pediatric anesthesia [1]. Children are particularly vulnerable to perioperative heat loss because the thermoregulatory center is immature, the body surface area-to-mass ratio is high, subcutaneous fat is thin, and brown-fat thermogenic reserves are limited; as a result, children lose heat more readily than adults under identical operating-room conditions [2, 3]. Reported incidence varies widely with case mix, ranging from approximately 4–18% in general elective pediatric surgical populations to more than 50% in major neonatal and transplant surgery [4, 5].

Intraoperative hypothermia is not a benign physiological fluctuation. It impairs platelet function and the coagulation cascade, promotes postoperative shivering and surgical-site infection, and is associated with pulmonary complications, prolonged hospitalization, and increased perioperative costs [6, 7]. Early identification of children at high risk, followed by targeted warming interventions, is therefore a core component of perioperative anesthesia management and quality nursing care.

Current pediatric risk-assessment tools are based predominantly on logistic regression nomograms and risk-rating scales [4, 5, 8]. Although such tools are transparent and easy to use, they model only linear predictor–outcome relationships and cannot capture the complex interactions and nonlinear effects that characterize heterogeneous clinical data [8, 9]. Machine-learning approaches, particularly ensemble algorithms such as random forest and XGBoost, can exploit high-dimensional clinical data and often outperform conventional statistical models in prediction tasks [10, 11]. However, machine-learning studies of intraoperative hypothermia in children remain scarce, are mostly restricted to a single surgical procedure, and rarely include interpretability analyses, which limits clinical translation [12].

We therefore used a large single-center cohort of pediatric surgical patients to (1) select core predictors by univariate screening and LASSO regression; (2) construct and compare five machine-learning prediction models spanning linear, tree-based, bagging, and boosting paradigms; (3) evaluate discrimination, calibration, and clinical utility, with bootstrap internal validation; and (4) elucidate the decision logic of the optimal model using SHAP values. The study is reported according to the TRIPOD+AI statement for clinical prediction models [13].

Methods

Study design and setting

This retrospective cohort study was conducted at The First People's Hospital of Nanning, a tertiary grade-A hospital in Guangxi, China. Children who underwent surgery between January 2020 and December 2025 were eligible. The study was approved by the hospital ethics committee (No. KY2026-Yijian-023) and conducted in accordance with the Declaration of Helsinki; the requirement for individual informed consent was waived because of the retrospective, de-identified nature of the data.

Participants

Inclusion criteria were: (1) age 0–14 years; (2) elective or emergency surgery; and (3) complete perioperative clinical and anesthesia records. Exclusion criteria were: (1) preoperative fever (≥ 37.5 °C) or preexisting hypothermia; (2) severe infection, thyroid disease, or other conditions markedly impairing thermoregulation; and (3) absence of the intraoperative temperature record required to define the outcome.

Outcome definition

The outcome was intraoperative hypothermia, defined as a core temperature < 36.0 °C recorded at any time during surgery (0 = no, 1 = yes), consistent with the definition used in previous pediatric studies [1, 3]. Core temperature was measured continuously throughout surgery with a temperature probe connected to a multiparameter patient monitor (BeneView T5, Mindray Bio-Medical Electronics, Shenzhen, China), following routine institutional protocol. Non-invasive blood pressure was recorded every 10 min throughout surgery.

Candidate predictors

Fourteen perioperative variables were considered as candidate predictors based on literature review and expert consultation: demographic variables (sex, age, BMI); preoperative variables (surgical department, surgical category, surgical grade, ASA classification, anesthesia method, surgical site, anesthesiologist seniority, systolic and diastolic blood pressure); and intraoperative variables (surgical duration, intraoperative blood loss).

Missing data

Before analysis, missingness was examined for all variables. Systolic and diastolic blood pressure were missing in 2.12% and 2.28% of records, respectively; all other variables had < 1.5% missing, and the outcome had no missing values. Little's test indicated data missing at random. Missing values were handled by multiple imputation by chained equations (MICE) [14] with five imputations, including all candidate predictors and the outcome in the imputation models; inferences were pooled by Rubin's rules, and a complete-case sensitivity analysis yielded consistent results. Imputation was performed after data splitting: imputation models were fitted on the training set only and applied to the validation set to prevent information leakage between partitions. All six predictors retained in the final models were complete in every record; the imputed blood-pressure values were used only for the descriptive comparisons in Table 1.

Sample size

A total of 177 hypothermia events occurred. With six predictors entering the final models, the events-per-variable (EPV) ratio was 29.5, comfortably exceeding the conventional minimum of 10 events per predictor parameter used in prediction model development; recent methodological work nevertheless recommends that sample size be justified by formal shrinkage- and fit-based criteria rather than by this rule of thumb alone [15].

Data splitting and variable selection

The cohort was randomly split into a training set and a validation set at a ratio of 7:3, stratified by outcome to preserve the hypothermia rate across partitions. In the training set, variables with $P < .05$ in univariate between-group comparisons entered LASSO logistic regression; the penalty parameter λ was selected by 10-fold cross-validation [16], and variables with non-zero coefficients were retained for model development.

Model development

Five algorithms were trained on the selected predictors: (1) logistic regression, fitted by maximum likelihood (maximum 5,000 iterations), as the conventional benchmark; (2) decision tree (maximum depth = 4; minimum samples per leaf = 10); (3) random forest (500 trees; maximum depth = 6; minimum samples per leaf = 10), a bagging ensemble; (4) AdaBoost (100 estimators), an adaptive boosting ensemble; and (5) XGBoost (80 boosting rounds; maximum depth = 3; learning rate = 0.1; minimum child weight = 5; L1

regularization = 0.5; L2 regularization = 1.0; subsample = 0.8; column subsample = 0.8) [17]. All predictors entered the models as numeric variables; ordinal categorical predictors (ASA classification, surgical grade, and surgical site) were coded as ordered integers and modeled as single continuous terms, consistent with the multivariable logistic model (Table 2).

Prediction time point and intended use

Two of the six final predictors—surgical duration and intraoperative blood loss—are determined during surgery and are not available before skin incision. The models were therefore specified for risk stratification at the point of surgery, in order to inform intraoperative warming decisions and to support targeted active warming, rather than for preoperative risk assessment. Risk scores should not be generated from these models before incision; a preoperative model would require refitting with preoperative variables alone. Both time-dependent quantities are themselves associated with the operations in which hypothermia is most likely to occur, so their inclusion also reflects the clinical reality that risk becomes apparent during the operation.

Model evaluation

Discrimination was assessed by the area under the receiver operating characteristic curve (AUC) in both training and validation sets, supplemented by accuracy, precision, recall, and F1 score, computed at the default probability threshold of 0.5. Calibration was assessed graphically by calibration curves and quantified by the Hosmer–Lemeshow (HL) test, where $P > .05$ indicates acceptable calibration; because this test has limited power and depends on arbitrary grouping, it was treated as a supplementary check alongside the calibration curves rather than as the sole measure of agreement [18]. Clinical utility was evaluated by decision curve analysis (DCA) [19], which estimates the net benefit of guiding interventions on model-predicted probabilities: net benefit = true-positive rate − false-positive rate × [threshold probability / (1 − threshold probability)], compared against the “treat all” and “treat none” strategies.

Internal validation

Optimism and internal validity were quantified by bootstrap resampling [20]: 200 resamples of the training set were used to estimate each model’s optimism (apparent AUC − bootstrap AUC) and optimism-corrected AUC, and 1,000 resamples of the validation set were used to

derive 95% CIs for validation AUCs. Performance of the optimal model was further examined in prespecified subgroups defined by sex and age band (infants and toddlers 0–3 years; preschool 4–6 years; school-age 7–14 years).

Interpretability analysis

For the best-performing model, SHAP values were computed with TreeExplainer, the reference implementation for tree ensembles; SHAP has become the most frequently used explanation method in tabular electronic health record research [21]. A global feature-importance bar chart ranked predictors by mean absolute SHAP value, and a beeswarm plot depicted the direction and magnitude of each predictor's effect on individual predictions; positive SHAP values indicate increased hypothermia risk and negative values decreased risk.

Statistical analysis

All analyses were performed in Python 3.9 (scikit-learn, statsmodels, XGBoost, and SHAP libraries). Normally distributed continuous variables are presented as mean ± SD and compared by independent-samples t tests; skewed variables as median (P25, P75) and compared by Mann–Whitney U tests; categorical variables as n (%) and compared by χ^2 tests or Fisher's exact test as appropriate. Multivariable logistic regression yielded adjusted coefficients, odds ratios (ORs), and 95% CIs. All tests were two-sided, with $P < .05$ considered statistically significant.

Use of large language models AI-based tools were used solely for language editing and stylistic refinement of the manuscript; they were not used to design the study, analyse data, generate results, or draw conclusions. No large language model is listed as an author. The authors take full responsibility for the content of the article.

Patient and public involvement

No patients or members of the public were involved in the design, conduct, or reporting of this research.

Results

Cohort characteristics

Of 4,124 children who underwent surgery during the study period, 487 were excluded because they were older than 14 years, leaving 3,637 children in the analysis (2,784 boys [76.55%] and 853 girls [23.45%]; age 0–14 years, mean 7.02 ± 3.88 years) (Fig. 1). Intraoperative hypothermia occurred in 177 children (4.87%). The cohort was split into a training set of 2,545 children (124 events) and a validation set of 1,092 children (53 events).

Univariate analysis

Surgical duration, BMI, systolic and diastolic blood pressure, intraoperative blood loss, sex, surgical department, surgical grade, ASA classification, and surgical site differed significantly between the hypothermia and non-hypothermia groups (all $P < .05$), whereas age, surgical category, anesthesia method, and anesthesiologist seniority did not (Table 1).

Predictor selection by LASSO regression

Eight variables significant in training-set univariable screening (surgical duration, BMI, intraoperative blood loss, sex, surgical department, surgical grade, ASA classification, and surgical site) were entered into LASSO regression with 10-fold cross-validation (Fig. 2). Six variables with non-zero coefficients were retained: surgical duration, intraoperative blood loss, ASA classification, surgical grade, surgical site, and BMI; the coefficients of sex and surgical department were shrunk to zero and excluded. Systolic and diastolic blood pressure, although significant in the full-cohort comparisons of Table 1, did not reach $P < .05$ in the training-set screening and therefore did not enter LASSO regression.

Multivariable logistic regression

In the multivariable logistic model (Table 2), longer surgical duration (OR 1.005 per minute, 95% CI 1.002–1.008), greater intraoperative blood loss (OR 1.007 per mL, 95% CI 1.004–1.010), higher ASA classification (OR 1.689 per class, 95% CI 1.299–2.196), and higher surgical grade (OR 3.211 per grade, 95% CI 2.241–4.602) were independent risk factors, whereas higher BMI was independently protective (OR 0.860 per kg/m², 95% CI 0.790–0.936) (all $P < .001$). The independent effect of surgical site was not statistically significant ($P = .64$).

Model performance

Training-set and validation-set ROC curves are shown in Figs. 3 and 4, and performance metrics are summarized in Table 3. In the validation set, AdaBoost achieved the numerically highest AUC (0.885) but showed severe miscalibration (HL $P < .001$); the random forest model achieved an AUC of 0.882 with good calibration (HL $P = .58$), representing the best overall performance; XGBoost followed (AUC 0.878; HL $P = .63$).

Bootstrap internal validation and subgroup analysis

Bootstrap internal validation confirmed low optimism for the random forest model (optimism 0.048; optimism-corrected AUC 0.885, closely matching the validation AUC of 0.882; validation AUC 95% CI 0.824–0.946), indicating limited overfitting and stable generalization (Table 4). Optimism-corrected AUCs for the other models were 0.818 (logistic regression), 0.799 (decision tree), 0.830 (AdaBoost), and 0.856 (XGBoost). In subgroup analyses, the random forest model maintained stable discrimination in boys (AUC 0.895) and girls (0.886), and across age bands—infants and toddlers (0.871), preschool children (0.858), and school-age children (0.913); subgroup event counts were small (12–38 events), so these estimates warrant confirmation in larger samples.

Calibration

Calibration curves for the validation set are shown in Fig. 5. The random forest, XGBoost, and decision tree curves tracked the diagonal closely, indicating good agreement between predicted probabilities and observed event rates; logistic regression deviated in the mid-to-high probability range; AdaBoost showed the largest calibration deviation, consistent with its HL test result.

SHAP interpretability analysis

SHAP analysis of the random forest model (Figs. 6 and 7) ranked global feature importance as: surgical grade > intraoperative blood loss > ASA classification > BMI > surgical duration > surgical site. The beeswarm plot showed that higher surgical grade, greater blood loss, higher ASA classification, and longer surgical duration shifted SHAP values positively (increasing predicted risk), whereas higher BMI shifted SHAP values negatively; surgical site contributed little overall. Notably, the effect of surgical grade exhibited a marked jump at grade IV rather than a linear increment.

Decision curve analysis

DCA on the validation set (Fig. 8) showed that the random forest, XGBoost, and decision tree models provided net benefit above both the "treat all" and "treat none" strategies across threshold probabilities of 0.02–0.60 (logistic regression, 0.01–0.53). Within the clinically relevant 0.05–0.20 threshold range, the random forest model's net benefit was 0.027–0.034, comparable to XGBoost—corresponding to approximately three additional correctly identified hypothermia cases per 100 operated children without additional false positives. AdaBoost yielded positive net benefit only within a narrow 0.41–0.54 range and was inferior to "treat none" elsewhere, consistent with its poor calibration and arguing against direct clinical use of its probabilities.

Discussion

Principal findings

In this cohort of 3,637 children, six routinely documented predictors sufficed to build models with good predictive performance for intraoperative hypothermia. The random forest model offered the best combination of discrimination (validation AUC 0.882) and calibration (HL P = .58), with minimal bootstrap optimism (0.048) and stable subgroup performance. SHAP analysis aligned fully with the direction of multivariable logistic regression estimates and revealed a nonlinear jump in risk at surgical grade IV, illustrating how ensemble models capture effects that linear models cannot.

Incidence and its context

The 4.87% incidence of intraoperative hypothermia was lower than the 17.31% reported for general pediatric surgical populations [5] and the 14.89% reported for laparoscopic pediatric surgery [4], and far below the 67.3% observed in neonates undergoing surgery for necrotizing enterocolitis [1] and 55.09% in infant living-donor liver transplantation [2]. These differences likely reflect case mix — our cohort was dominated by lower-acuity elective surgery with relatively few neonatal or critically ill patients — as well as adherence to routine baseline warming measures at our center. The consequences, however, are consistent across settings: intraoperative hypothermia has been linked to coagulopathy and prolonged hospitalization

[1], postoperative pulmonary complications [6], and increased morbidity and cost [7], reinforcing the value of perioperative risk stratification.

Risk factors and mechanisms

Surgical grade was the strongest predictor. Higher-grade operations involve larger exposures and prolonged open body cavities, amplifying heat loss by radiation, conduction, and evaporation, while greater surgical stress suppresses hypothalamic thermogenic compensation; this accords with findings in infant liver transplantation [2] and pediatric burn surgery [22]. Intraoperative blood loss ranked second: hemorrhage dissipates body heat directly, and infusion of room-temperature fluids and blood products further dilutes core temperature, as reported in pediatric liver transplantation [6] and neonatal necrotizing enterocolitis [1]. Higher ASA classification reflects diminished physiological reserve for thermoregulation and is a core component of existing pediatric hypothermia risk scores [7, 9] and multicenter analyses [23]. Lower BMI was independently protective: thin children have less insulating adipose tissue and relatively larger surface area, mirroring reports in very-low-birth-weight neonates [24] and infants [3]. Longer surgical duration compounds cumulative heat loss [3, 7, 8]. These effect directions are physiologically coherent and clinically actionable—for example, anticipating fluid warming and active insulation when major surgery, large blood loss, or long duration is expected in an underweight child.

Comparison with previous models

Existing pediatric tools are predominantly logistic-regression nomograms: AUC 0.698 in a general pediatric cohort [5] and 0.895–0.907 in laparoscopic surgery [4]. Our logistic benchmark reached 0.847 but with suboptimal calibration, and the ensemble models outperformed it, consistent with the general superiority of machine learning in heterogeneous clinical data [10, 11]. Our random forest (AUC 0.882) compares favorably with machine-learning models in neonates and young children (AUC 0.82–0.85 [25]) and neonates (AUC 0.81 [26]). Importantly, our head-to-head comparison illustrates why calibration must accompany discrimination: AdaBoost had the highest validation AUC (0.885) yet catastrophic miscalibration (HL $P < .001$), rendering its predicted probabilities unusable for clinical decisions. The robustness of random forest's bagging architecture to noise and class

imbalance (events < 5%) plausibly underlies its superior calibration, in line with findings in pediatric thoracoscopic surgery [12].

Interpretability and clinical implications

The "black box" nature of machine-learning models is a major barrier to clinical adoption. By coupling the random forest model with SHAP analysis, we show that its learned relationships reproduce the effects identified by conventional logistic regression—evidence that the model captured genuine clinical regularities rather than noise—while also exposing nonlinear behavior (the grade-IV risk jump) invisible to linear models. In practice, the models are intended to run inside the hospital information system, automatically extracting BMI, ASA classification, and scheduled surgical attributes and, once the operation has started, continuously updating an individual hypothermia probability that is displayed to anesthesiologists and operating-room nurses so that warming measures can be intensified for patients at high risk. Warming evidence supports this workflow: forced-air warming blankets reduce perioperative hypothermia in children (meta-analysis [27]); evidence summaries recommend stratified warming and continuous temperature monitoring [28]; and closed-loop temperature management reduces hypothermia in pediatric laparoscopic surgery [29]. The DCA quantifies the value of this triage: at the 0.05–0.20 threshold range, risk stratification by the random forest model yields a net benefit of 0.027–0.034—roughly three additional true positives per 100 children at no added false-positive burden—allowing warming resources to be concentrated on high-risk children while avoiding overtreatment of low-risk children. Model outputs are intended to support, not replace, clinical judgment.

Strengths and limitations

Strengths include the large sample for this outcome, a systematic comparison of five algorithm families, a comprehensive evaluation spanning discrimination, calibration, clinical utility, and internal validity, and SHAP-based interpretability, reported transparently according to TRIPOD+AI. Several limitations should be acknowledged. First, this was a single-center retrospective study; although bootstrap internal validation showed low optimism, external validation in independent and multicenter cohorts is required before broader deployment. Second, potentially important predictors — operating-room ambient temperature, infused fluid temperature, and active warming measures — were not captured

and could improve model accuracy. Third, the outcome was relatively rare (4.87%); despite an EPV of 29.5 and stable subgroup estimates, larger cohorts would sharpen precision, particularly in the youngest age bands. Fourth, predicted probabilities reflect the case mix and warming practices of the source center and may require recalibration elsewhere. Fifth, because two of the six predictors are recorded during surgery, the models inform intraoperative risk stratification; a preoperative model would require refitting with preoperative predictors alone.

Future directions

Priorities are multicenter external validation, incorporation of operating-room environmental and warming-intervention variables, and prospective implementation studies evaluating whether model-guided warming reduces intraoperative hypothermia and downstream complications.

Conclusions

Among five machine-learning models built from six routinely available predictors, the random forest model demonstrated the best combination of discrimination and calibration for predicting intraoperative hypothermia in children, with low bootstrap-estimated optimism and interpretable SHAP feature effects. Surgical grade, intraoperative blood loss, ASA classification, BMI, and surgical duration were the core determinants of risk. The models can stratify hypothermia risk during surgery to guide individualized perioperative temperature management. Applied before incision, they would require refitting with preoperative predictors alone; clinical implementation requires external validation.

List of abbreviations

ADA: AdaBoost; ASA: American Society of Anesthesiologists; AUC: area under the receiver operating characteristic curve; BMI: body mass index; CI: confidence interval; DCA: decision curve analysis; DT: decision tree; EPV: events per variable; HL: Hosmer–Lemeshow; LASSO: least absolute shrinkage and selection operator; LR: logistic regression; MAR: missing at random; MICE: multiple imputation by chained equations; OR: odds ratio; RF: random forest; ROC: receiver operating characteristic; SHAP: SHapley Additive

exPlanations; TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence

Declarations

Ethics approval and consent to participate **This study was approved by the Ethics Committee of The First People's Hospital of Nanning (approval No. KY2026-Yijian-023) and conducted in accordance with the Declaration of Helsinki. The requirement for informed consent to participate was waived by the Ethics Committee of The First People's Hospital of Nanning owing to the retrospective, de-identified design.**

Consent for publication. Not applicable.

Availability of data and materials **All de-identified participant data and the analysis code supporting the conclusions of this article are available from the corresponding author on reasonable request. The datasets analysed during the current study are not publicly available because individual privacy could be compromised, but are available from the corresponding author on reasonable request.**

Competing interests **The authors declare that they have no competing interests.**

Funding **This work was supported by the Self-Funded Research Project of the Health Commission of Guangxi Zhuang Autonomous Region, China (No. Z-A20261049). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.**

Authors' contributions **SM curated the data and drafted the manuscript. JD verified the data and revised the manuscript. BL performed the statistical analysis. CL verified the data and revised the manuscript. All authors read and approved the final manuscript.**

Acknowledgements **Not applicable.**

Authors' information **Not applicable.**

References

1. Ran Q, Wei H, Shi Y, Wang ZL. Effect of intraoperative hypothermia on short-term prognosis in neonatal necrotizing enterocolitis [in Chinese]. Chin J Contemp Pediatr. 2026;28(5):556–62. doi:10.7499/j.issn.1008-8830.2510016.

2. Huang YL, Yang Y, Chen GL. Trend of the temperature variation for hypothermia during pediatric living-donor liver transplantation and nursing implications [in Chinese]. Chin J Nurs. 2025;60(21):2572–8. doi:10.3761/j.issn.0254-1769.2025.21.002.

3. Zhao J, Le Z, Chu L, Gao Y, Zhang M, Fan J, et al. Risk factors and outcomes of intraoperative hypothermia in neonatal and infant patients undergoing general anesthesia and surgery. Front Pediatr. 2023;11:1113627. doi:10.3389/fped.2023.1113627.

4. Lan HL, Zhang SY, Zhang J, Cheng MJ, Jin CY. Construction and validation of a nomogram for predicting hypothermia during laparoscopic surgery in children [in Chinese]. J Laparosc Surg. 2026;31(2):128–35. doi:10.13499/j.cnki.fqjwkzz.2026.02.128.

5. Ji B, Xu WH, Chen H, Yang LC, Luo WY, Shen NP. Construction of a nomogram prediction model for intraoperative hypothermia in children [in Chinese]. J Nurs Sci. 2022;37(1):39–41. doi:10.3870/j.issn.1001-4152.2022.01.039.

6. Zhang GC, Liu YX, Yu HL, Yu WL. Intraoperative hypothermia and early postoperative pulmonary complications in pediatric living-related liver transplantation [in Chinese]. Chin J Anesthesiol. 2025;45(3):281–5. doi:10.3760/cma.j.cn131073-20240711-00304.

7. Saezhang W, Oofuvong M, Kovitwanawong N, Yongyukantorn K. Perioperative outcomes associated with intraoperative hypothermia in pediatric patients with preserved functional capacity undergoing anesthesia: a multivariate analysis. J Clin Med. 2025;14(20):7320. doi:10.3390/jcm14207320.

8. Lü Y, Wu HY, Li S, Zhao SY, Chen XR, Liu YL. Development and validation of a risk assessment scale for intraoperative hypothermia in infants and young children [in Chinese]. J Nurs Sci. 2025;40(18):38–42. doi:10.3870/j.issn.1001-4152.2025.18.038.

9. Saezhang W, Oofuvong M, Kovitwanawong N, Yongyukantorn K, Pichayanukul N, Lakateb C. Development of a risk predictive score for intraoperative hypothermia in pediatric patients: a retrospective cohort study. PLoS One. 2025;20(10):e0335796. doi:10.1371/journal.pone.0335796.

10. Wang Q, Ding X, Luo Z, Shi J, Zheng L. Advances of artificial intelligence in nursing quality management [in Chinese]. Chin J Nurs. 2024;59(23):2933–9. doi:10.3761/j.issn.0254-1769.2024.23.019.

11. Zhao R, Fan WQ, Liu XX, Ge LN. Establishment and comparison of 3 compassion fatigue risk prediction models for obstetrics and gynecology nurses [in Chinese]. Chin J Nurs. 2025;60(3):347–54. doi:10.3761/j.issn.0254-1769.2025.03.013.

12. Liu X, Lin H, Li Y, Dong C, Wang T, Wang X, et al. Construction and validation of a risk prediction model for intraoperative hypothermia in children undergoing thoracoscopic lobectomy. Sci Prog. 2025;108(4):368504251386310. doi:10.1177/00368504251386310.

13. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.

14. Legendre B, Cerasuolo D, Dejardin O, Boyer A. How to deal with missing data? Multiple imputation by chained equations: recommendations and explanations for clinical practice. Nephrol Ther. 2023;19(3):171–4. doi:10.1684/ndt.2023.24.

15. Martin GP, Riley RD, Collins GS, Sperrin M. Developing clinical prediction models when adhering to minimum sample size recommendations: the importance of quantifying bootstrap variability in tuning parameters and predictive performance. Stat Methods Med Res. 2021;30(12):2545–61. doi:10.1177/09622802211046388.

16. Friedrich S, Groll A, Ickstadt K, Kneib T, Pauly M, Rahnenführer J, Friede T. Regularization approaches in clinical biostatistics: a review of methods and their applications. Stat Methods Med Res. 2023;32(2):425–40. doi:10.1177/09622802221133557.

17. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? Adv Neural Inf Process Syst. 2022;35:507–20.

18. Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Med Res Methodol. 2022;22:316. doi:10.1186/s12874-022-01801-8.

19. Vickers AJ, Van Claster B, Wynants L, Steyerberg EW. Decision curve analysis: confidence intervals and hypothesis testing for net benefit. Diagn Progn Res. 2023;7(1):11. doi:10.1186/s41512-023-00148-y.

20. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. BMJ. 2024;386:e078276. doi:10.1136/bmj-2023-078276.

21. Caterson J, Lewin A, Williamson E. The application of explainable artificial intelligence (XAI) in electronic health record research: a scoping review. Digit Health. 2024;10:20552076241272657. doi:10.1177/20552076241272657.

22. Sahutoglu C, Ozok G, Bor C. Risk factors for intraoperative hypothermia in pediatric burn surgery: a single-center university hospital experience. Ther Hypothermia Temp Manag. 2026. doi:10.1177/21537658261474905.

23. Mekonnen MT, Mustofa SY, Berhe YW, Arefayne NR, Tegegne BA. Intraoperative pediatric hypothermia and associated factors during general anesthesia at comprehensive specialized hospitals, Northwest Ethiopia: a multicenter follow-up study. PLoS One. 2025;20(5):e0320950. doi:10.1371/journal.pone.0320950.

24. Cui Y, Zhang D, Gong T, Huang Q, Zhang C. Incidence and risk factors for intraoperative hypothermia in very low birth weight (≤ 1500 g) neonates undergoing general anesthesia and digestive surgery. Paediatr Anaesth. 2026;36(1):100–3. doi:10.1111/pan.70017.

25. Ruan J, Dai K, Wu Y, Zhang Y, Mai J, Qin L, et al. Development and validation of a predictive model for intraoperative hypothermia in neonates and young children undergoing general anesthesia. J Perianesth Nurs. 2026;41(1):75–84. doi:10.1016/j.jopan.2025.04.013.

26. Li B, Zhen H, Wu W, Qiao X. Formulation and implementation of a risk prediction model for intraoperative hypothermia in neonates. Asian J Surg. 2024;47(9):4243–5. doi:10.1016/j.asjsur.2024.05.090.

27. Ye PY, Chen XY, Wu XC, Nian JX. Effectiveness of inflatable warming blankets for preventing perioperative hypothermia in children: a meta-analysis. World J Surg. 2025;49(10):2807–15. doi:10.1002/wjs.70038.

28. Chen GL, Huang YL, Yang Y. Evidence summary and audit indicators for preventing intraoperative hypothermia in infant liver transplantation [in Chinese]. Milit Nurs. 2026;43(3):26–30. doi:10.3969/j.issn.2097-1826.2026.03.007.

29. Ye GM, Chen YJ, Wang J, Ma WL, Yang Q, Chen JJ. Effect of closed-loop temperature management on intraoperative body temperature in pediatric patients undergoing laparoscopic surgery [in Chinese]. J Clin Anesthesiol. 2025;41(8):803–6. doi:10.12089/jca.2025.08.003.

Tables

Table 1 Baseline characteristics of children by intraoperative hypothermia status

Data are mean ± SD, median (P25, P75), or n (%). ^a^ Independent-samples t test; ^b^ Mann–Whitney U test; ^c^ χ^2 test; ^d^ Fisher's exact test. For ASA classification and surgical site, fewer than 20% of cells had expected counts < 5 (all > 1), and conclusions were confirmed by Fisher's exact test; surgical department and anesthesia method contained cells with expected counts < 1, so Fisher's exact test was used and no χ^2 statistic is reported. Surgical duration and intraoperative blood loss are intraoperative variables that cannot be known before incision; they are presented for descriptive comparison and were treated as candidate predictors rather than as baseline characteristics.

Table 2 Multivariable logistic regression analysis of intraoperative hypothermia

β, regression coefficient; SE, standard error; OR, odds ratio; CI, confidence interval; df, degrees of freedom. Each predictor was entered with a single degree of freedom.

Table 3 Performance of the five prediction models

AUC, area under the receiver operating characteristic curve; HL, Hosmer–Lemeshow. Accuracy, precision, recall, and F1 score were computed at the default probability threshold of 0.5. The decision tree's training-set HL χ^2 equals exactly 0 because each leaf's predicted probability equals the observed event rate within that leaf (perfect in-sample calibration by construction).

Table 4 Bootstrap internal validation of the five models

Optimism and optimism-corrected AUC were estimated from 200 bootstrap resamples of the training set; 95% CIs for validation AUCs were derived from 1,000 bootstrap resamples. Apparent AUCs in this table are averaged across the 200 resamples used to estimate optimism and therefore differ slightly from the single-split training-set AUCs in Table 3 (random forest 0.935 vs 0.933; XGBoost 0.910 vs 0.911). Subgroup AUCs of the random forest model in the validation set: boys 0.895; girls 0.886; infants/toddlers (0–3 y) 0.871; preschool (4–6 y) 0.858; school-age (7–14 y) 0.913.

Figure legends

Figure 1 Study flow diagram. Of 4,124 children assessed for eligibility, 487 were excluded because they were older than 14 years (15 years, n = 146; 16 years, n = 150; 17 years, n = 191). The remaining 3,637 children (177 intraoperative hypothermia events, 4.87%) were randomly split 7:3, stratified by outcome, into a training set (n = 2,545; 124 events) used for predictor selection, model development, and bootstrap internal validation, and a validation set (n = 1,092; 53 events) used for performance evaluation. LR, logistic regression; DT, decision tree; RF, random forest; ROC, receiver operating characteristic; AUC, area under the ROC curve; HL, Hosmer–Lemeshow; DCA, decision curve analysis.

Figure 2 LASSO regression variable selection. (A) Cross-validation error curve; the red dashed line marks the optimal penalty λ. (B) Coefficient regularization paths; as log(λ) increases, coefficients of weak predictors are progressively shrunk to zero.

Figure 3 Receiver operating characteristic (ROC) curves of the five models in the training set. Model performance values are provided in Table 3.

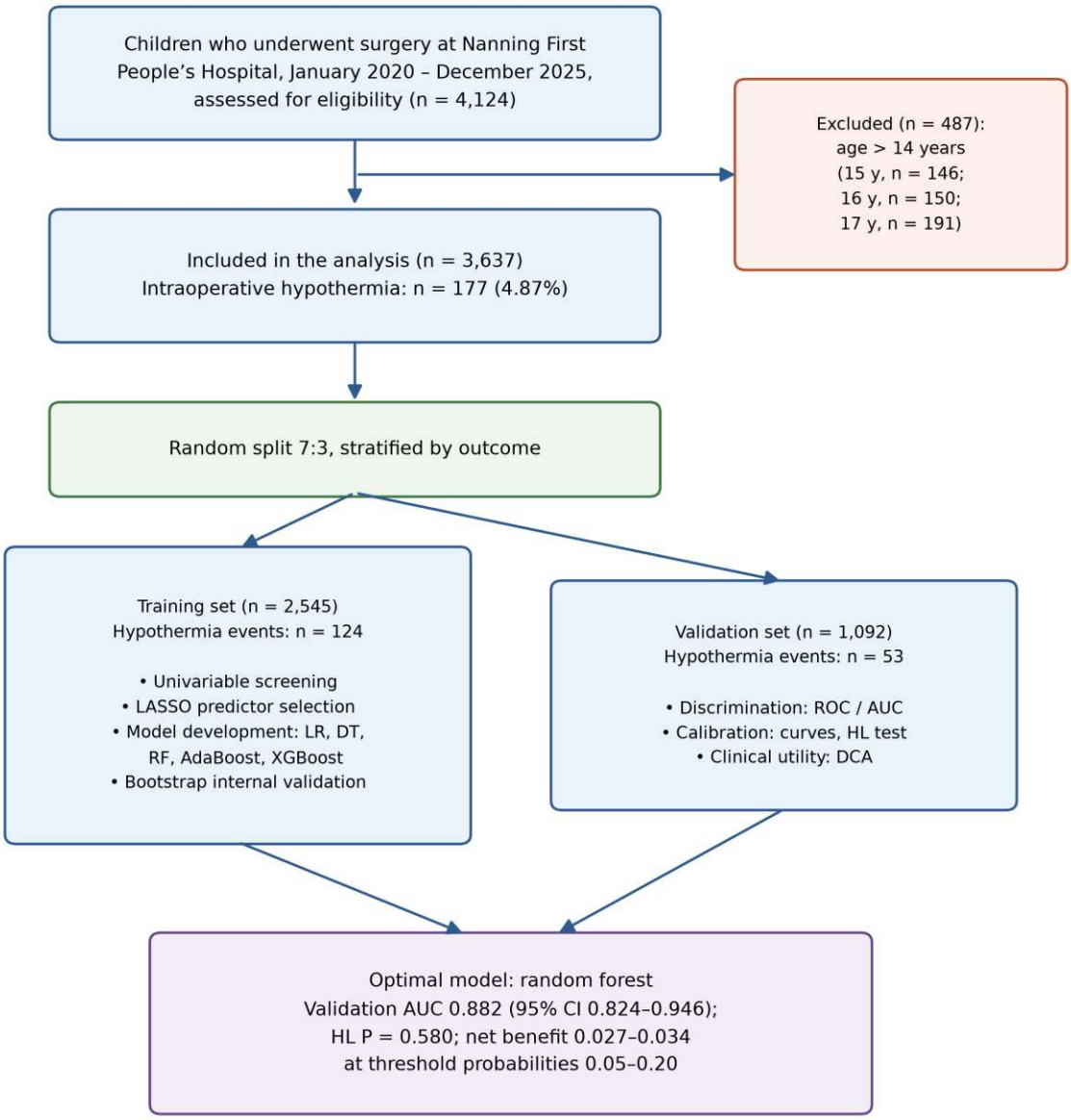
Figure 4 ROC curves of the five models in the validation set. Model performance values are provided in Table 3.

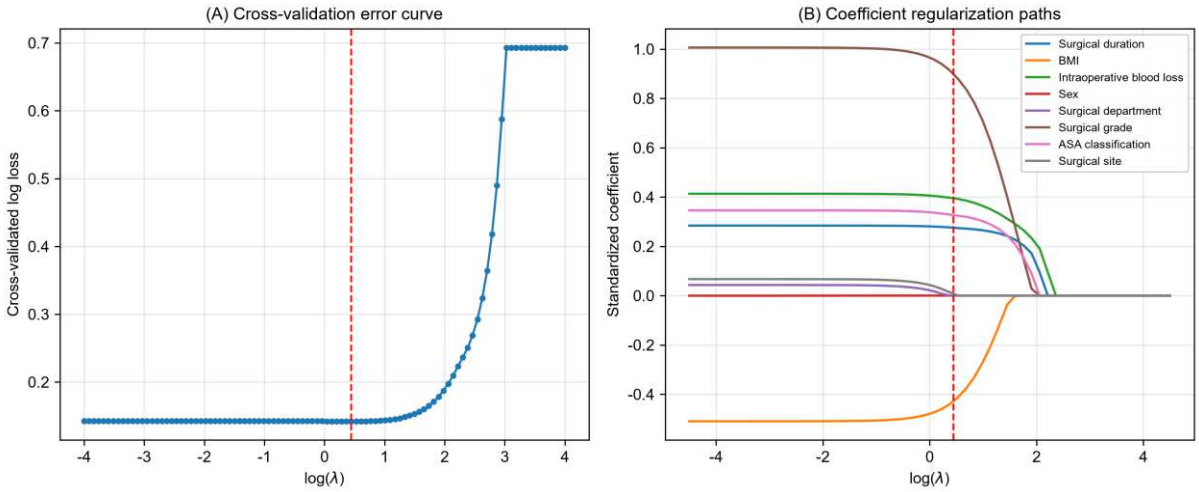
Figure 5 Calibration curves of the five models in the validation set. The diagonal represents perfect agreement between predicted probability and observed event rate; the closer the curve lies to the diagonal, the better the calibration.

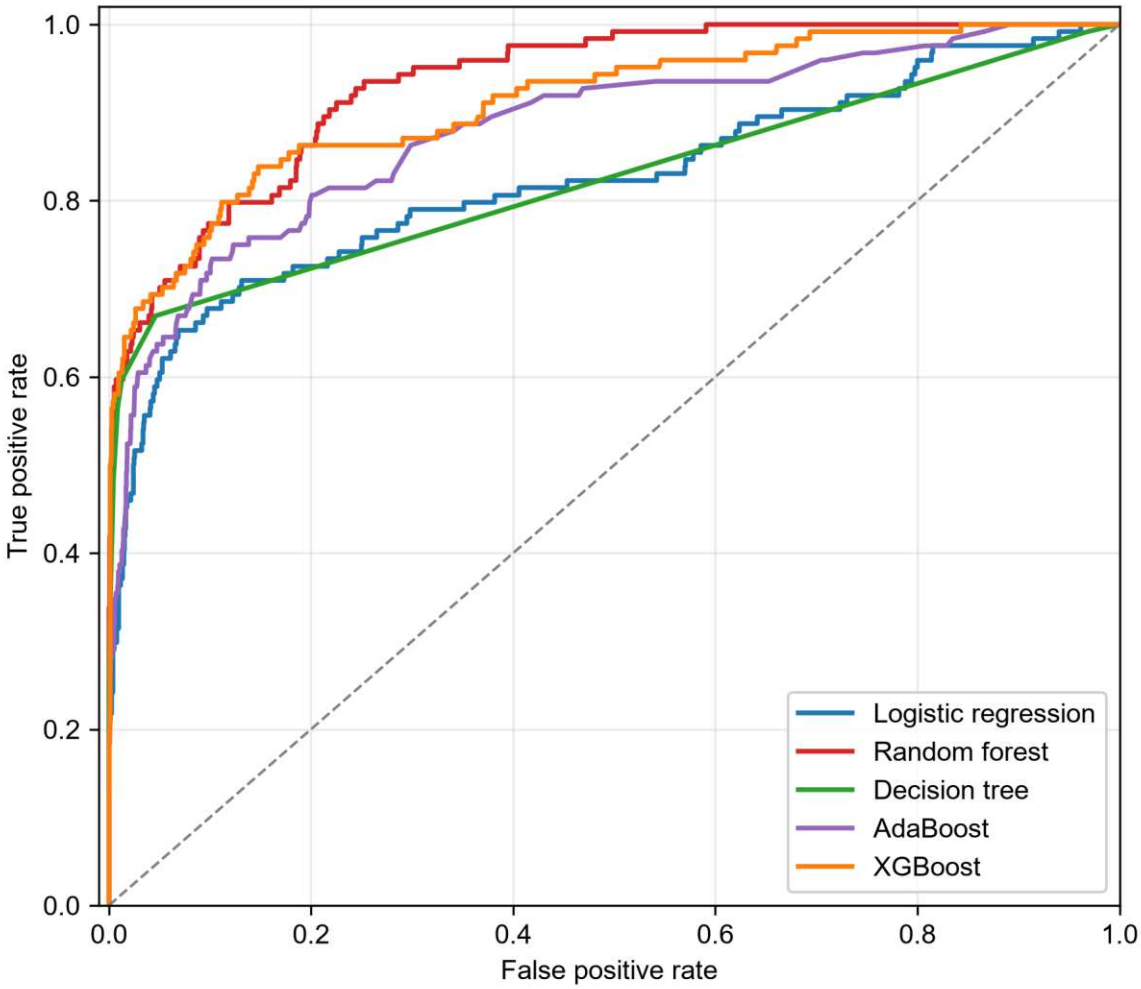
Figure 6 Global SHAP feature importance of the random forest model. Predictors are ordered by mean absolute SHAP value; longer bars indicate greater overall contribution to model predictions.

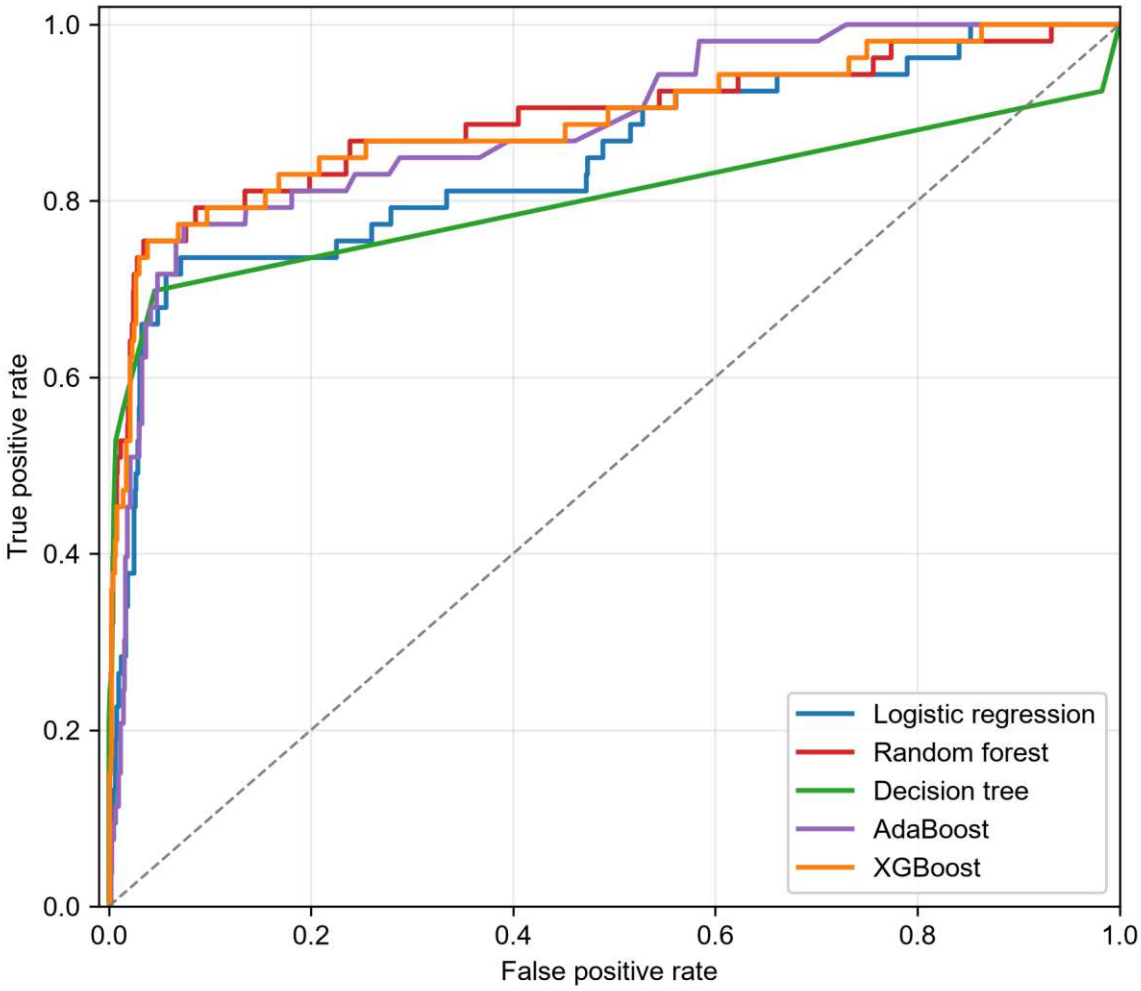
Figure 7 SHAP beeswarm plot of the random forest model. The horizontal axis shows the SHAP value: positive values increase and negative values decrease the predicted risk of intraoperative hypothermia. Each point is one child; color encodes the predictor value (red = high, blue = low).

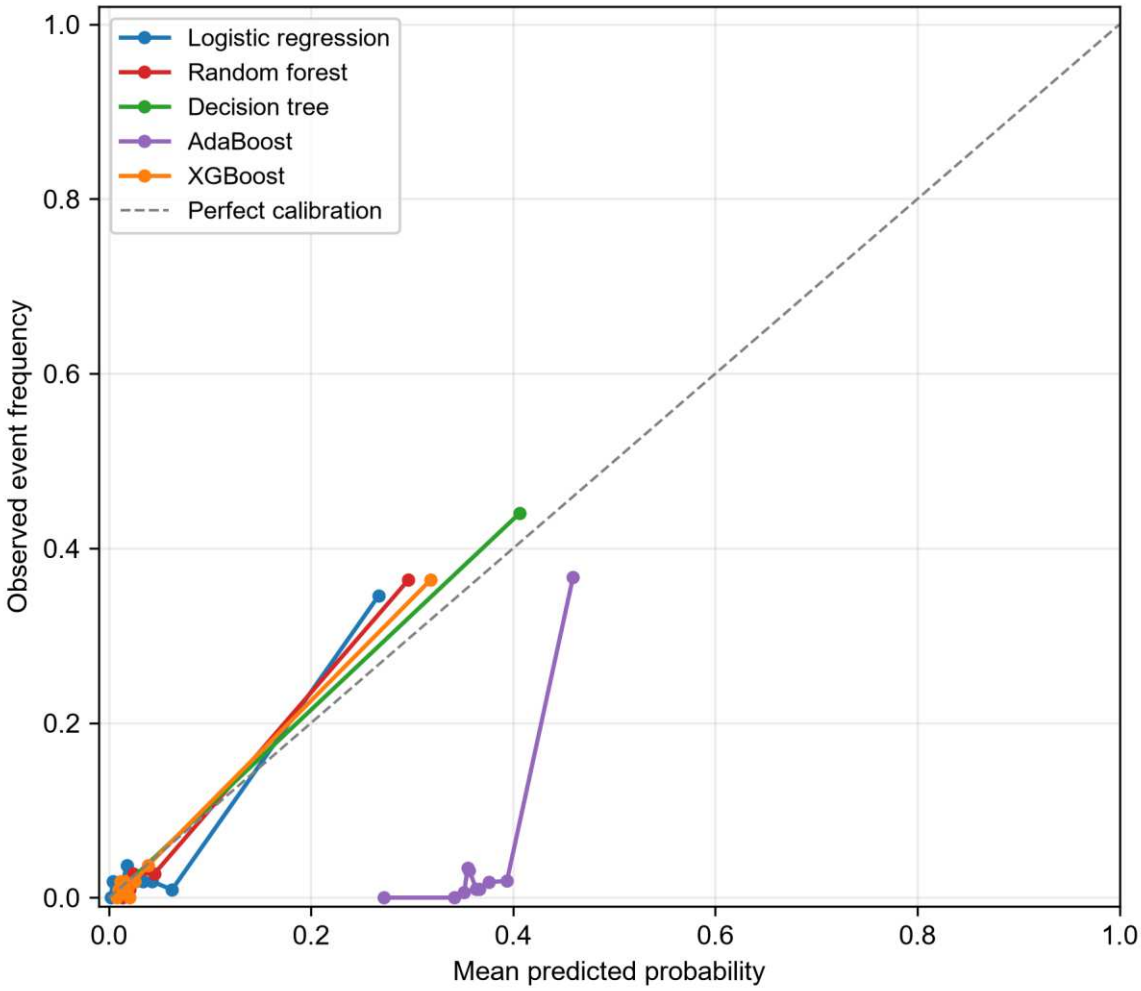
Figure 8 Decision curve analysis of the five models in the validation set. Net benefit of the five models is compared with the “treat all” and “treat none” strategies across threshold probabilities.

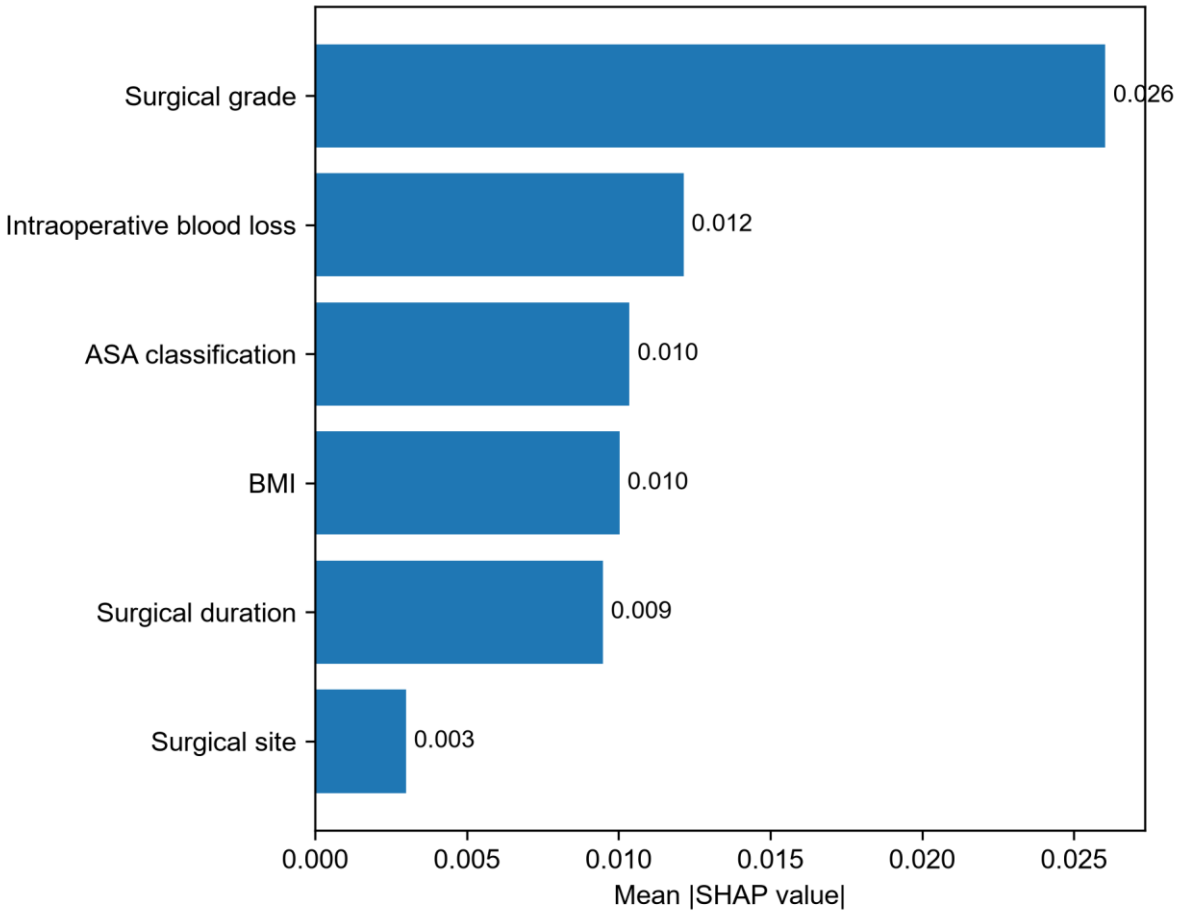

Children who underwent surgery at Nanning First People's Hospital, January 2020 – December 2025, assessed for eligibility (n = 4,124)
Excluded (n = 487): age > 14 years (15 y, n = 146; 16 y, n = 150; 17 y, n = 191)
Included in the analysis (n = 3,637) Intraoperative hypothermia: n = 177 (4.87%)
Random split 7:3, stratified by outcome
Training set (n = 2,545) Hypothermia events: n = 124
• Univariable screening
• LASSO predictor selection
• Model development: LR, DT, RF, AdaBoost, XGBoost
• Bootstrap internal validation
Validation set (n = 1,092) Hypothermia events: n = 53
• Discrimination: ROC / AUC
• Calibration: curves, HL test
• Clinical utility: DCA
Optimal model: random forest
Validation AUC 0.882 (95% CI 0.824–0.946);
HL P = 0.580; net benefit 0.027–0.034
at threshold probabilities 0.05–0.20

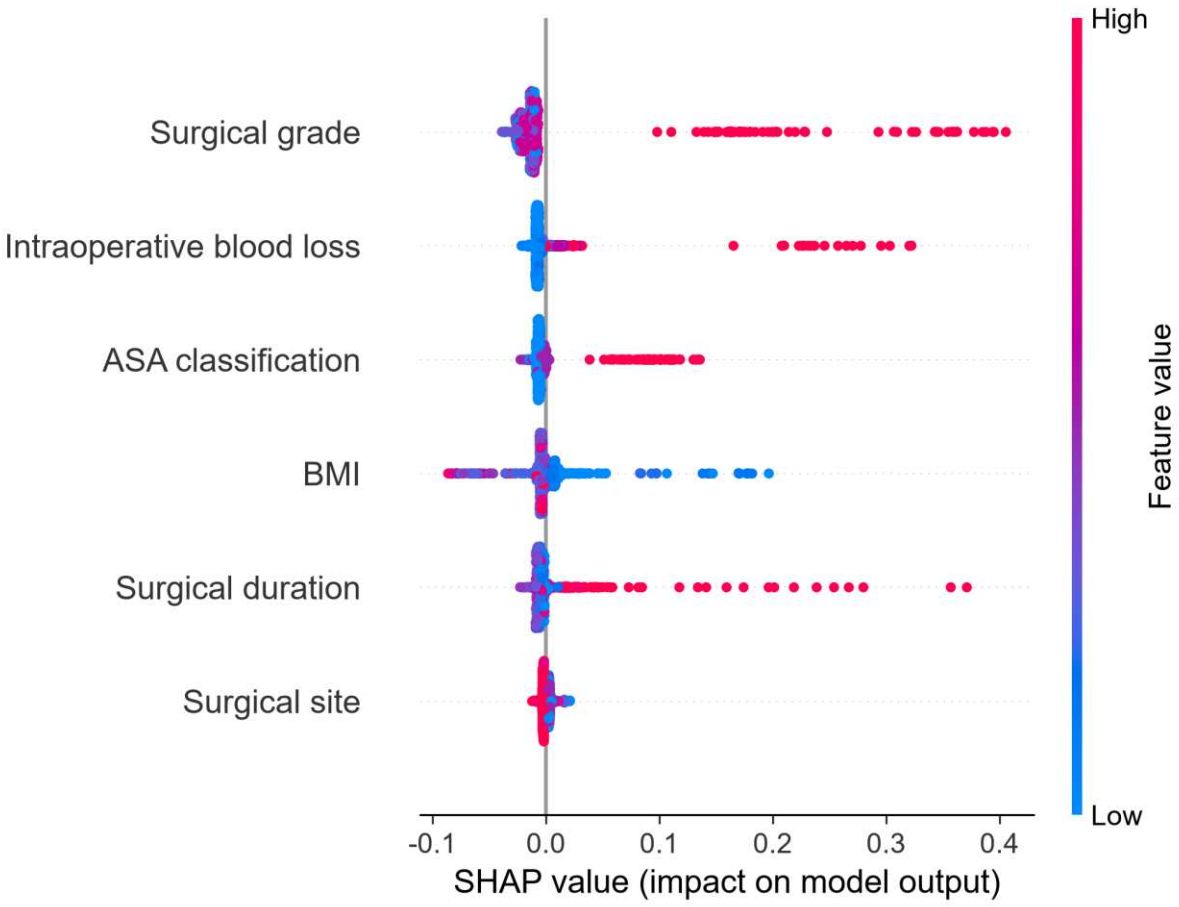

(A) Cross-validation error curve
Cross-validated log loss
log(λ)
(B) Coefficient regularization paths
Standardized coefficient
log(λ)
Surgical duration
BMI
Intraoperative blood loss
Sex
Surgical department
Surgical grade
ASA classification
Surgical site

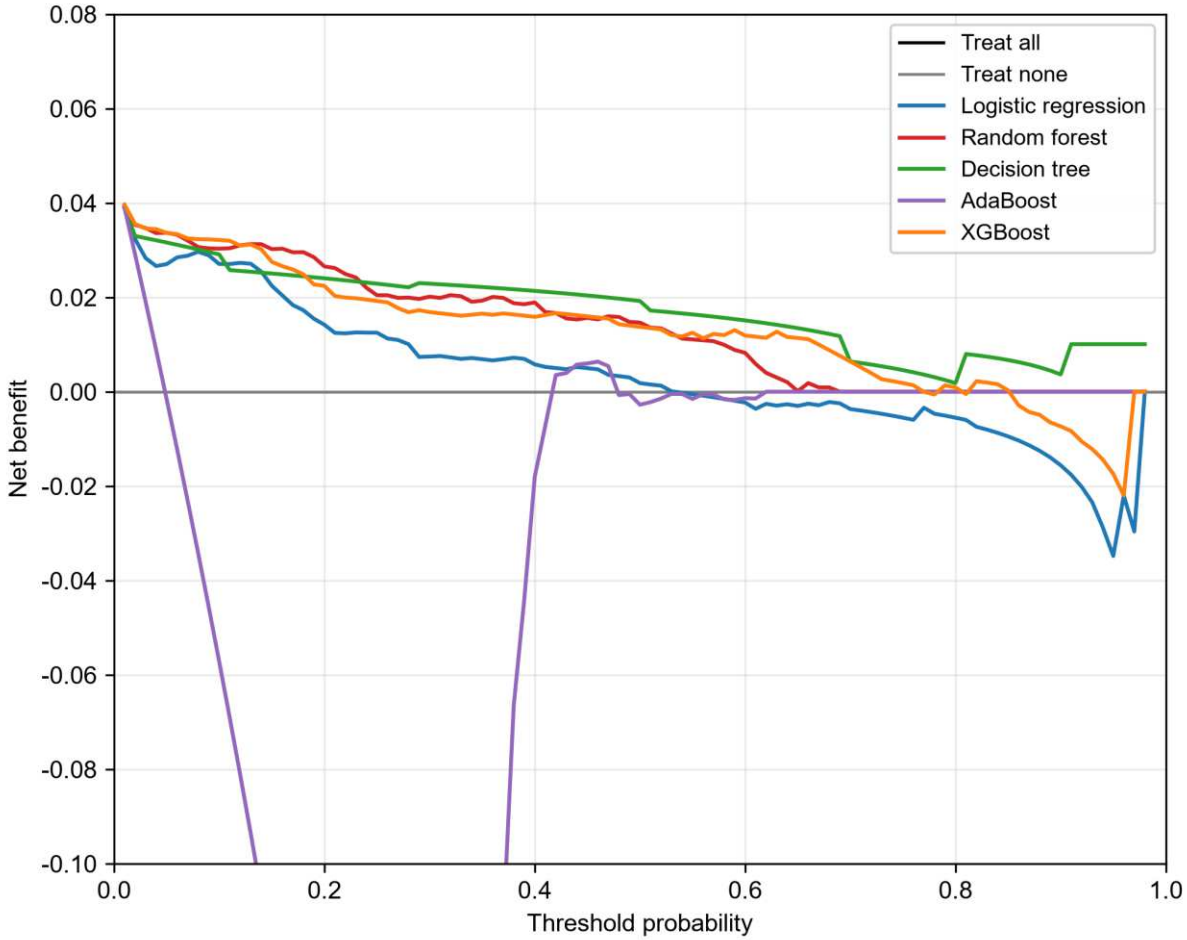

True positive rate
False positive rate
Logistic regression
Random forest
Decision tree
AdaBoost
XGBoost


True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic regression
Random forest
Decision tree
AdaBoost
XGBoost


Logistic regression
Random forest
Decision tree
AdaBoost
XGBoost
Perfect calibration
Observed event frequency
Mean predicted probability
0.0
0.2
0.4
0.6
0.8
1.0


Surgical grade
Intraoperative blood loss
ASA classification
BMI
Surgical duration
Surgical site
0.026
0.012
0.010
0.010
0.009
0.003
0.000
0.005
0.010
0.015
0.020
0.025
Mean |SHAP value|


Surgical grade
Intraoperative blood loss
ASA classification
BMI
Surgical duration
Surgical site
-0.1
0.0
0.1
0.2
0.3
0.4
SHAP value (impact on model output)
High
Feature value
Low


Treat all
Treat none
Logistic regression
Random forest
Decision tree
AdaBoost
XGBoost
Net benefit
Threshold probability
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.08
-0.10
0.0
0.2
0.4
0.6
0.8
1.0

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.docx