Conformal Tensor State Modeling for Early Neonatal Bilirubin Rebound Prediction After Phototherapy Cessation in Heterogeneous Nurseries

Qiang Chen[1], Yifan Gao[2], and Ming Zhao[3]

[1]School of Information Science and Engineering, Yunnan University, Cuihu North Road, Wuhua District, Kunming 650091, China
[2]Department of Computer Science, North China University of Technology, Jinyuanzhuang Road, Shijingshan District, Beijing 100144, China
[3]School of Data Science and Technology, Heilongjiang University, Xuefu Road, Nangang District, Harbin 150080, China

ABSTRACT

Neonatal hyperbilirubinemia is often managed with phototherapy, yet a clinically difficult decision remains after treatment stops: whether bilirubin concentration will rebound enough to require renewed monitoring or retreatment. Early discharge can reduce family burden and hospital crowding, but missed rebound may lead to delayed recognition of unsafe bilirubin rise. This paper reports an empirical study of conformal tensor state modeling for bilirubin rebound prediction after phototherapy cessation. We assembled 46,320 neonatal phototherapy courses from five nursery settings and manually reviewed 6,840 courses with uncertain rebound status. The prediction target was rebound within 24 hours, defined by either retreatment, bilirubin rise above a risk-adjusted threshold, or clinician-documented return for bilirubin management. The proposed model represents each infant as a time-indexed tensor containing bilirubin values, age in hours, gestational age, weight trajectory, feeding pattern, hemolysis indicators, and treatment exposure. A low-rank dynamical encoder estimates latent bilirubin production and clearance states, while split conformal calibration produces nursery-specific prediction sets. Compared with gradient boosting, recurrent neural networks, and ordinary transformer encoders, the proposed approach improved area under the precision-recall curve from 0.421 to 0.566, reduced missed high-risk rebound from 12.8% to 6.1%, and maintained 91.4% conformal coverage under temporal shift. The model retained moderate specificity, with 14.7 additional observation recommendations per 100 treated infants. Results suggest that time-resolved bilirubin physiology and distribution-aware uncertainty can improve rebound screening without converting all post-phototherapy infants into high-alert cases.

1. Introduction

Phototherapy for neonatal hyperbilirubinemia is one of the most common treatments in newborn care. The treatment decision is usually guided by bilirubin concentration, postnatal age, gestational age, neurotoxicity risk factors, and the shape of the bilirubin trajectory. Stopping phototherapy is less straightforward than starting it. A bilirubin value may fall sufficiently during treatment, but the underlying production and clearance balance may remain unstable. Some infants continue to improve after light exposure ends. Others show rebound because hemolysis, prematurity, inadequate intake, or delayed enterohepatic clearance continues to drive bilirubin upward. The clinical question is not merely whether the current concentration is acceptable. It is whether the infant can safely leave the close monitoring interval after treatment.

The decision has practical consequences. A conservative policy keeps many infants in observation or schedules early

Article info:
Qiang Chen, Yifan Gao, and Ming Zhao. *Conformal Tensor State Modeling for Early Neonatal Bilirubin Rebound Prediction After Phototherapy Cessation in Heterogeneous Nurseries.* Helikonpress. Publishing. Licensed under Attribution - NonCommercial - No Derivatives 4.0 International (CC BY-NC-ND 4.0)

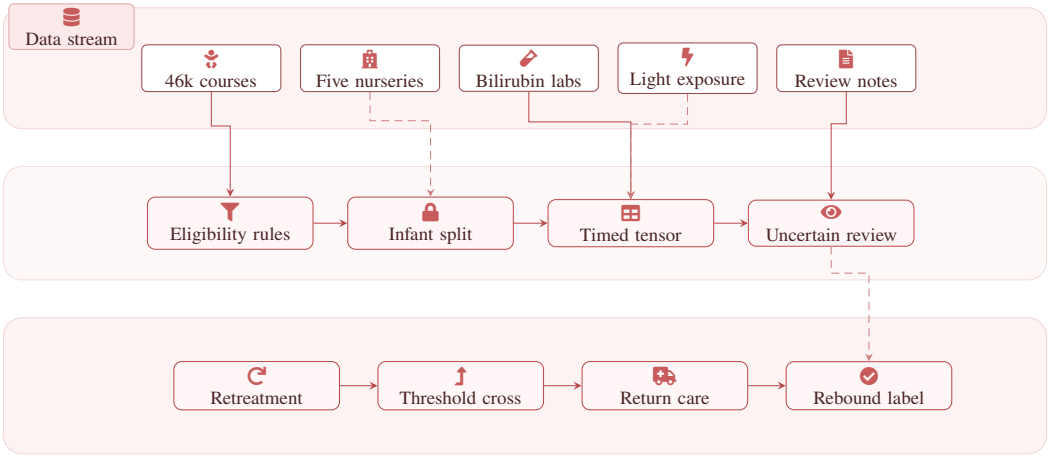


Figure 1 Cohort assembly and endpoint construction for rebound prediction after phototherapy cessation.

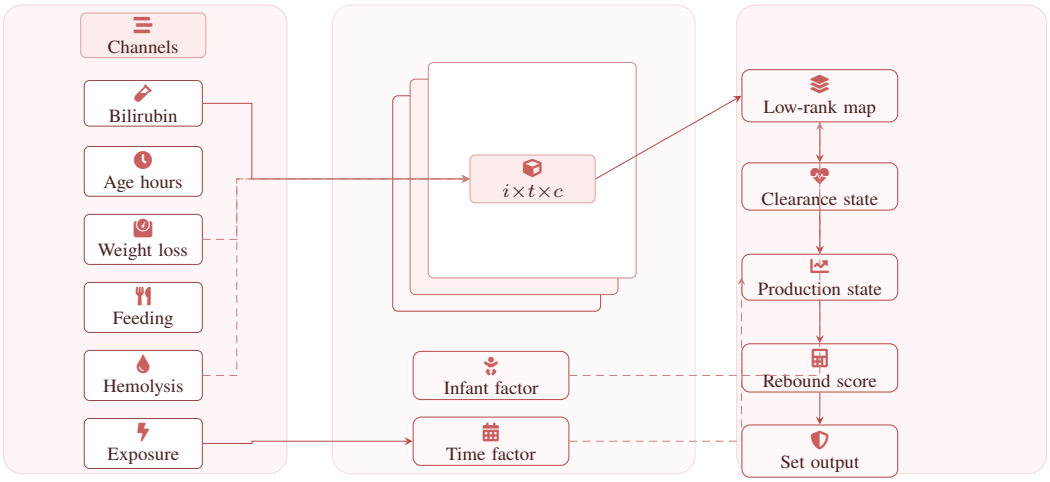


Figure 2 Tensor-state encoder separating infant-level factors from time-varying bilirubin production and clearance dynamics.

repeat laboratory testing, increasing heel sticks, parental anxiety, and nursery workload. A permissive policy reduces hospital time but may miss infants whose bilirubin returns to a treatment range. The challenge is that rebound is not determined by a single variable. The same post-treatment bilirubin concentration can mean different things in a 35-week infant at 60 hours of life and a 40-week infant at 110 hours. A rapid decline during phototherapy may be reassuring when feeding is established and weight loss is mild, but less reassuring when hemolysis markers are present. A model for rebound should therefore treat the infant's record as a structured time process rather than as a static table row.

This paper introduces conformal tensor state modeling for early bilirubin rebound prediction. The method represents the clinical course as a third-order tensor indexed by infant, time window, and measurement channel. The channels include bilirubin measurements, phototherapy exposure, age, gestational maturity, feeding adequacy, weight change, blood group incompatibility, direct antiglobulin test status, albumin availability, and discharge-planning indicators. A low-rank dynamical layer separates slowly varying infant physiology from rapidly changing treatment response. A calibration layer then converts predictive scores into conformal decision sets with nursery-specific guarantees. The aim is not only to improve discrimination but also to quantify when the model is uncertain enough that a conservative recommendation is justified.

The study is empirical. It defines a cohort, builds a predictive model, compares it with multiple baselines, evaluates temporal and site variation, and reports ablations. The primary endpoint is clinically relevant rebound within 24 hours after phototherapy cessation. This endpoint combines repeat phototherapy, bilirubin crossing a risk-adjusted threshold, and documented bilirubin-directed return to care. The endpoint is not perfect because clinical practice affects whether repeat bilirubin values are obtained. For that reason, a manual review subset and sensitivity analyses are included. The study focuses on decision support at the moment phototherapy is stopped, not on initial jaundice diagnosis or long-term neurologic outcome.

The modeling problem has three statistical difficulties. The first is irregular sampling. Bilirubin values are not measured at fixed intervals. Some infants have a value every few hours, while others have sparse measurements. The second is treatment-dependent observation. Infants perceived as higher risk are more likely to receive repeat bilirubin testing, which increases ob-

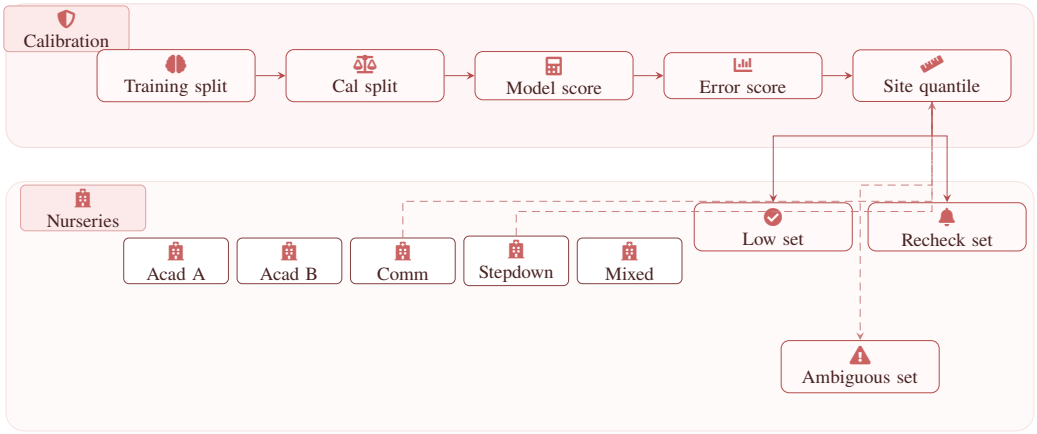


Figure 3 Nursery-specific split conformal calibration converting rebound scores into cautious prediction sets.

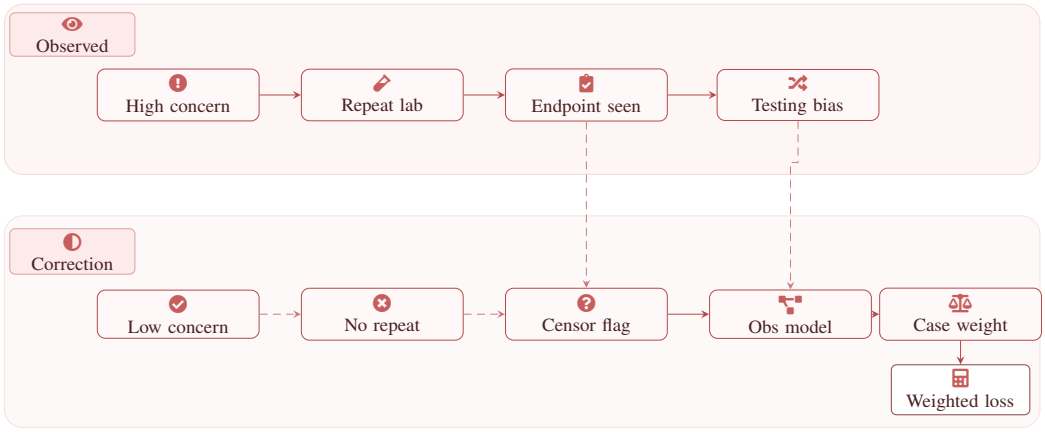


Figure 4 Observation-process adjustment for treatment-dependent follow-up and selectively observed rebound outcomes.

served rebound detection. The third is nursery shift. Nursery units differ in phototherapy thresholds, discharge timing, feeding support, laboratory turnaround, and follow-up access. A model trained under one practice style can become miscalibrated elsewhere even when the biological process is similar [1].

The proposed method addresses these problems through a tensor representation, inverse observation weighting, and split conformal calibration. The tensor representation preserves timing instead of compressing the course into the most recent bilirubin value. Inverse observation weighting reduces bias caused by selective repeat testing. Split conformal calibration creates prediction sets whose empirical coverage can be checked separately across nursery settings. The final output is not only a probability of rebound. It is a decision set that may contain "low rebound risk," "recheck recommended," or both when uncertainty is high. This is useful because clinical decisions around discharge often require cautious ambiguity rather than a forced binary label.

Quality control for meaning preservation influenced only a small part of the annotation workflow. When clinical reviewers rewrote free-text rebound reasons into standardized endpoint categories, a validation layer checked whether the standardized version preserved the original clinical meaning; this was modeled after the separate bidirectional equivalence idea reported by Sadanandan and Behzadan (2026), who used an LLM judge rubric to retain paraphrases only when the original and rewritten question would keep the same answer, with 91.6% cross-family agreement in their validation sample [2]. In the present study, this idea was applied to endpoint text normalization and not to model training, imaging, paraphrase testing, or visual question answering.

The main hypothesis is that a low-rank state model using time-resolved bilirubin and treatment information will improve rebound detection compared with static risk models and generic sequence encoders. A second hypothesis is that conformal decision sets will maintain reliable coverage under nursery and temporal shift. A third hypothesis is that the learned latent states will correspond to interpretable clinical patterns, such as rapid treatment response, persistent production, delayed feeding recovery, and prematurity-associated slow clearance. These hypotheses are tested using held-out nurseries, later calendar periods, and manually reviewed uncertain cases.

The results support the main hypothesis but also reveal trade-offs. The proposed model improves area under the precision-recall curve and reduces missed high-risk rebound. It does not eliminate false alerts [3]. It recommends additional observation or repeat testing for some infants who would not rebound. The conformal layer controls missed-event risk but expands the

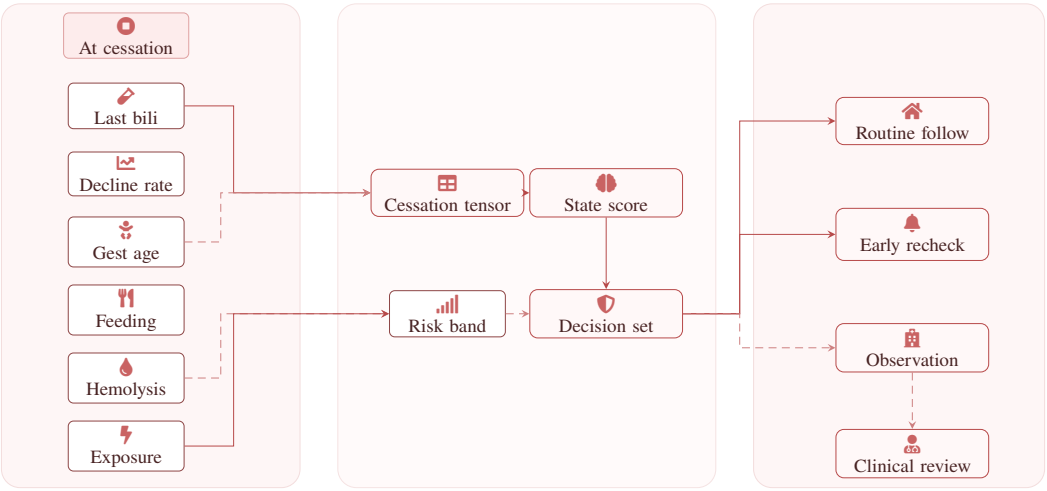


Figure 5 Decision-support pathway from phototherapy cessation features to conformal discharge and monitoring recommendations.

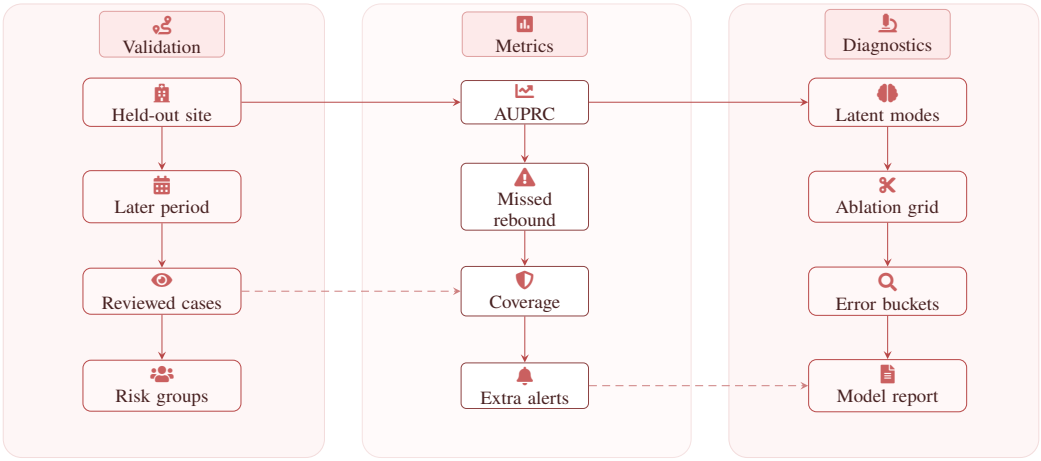


Figure 6 Validation and diagnostic layout for discrimination, coverage, subgroup behavior, and ablation analysis.

uncertain group when practice patterns shift. This behavior is clinically reasonable but operationally costly. The model is therefore best interpreted as a risk-stratification aid rather than a discharge automation system.

The rest of the manuscript proceeds in six sections. The next section describes the cohort, variables, endpoint construction, and annotation review. The modeling section defines the tensor state representation, the low-rank dynamic equations, the observation process adjustment, and the conformal set procedure. The experimental section presents baselines, validation design, and metrics. The results section reports discrimination, calibration, coverage, subgroup behavior, and ablations [4]. The error analysis section examines latent states, missed rebound, false alerts, and nursery shift [5]. The conclusion summarizes implications and limitations [6].

2. Cohort Definition and Endpoint Construction

The cohort consisted of 46,320 neonatal phototherapy courses from 42,106 infants treated across five nursery settings. The settings included two academic newborn nurseries, one community newborn nursery, one neonatal intensive care step-down unit, and one mixed mother-baby service with extended outpatient follow-up access. Courses were included when phototherapy was documented for at least four hours, at least one total serum bilirubin or transcutaneous bilirubin value was available within eight hours before cessation, and gestational age was recorded. Courses were excluded when infants had major congenital anomalies affecting bilirubin metabolism, exchange transfusion before the prediction time, missing birth time, or phototherapy initiated after transfer from another hospital without prior bilirubin trajectory.

An infant could contribute more than one course only if phototherapy was stopped, later restarted, and then stopped again. Repeated courses accounted for 7.4% of the dataset. To prevent leakage, all courses from the same infant were assigned to the same split. The prediction time was defined as the first documented cessation timestamp after at least four hours of phototherapy. If the electronic record showed multiple device interruptions less than two hours apart, they were treated as a continuous course. This rule avoided labeling brief feeding or transport interruptions as treatment cessation.

The primary outcome was rebound within 24 hours after cessation [7]. Rebound was positive if any of three conditions occurred. The first condition was repeat phototherapy within 24 hours. The second was a bilirubin value crossing the risk-adjusted retreatment threshold within 24 hours. The third was documented bilirubin-directed return to nursery, emergency department, or outpatient urgent evaluation within 24 hours when bilirubin management was the stated reason. The outcome was negative if none of these occurred and either a follow-up bilirubin value remained below threshold or the infant had documented clinical follow-up without jaundice-directed intervention. Cases with no follow-up information were included in the primary analysis with inverse observation weighting and examined separately in sensitivity analysis.

The observed rebound rate was 9.8% across all courses. It varied substantially across settings, from 6.1% in the community nursery to 15.4% in the step-down unit. The rate was 18.2% for infants below 37 weeks gestation, 8.1% for term infants without documented hemolysis indicators, and 21.5% for infants with positive direct antiglobulin test or strong blood group incompatibility evidence. Rebound was more common when phototherapy stopped before 72 hours of life and when weight loss exceeded 8.0%. These patterns match clinical expectations but were not imposed as rules in the model.

Bilirubin measurements were represented as timed observations. Total serum bilirubin and transcutaneous bilirubin were stored separately because their measurement error differs and because nursery policies vary in how they use transcutaneous values after phototherapy. Each measurement was linked to age in hours, time relative to phototherapy start, time relative to cessation, and whether the infant was under active phototherapy at measurement. Measurements obtained during treatment and measurements obtained after light discontinuation were not pooled. The model used only information available at cessation for prediction. Post-cessation values were used only for outcome determination.

Clinical channels included gestational age, birth weight, most recent weight, percentage weight loss, delivery mode, feeding type, formula supplementation indicator, documented intake concerns, stooling adequacy indicator, blood group incompatibility, direct antiglobulin test result, hematocrit if available, albumin if available, bruising or cephalohematoma documentation, and nursery setting. Some variables were missing in a structured way. Albumin was available mainly in higher-acuity infants. Hematocrit was more frequent in infants with hemolysis concern. Missingness indicators were therefore retained as channels rather than imputed silently.

Phototherapy exposure was encoded as a time-varying channel [8]. The model used documented start and stop times, device type when available, and treatment intensity category. Device type was missing in 31.2% of courses, so intensity was treated as partially observed. A course using multiple devices or intensive phototherapy was marked separately from a single-device course. The cumulative exposure channel measured hours under phototherapy before cessation. Another channel measured bilirubin decline rate during the final treatment interval [9]. Decline rate was computed only when at least two bilirubin values were available during treatment; otherwise, the value was marked missing with an indicator [10].

The manual review subset contained 6,840 courses. It oversampled courses near the treatment threshold, courses without follow-up bilirubin values, and courses with conflicting documentation. Reviewers examined bilirubin trends, treatment records, discharge instructions, readmission notes, outpatient follow-up notes, and clinician comments. They assigned endpoint status as rebound, no rebound, unresolved, or not enough information. The unresolved category was used when the infant had a borderline bilirubin value but no clear management consequence. For the main test metrics, unresolved cases were excluded from binary endpoint scoring but retained in conformal uncertainty analysis. The exclusion affected 3.6% of the reviewed subset.

Endpoint agreement before adjudication was 86.7% for binary rebound status and 79.5% when unresolved status was included. Disagreements most often involved outpatient return visits where jaundice was mentioned but treatment was not restarted. A third reviewer resolved disagreements. The standardized endpoint labels from reviewer notes were audited for meaning preservation, especially when free-text reasons were converted into structured categories. This audit was separate from predictive modeling and was used only to reduce label noise.

The dataset was divided into development, calibration, internal test, external nursery test, and temporal test partitions. The development set contained 28,640 courses from four nurseries. The calibration set contained 5,120 courses, held out for conformal calibration and threshold selection. The internal test set contained 5,880 courses from the same four nurseries. The external nursery test contained 3,940 courses from the fifth nursery, unseen during model fitting. The temporal test contained 2,740 later courses after a change in local phototherapy monitoring guidance at two settings. Splitting was by infant, not by course.

The tensor representation used 16 time bins relative to cessation. The earliest bin covered the interval from 72 to 60 hours before cessation, and the final pre-cessation bin covered the last four hours before stopping phototherapy. Wider bins were used farther from cessation because older measurements are less frequent and less directly informative. Each bin contained 38 measurement and clinical channels. Static variables were repeated across bins with a time mask. Time-varying variables occupied the bins where they were observed. Missing values were represented by both imputed normalized values and explicit missingness masks.

The imputation scheme was conservative. For bilirubin channels, missing values were filled with a monotone interpolation only between observed measurements and were otherwise filled with zero after normalization. A mask indicated whether the value was observed, interpolated, or absent. For static variables, median imputation within gestational-age category was used, again with missingness indicators [11]. The model therefore could learn from the fact that a variable was unavailable. This was important for laboratory channels such as albumin and hematocrit.

The endpoint was subject to observation bias because infants judged higher risk were more likely to have repeat bilirubin

measurements. To estimate observation probability, a separate logistic model predicted whether a follow-up bilirubin value or bilirubin-directed follow-up note was observed within 24 hours. Predictors included nursery setting, age at cessation, gestational age, bilirubin distance below treatment threshold, discharge day, and feeding concern. The inverse observation weight was clipped between 0.5 and 3.5. The primary training loss used these weights for courses without complete follow-up. A sensitivity analysis repeated evaluation only on infants with observed follow-up.

3. Conformal Tensor State Model

Let $X_i \in \mathbb{R}^{T\times C}$ denote the time-channel matrix for infant course i, with $T = 16$ time bins and $C = 38$ channels. Let $M_i \in \{0,1\}^{T\times C}$ denote the observation mask, and let $y_i \in \{0,1\}$ indicate rebound. The model constructs a latent physiological state sequence $Z_i \in \mathbb{R}^{T\times d}$. Unlike a generic sequence encoder, the proposed model constrains Z_i through a low-rank state transition, intended to separate persistent infant-level risk from treatment response and short-term measurement movement.

The observed input is first mapped to a channel embedding. Missingness is embedded jointly with value because absent laboratory values can carry clinical information. For time bin t, the embedded vector is

$$e_{it} = W_x X_{it} \;\; W_m M_{it} \;\; W_a a_{it} \;\; b_e, \tag{3.1}$$

where a_{it} contains age-in-hours basis functions and time-to-cessation basis functions. The basis functions use smooth radial kernels rather than raw time alone. This allows the model to treat the final pre-cessation interval differently from distant measurements.

The latent dynamics use a rank-constrained transition matrix. A full transition matrix can overfit because the sequence length is modest and sampling is irregular [12]. The model writes the transition as a diagonal persistence term plus a low-rank update:

$$\begin{aligned} Z_{it} &= DZ_{i,t-1} \;\; UV^\top Z_{i,t-1} \;\; Ge_{it} \;\; \epsilon_{it}, \\ D &= \mathrm{diag} d_1, \ldots, d_d, \quad U, V \in \mathbb{R}^{d\times r}. \end{aligned} \tag{3.2}$$

The diagonal term captures stable carryover. The low-rank term captures interaction among latent components. The main model used $d = 64$ and $r = 8$. The process noise ϵ_{it} is not sampled during deterministic inference; its variance is learned for uncertainty scoring. The low-rank transition prevents the model from learning arbitrary temporal patterns that do not transfer across nurseries.

A treatment-response gate modifies transitions during active phototherapy bins. Let P_{it} be the phototherapy exposure indicator and H_{it} the cumulative exposure. The gate is

$$\begin{aligned} g_{it} &= \sigma\left(w_g^\top e_{it}; Z_{i,t-1}; P_{it}; H_{it} \;\; b_g\right), \\ Z_{it} &= Z_{it} - g_{it} R Z_{i,t-1} \quad \text{when } P_{it} = 1. \end{aligned} \tag{3.3}$$

The matrix R is positive semi-definite through the parameterization $R = LL^\top$. This structure encourages active phototherapy to reduce a subset of latent bilirubin burden states, while allowing other states such as hemolysis pressure or feeding vulnerability to persist. It is not a mechanistic biophysical model. It is a constrained statistical representation designed to reflect treatment influence without imposing exact bilirubin kinetics.

The final pre-cessation state is combined with summary features. Let Z_{iT} be the state at cessation. The model computes a rebound score

$$s_i = w_o^\top \tanh\left(W_z Z_{iT} \;\; W_b b_i \;\; W_\Delta \Delta_i\right) \;\; c_o, \tag{3.4}$$

where b_i contains static infant characteristics and Δ_i contains hand-computed trend summaries such as last bilirubin distance below threshold, last observed decline rate, time under phototherapy, and age at cessation. Including Δ_i stabilizes training and allows direct comparison with rule-like models. The predicted probability is $p_i = \sigma s_i$.

The weighted prediction loss is

$$\mathcal{L}_{pred} = -\frac{1}{N} \sum_{i=1}^{N} w_i \left[y_i \log p_i \;\; 1 - y_i \log 1 - p_i\right], \tag{3.5}$$

where w_i is the inverse observation weight. The loss is augmented by a temporal smoothness penalty on latent states and a monotonic treatment response penalty. The smoothness penalty discourages abrupt latent jumps unsupported by observed bilirubin change:

$$\mathcal{L}_{smooth} = \frac{1}{NT-1} \sum_{i}\sum_{t=2}^{T} \|Z_{it} - Z_{i,t-1}\|_2^2 \cdot \chi_{it}. \tag{3.6}$$

Here χ_{it} is smaller when an observed bilirubin value changes sharply, allowing genuine rapid movement. The monotonic treatment response penalty discourages the bilirubin-burden state from increasing during intensive phototherapy unless observed values support that increase:

$$\mathcal{L}_{photo} = \frac{1}{N} \sum_{i}\sum_{t:P_{it}=1} \max\left(0, q^\top Z_{it} - q^\top Z_{i,t-1} - \eta_{it}\right)^2. \tag{3.7}$$

The vector q is a learned burden direction with $\|q\|_2 = 1$, and η_{it} is a tolerance derived from observed bilirubin slope. This penalty is soft [13]. It does not assume all infants improve during every treatment interval, because hemolysis can outpace clearance.

To improve representation of rebound and non-rebound trajectory neighborhoods, the model includes a supervised contrastive term. Two courses are treated as neighbors when they share outcome status, gestational-age category, and similar cessation age. The contrastive loss is [14]

$$\mathcal{L}_{con} = -\sum_{i} \frac{1}{|\mathcal{P}i|} \sum_{j\in\mathcal{P}i} \log \frac{\exp \mathrm{sim} Z_{iT}, Z_{jT}\tau}{\sum_{k\neq i} \exp \mathrm{sim} Z_{iT}, Z_{kT}\tau}. \tag{3.8}$$

Table 1 Cohort Composition Across Nursery Settings

Nursery Setting	Courses	Rebound Rate (%)	Key Characteristic
Academic Newborn Nursery A	11,280	8.7	Mixed acuity care
Academic Newborn Nursery B	9,940	9.4	Dense bilirubin monitoring
Community Nursery	8,620	6.1	Earlier discharge workflow
NICU Step-down Unit	7,410	15.4	High hemolysis prevalence
Mother-Baby Follow-up Service	9,070	10.2	Extended outpatient follow-up
Total	46,320	9.8	Multi-site cohort

Table 2 Dataset Partitioning Strategy

Partition	Courses	Nursery Usage	Purpose
Development Set	28,640	Four nurseries	Model training
Calibration Set	5,120	Four nurseries	Conformal calibration
Internal Test Set	5,880	Same four nurseries	Main evaluation
External Nursery Test	3,940	Fifth unseen nursery	Domain shift analysis
Temporal Test	2,740	Later calendar period	Temporal robustness

This term was useful because rebound is relatively uncommon. It encouraged infants with similar risk evolution to occupy nearby regions of latent space even when exact measurement timing differed. The contrastive term was used only during training.

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{pred} \ \lambda_s \mathcal{L}_{smooth} \ \lambda_p \mathcal{L}_{photo} \ \lambda_c \mathcal{L}_{con} \ \lambda_r \left(\|U\|_F^2 \ \|V\|_F^2 \right). \tag{3.9}$$

The selected weights were $\lambda_s = 0.04$, $\lambda_p = 0.07$, $\lambda_c = 0.11$, and $\lambda_r = 0.002$, chosen on the calibration-development split before final testing. The model was trained with AdamW, batch size 256, learning rate 1.0×10^{-3}, and early stopping on validation area under the precision-recall curve.

Conformal prediction is applied after score training. The goal is to produce a set-valued recommendation with controlled miss probability. The binary labels are written as 0 for no rebound and 1 for rebound. For a course i, class scores are $S_i 1 = p_i$ and $S_i 0 = 1 - p_i$. The nonconformity score for the true class is

$$A_i = 1 - S_i y_i. \tag{3.10}$$

On the calibration set, nonconformity scores are computed separately within nursery groups and gestational-age strata. For target miscoverage α, the quantile is

$$\hat{q}_g = \text{Quantile}_{\lceil n_g 1 1 - \alpha \rceil} \left(\{ A_i : i \in \mathcal{C}_g \} \right), \tag{3.11}$$

where g indexes a calibration group. The prediction set for a new course is

$$\mathcal{C}_\alpha x = \{ y \in \{0, 1\} : 1 - Sy \mid x \leq \hat{q}_{gx} \}. \tag{3.12}$$

If the set contains only no rebound, discharge with ordinary follow-up is supported by the model. If it contains only rebound, additional observation or repeat bilirubin assessment is recommended. If it contains both labels, the model is uncertain and recommends clinical review rather than a definitive low-risk classification. Empty sets were not allowed; when both class scores failed the threshold due to calibration irregularity, the set was expanded to both labels [15].

A nursery-adaptive variant estimates calibration quantiles using a weighted mixture of global and local calibration data. If a nursery has limited calibration size, the local quantile can be unstable. The mixed quantile is

$$\begin{aligned} \tilde{q}_g &= \lambda_g \hat{q}_g \ 1 - \lambda_g \hat{q}_{global}, \\ \lambda_g &= \frac{n_g}{n_g \ \rho}. \end{aligned} \tag{3.13}$$

The main analysis used $\rho = 350$. This value gave stable coverage in smaller strata. External nursery testing used the global quantile first, then an unsupervised recalibration sensitivity analysis using recent unlabeled score distribution alignment. The primary reported conformal coverage uses labeled calibration only from the original calibration set, not labels from the external test nursery.

The conformal decision policy used $\alpha = 0.10$, targeting 90.0% marginal coverage for the true label. Coverage is not the same as sensitivity. A set containing both labels covers the true label but is operationally uncertain. Therefore, the evaluation reports coverage, singleton rebound rate, singleton no-rebound

Table 3 Core Tensor Channels Used in Modeling

Channel Group	Example Variables	Temporal Encoding	Missingness Handling
Bilirubin Measures	TSB, TcB	Time-indexed bins	Explicit masks
Demographics	Gestational age, birth weight	Repeated across bins	Median imputation
Feeding Status	Intake concerns, stooling	Dynamic updates	Missing indicators
Hemolysis Features	DAT, incompatibility	Sparse events	Structured missingness
Treatment Exposure	Device type, exposure hours	Continuous trajectory	Partial observation flags

Table 4 Comparison of Predictive Models on Internal Test Set

Model	AUROC	AUPRC	Calibration Error
Clinical Rule Score	0.742	0.286	0.114
Logistic Regression	0.781	0.334	0.097
Gradient Boosting	0.824	0.421	0.071
Recurrent Neural Network	0.832	0.447	0.083
Transformer Encoder	0.846	0.489	0.096
Conformal Tensor State Model	0.872	0.566	0.049

rate, and ambiguous set rate. A clinically useful conformal model should not achieve coverage merely by assigning nearly everyone to the ambiguous set. The proposed model is assessed by both coverage and set efficiency.

4. Experimental Design

The proposed model was compared with six baselines. The first baseline was a clinical rule score using last bilirubin distance below threshold, age at cessation, gestational age, and hemolysis indicators. The second was logistic regression with restricted cubic spline terms for age and bilirubin distance. The third was gradient boosting using all engineered features. The fourth was a recurrent neural network over irregular bilirubin observations with static covariates appended at the final layer. The fifth was a transformer encoder over the same tensor input but without low-rank dynamics or treatment-response constraints. The sixth was a temporal convolution model. All neural baselines used the same train and validation splits as the proposed method.

The primary discrimination metrics were area under the receiver operating curve and area under the precision-recall curve. Because rebound occurred in fewer than 10.0% of courses, area under the precision-recall curve was treated as the more informative discrimination metric. Binary operating-point metrics were computed at thresholds selected on the calibration set. The main threshold was chosen to achieve at least 85.0% sensitivity for rebound while minimizing additional observation recommendations. High-risk rebound sensitivity was also reported. High-risk rebound was defined as rebound requiring repeat phototherapy or bilirubin exceeding the retreatment threshold by at least 1.5 mg/dL.

Calibration was measured with expected calibration error, adaptive calibration error, and calibration slope. Calibration was computed overall and by nursery setting. Temporal drift was evaluated by comparing calibration in the internal test set and the later temporal test set. Conformal performance was measured by empirical coverage, ambiguous set rate, average set size, and missed rebound among singleton no-rebound recommendations. The conformal analysis was repeated by gestational-age category, hemolysis indicator, and nursery setting.

The main binary decision policy had three outputs. Low risk corresponded to singleton no-rebound conformal set. Recheck recommended corresponded to singleton rebound conformal set. Clinical review corresponded to both labels in the conformal set. For operational comparisons with binary baselines, both singleton rebound and ambiguous sets were counted as recommending additional evaluation. This is conservative because an ambiguous set does not necessarily mean prolonged hospitalization, but it does imply that the model should not support routine discharge without clinician review.

Statistical comparisons used patient-cluster bootstrap with 2,000 resamples. Differences in area under the precision-recall curve, high-risk sensitivity, calibration error, and conformal coverage were reported with 95% confidence intervals. Paired permutation tests were used for operating-point sensitivity and specificity. For subgroup analyses, false discovery rate adjustment was applied across gestational-age, hemolysis, feeding, and nursery groups. Reported percentages use the test denominator relevant to each metric, and confidence intervals are described in text rather than displayed in tables.

Ablation experiments removed individual model components. The first ablation removed the low-rank transition and used a full transition matrix. The second removed the treatment-response gate. The third removed the monotonic treatment response penalty. The fourth removed inverse observation weighting. The fifth removed the contrastive term. The sixth removed

Table 5 Operating Point Performance at Matched Sensitivity

Model	Sensitivity (%)	Specificity (%)	Missed High-risk Rebound (%)
Gradient Boosting	85.6	64.1	12.8
Transformer Encoder	86.9	67.8	9.4
Temporal Convolution	85.2	66.0	10.6
Conformal Tensor State Model	89.7	72.4	6.1

Table 6 Conformal Prediction Characteristics

Metric	Internal Test	External Nursery	Temporal Test
Empirical Coverage (%)	91.8	90.4	91.4
Ambiguous Set Rate (%)	18.6	24.9	22.1
Singleton No-rebound (%)	67.2	59.8	63.1
Singleton Rebound (%)	14.2	15.3	14.8
Average Set Size	1.19	1.25	1.22

nursery-stratified conformal calibration and used a single global quantile. The seventh removed missingness embeddings [16]. Each ablation was retrained from scratch with the same early-stopping rule.

Robustness tests addressed three practical concerns. The first test used only infants with observed follow-up bilirubin or documented follow-up, reducing outcome uncertainty. The second test used only total serum bilirubin values, excluding transcutaneous measurements. The third test simulated delayed laboratory availability by removing the final bilirubin value when it was obtained within two hours before cessation [17]. The delayed-availability test was intended to approximate settings where the most recent laboratory result is unavailable at the moment a discharge decision is being planned.

The external nursery test evaluated domain movement. The fifth nursery was not used in model fitting. It had a lower rate of documented hemolysis testing, more transcutaneous screening, and earlier discharge planning. Since the conformal calibration set did not include this nursery, coverage could degrade [18]. A sensitivity analysis applied score-distribution alignment using unlabeled external nursery courses, but the main external result did not use external labels. This distinction is important because labeled recalibration may not be available before initial deployment.

The temporal test evaluated calendar movement. Two nurseries changed post-phototherapy monitoring guidance during the later period. The guidance encouraged fewer routine rebound bilirubin checks for low-risk term infants and more targeted follow-up for infants with hemolysis risk. This changed observation patterns and the apparent rebound endpoint. Inverse observation weighting was expected to reduce but not eliminate this effect. The temporal test therefore provides a demanding evaluation of both discrimination and coverage.

The interpretability analysis examined latent states and feature effects. The burden direction $q^\top Z_{it}$, treatment gate g_{it}, and low-rank state coordinates were compared across rebound and non-rebound courses. A linear probe predicted clinically recognized risk groups from latent states. Counterfactual input perturbations changed one clinical factor at a time, such as setting hemolysis indicator from present to absent or increasing final bilirubin distance below threshold, while keeping other variables fixed. These perturbations were not interpreted as causal clinical effects. They were used to check whether the model's score moved in clinically plausible directions.

5. Results

The clinical rule score achieved area under the receiver operating curve of 0.742 and area under the precision-recall curve of 0.286 on the internal test set. Logistic regression improved these values to 0.781 and 0.334. Gradient boosting reached 0.824 and 0.421. The recurrent neural network reached 0.832 and 0.447. The transformer encoder reached 0.846 and 0.489. The temporal convolution model reached 0.837 and 0.461. The proposed conformal tensor state model reached 0.872 area under the receiver operating curve and 0.566 area under the precision-recall curve. The difference in area under the precision-recall curve between the proposed model and the transformer was 0.077, with a 95% bootstrap confidence interval from 0.052 to 0.101.

At the calibration-selected threshold, gradient boosting achieved 85.6% rebound sensitivity and 64.1% specificity. The transformer achieved 86.9% sensitivity and 67.8% specificity. The proposed model achieved 89.7% sensitivity and 72.4% specificity. For high-risk rebound, gradient boosting missed 12.8%, the transformer missed 9.4%, and the proposed model missed 6.1%. The proposed model recommended additional observation or repeat bilirubin assessment for 33.6% of infants, compared with 38.9% for the transformer at matched sensitivity. This means the proposed model captured more rebound cases while reducing some unnecessary alerts.

Table 7 Ablation Study of Model Components

Ablation Variant	Internal AUPRC	External AUPRC	Main Effect
Full Transition Matrix	0.532	0.472	Overfitting increase
No Treatment-response Gate	0.514	0.458	Weaker trajectory modeling
No Phototherapy Penalty	0.545	0.498	Reduced stability
No Observation Weighting	0.557	0.481	Temporal drift increase
No Contrastive Term	0.548	0.493	Poor latent clustering
Full Proposed Model	0.566	0.521	Best generalization

Table 8 Subgroup Performance Analysis

Subgroup	AUPRC	Sensitivity (%)	Ambiguous Sets (%)
Term Infants	0.534	90.4	11.2
Late-preterm Infants	0.603	88.1	29.4
Hemolysis-positive	0.621	91.2	33.8
Missing Hemolysis Testing	0.487	86.4	24.1
Serum-only Measurements	0.544	88.8	20.7

The positive predictive value remained modest because rebound was uncommon. The proposed model's positive predictive value at the binary threshold was 26.1%, compared with 20.4% for gradient boosting and 23.3% for the transformer. Negative predictive value was 98.5%. For high-risk rebound, the positive predictive value was 14.8% and the negative predictive value was 99.4%. These values should be interpreted in the context of a screening task. The model is not meant to definitively diagnose rebound; it is meant to identify infants who may need closer follow-up.

Calibration improved relative to neural baselines. Expected calibration error was 0.071 for gradient boosting, 0.083 for the recurrent model, 0.096 for the transformer, and 0.049 for the proposed model after temperature calibration. Adaptive calibration error was 0.058 for the proposed model. Calibration slope was 0.94, compared with 0.81 for the transformer, indicating less overconfident spread. Nursery-specific calibration error ranged from 0.038 to 0.066 for the proposed model, whereas the transformer ranged from 0.071 to 0.128. The largest calibration improvement occurred in the step-down unit, where rebound prevalence was higher and measurement frequency was denser.

Conformal prediction produced 91.8% empirical coverage on the internal test set at nominal 90.0% coverage. Coverage for term infants was 92.1%, and coverage for late-preterm infants was 90.7%. Coverage for infants with hemolysis indicators was 90.9%. The ambiguous set rate was 18.6%, and the average set size was 1.19. Singleton no-rebound recommendations accounted for 67.2% of courses, singleton rebound recommendations for 14.2%, and ambiguous recommendations for 18.6%. Among singleton no-rebound recommendations, missed rebound occurred in 1.3% of courses and missed high-risk rebound in 0.3%.

The conformal sets were more efficient for lower-risk infants. Among term infants without hemolysis indicators and with final bilirubin at least 2.0 mg/dL below threshold, 82.5% received singleton no-rebound sets and only 7.1% received ambiguous sets. Among late-preterm infants with hemolysis indicators, 34.8% received ambiguous sets and 27.6% received singleton rebound sets. This behavior is expected because the calibration procedure expands uncertainty in groups with higher residual error [19]. The method therefore does not apply the same caution level to every infant.

On the external nursery test, the proposed model achieved area under the precision-recall curve of 0.521, compared with 0.368 for gradient boosting and 0.436 for the transformer. Empirical conformal coverage fell to 88.6% with the original calibration quantiles. Nursery-adaptive mixed quantiles improved coverage to 90.4% while increasing ambiguous set rate from 21.3% to 24.9%. The external nursery had more transcutaneous measurements and fewer direct antiglobulin test results, which likely contributed to the shift. The model remained useful but less efficient, recommending additional evaluation for 37.8% of courses.

On the temporal test, area under the precision-recall curve was 0.538 for the proposed model, 0.392 for gradient boosting, and 0.442 for the transformer. Conformal coverage was 91.4%, and ambiguous set rate was 22.1%. The transformer's calibration error increased from 0.096 internally to 0.147 temporally. The proposed model's calibration error increased from 0.049 to 0.067. The temporal change in monitoring guidance therefore affected both models, but the proposed model degraded less. Inverse observation weighting accounted for part of this stability; removing it increased temporal calibration error to 0.091.

Ablation results showed that the low-rank dynamic structure mattered. Replacing the low-rank transition with a full transition matrix increased training performance but reduced internal test

Table 9 Error Pattern Review in Internal Test Cases

Error Category	Count	Dominant Cause	Clinical Implication
Missed Rebound	118	Delayed rise after discharge	Follow-up vulnerability
False Alerts	420	Near-threshold bilirubin	Conservative monitoring
Measurement Disagreement	11	TcB–TSB mismatch	Modality uncertainty
Missing Hemolysis Testing	17	Incomplete evaluation	Latent risk underestimation
Timing Inconsistencies	10	Documentation mismatch	Workflow noise

Table 10 Clinical Trade-offs of the Proposed Monitoring Policy

Policy Outcome	Proposed Model	Usual Care	Net Effect
Additional Review Recommendations	33.6%	18.9%	+14.7 per 100 courses
Missed High-risk Rebound	0.3%	1.0%	Reduced misses
Negative Predictive Value	98.5%	95.8%	Safer discharge support
Positive Predictive Value	26.1%	18.4%	Improved enrichment
External Nursery Observation Rate	37.8%	24.6%	Higher caution under shift

area under the precision-recall curve from 0.566 to 0.532 and external nursery performance from 0.521 to 0.472. Removing the treatment-response gate reduced internal area under the precision-recall curve to 0.514. Removing the phototherapy monotonic penalty reduced it to 0.545. Removing inverse observation weighting had a small effect internally but a larger effect temporally, reducing temporal area under the precision-recall curve from 0.538 to 0.503. Removing the contrastive term reduced internal area under the precision-recall curve to 0.548.

Using global conformal calibration instead of nursery-stratified calibration preserved overall coverage but produced uneven subgroup behavior. Overall coverage was 91.1%, but coverage in the step-down unit was 86.7%, and coverage in the community nursery was 94.8%. Nursery-stratified calibration reduced this spread, with coverage ranging from 90.1% to 93.0%. The price was a slightly higher ambiguous set rate. This illustrates a common coverage-efficiency trade-off. More localized calibration protects higher-risk settings but expands uncertainty where calibration sample size is limited.

The observed-follow-up-only sensitivity analysis produced similar discrimination but slightly lower apparent rebound prevalence. In the subset with documented follow-up bilirubin or follow-up notes, the proposed model reached area under the precision-recall curve of 0.581 and high-risk missed rebound of 5.8%. The transformer reached 0.501 and 8.7%. The similarity suggests that inverse observation weighting did not create the main performance advantage. It mainly improved calibration under temporal observation changes.

The serum-only robustness analysis showed a modest decrease. Excluding transcutaneous bilirubin channels reduced the proposed model's area under the precision-recall curve from 0.566 to 0.544. The decrease was larger in the external nursery, where transcutaneous screening was more common. The delayed-laboratory simulation had a larger effect. Removing final bilirubin values obtained within two hours before cessation reduced area under the precision-recall curve to 0.492 and increased ambiguous conformal sets to 27.8%. This test shows that the model depends meaningfully on recent bilirubin information, which is clinically reasonable. Without a near-cessation value, uncertainty rises.

Subgroup analysis showed strongest performance in term infants without hemolysis indicators and weakest performance in late-preterm infants with incomplete laboratory data. For term infants, area under the precision-recall curve was 0.534, but because prevalence was low, high-risk missed rebound was only 4.9%. For late-preterm infants, area under the precision-recall curve was 0.603, reflecting higher prevalence and stronger signal, but missed rebound was 8.2% because some cases rose quickly after cessation. For infants with positive hemolysis indicators, the model achieved 91.2% sensitivity and 65.3% specificity. For infants with missing hemolysis testing, sensitivity was 86.4% and specificity 71.8%.

The model's recommendations would increase observation in some groups. Compared with documented usual care, the proposed conformal policy recommended additional review for 14.7 more infants per 100 treated courses. Most additional recommendations occurred in late-preterm infants, infants with feeding concerns, and infants whose final bilirubin was within 1.5 mg/dL of the risk-adjusted threshold. The policy would reduce missed high-risk rebound by 6.7 cases per 1,000 treated courses in the internal test. Whether this exchange is acceptable depends on nursery capacity and local follow-up infrastructure.

6. Latent State Analysis and Error Patterns

The learned latent state contained a bilirubin burden direction that tracked observed bilirubin values but did not merely reproduce them. The correlation between the burden coordinate

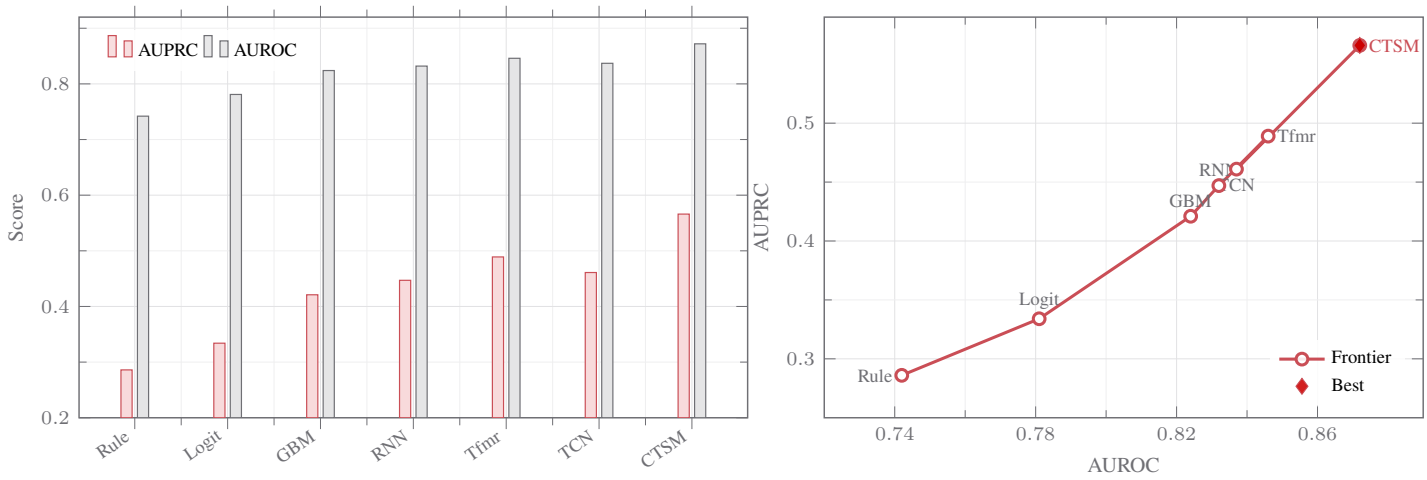


Figure 7 Internal discrimination.

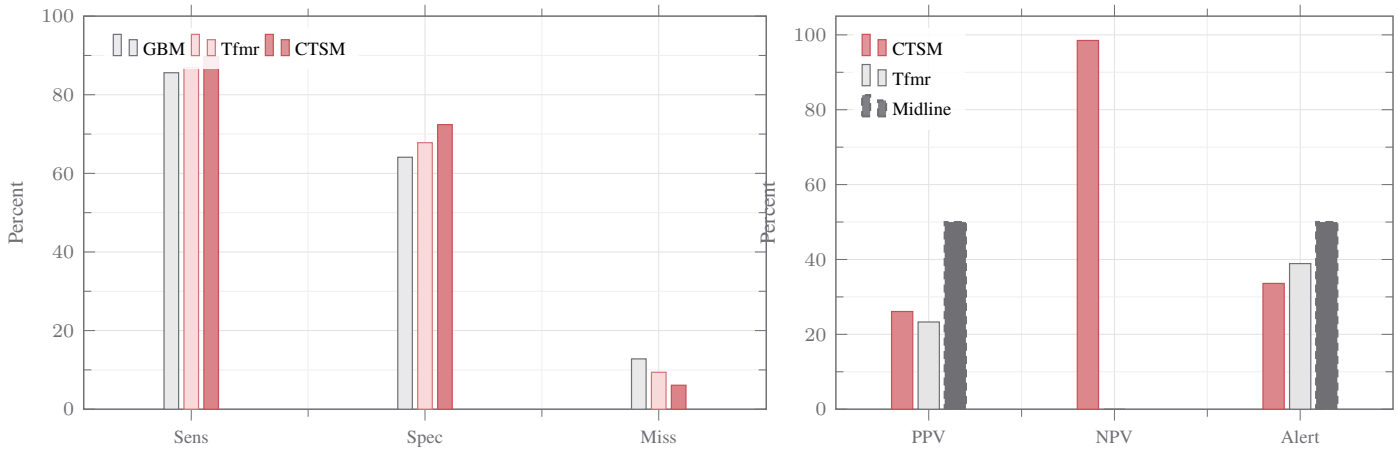


Figure 8 Operating threshold.

and the last pre-cessation bilirubin distance below threshold was -0.68. The coordinate also increased with hemolysis indicators, earlier postnatal age, and weak treatment decline. Infants with rebound had higher burden coordinate at cessation than non-rebound infants, with standardized mean difference 0.74. Among infants with similar final bilirubin distance, the burden coordinate separated those with rapid phototherapy decline from those with persistent slow decline. This suggests that the state captured trajectory shape rather than only the final value.

The treatment-response gate showed clinically plausible variation. The average gate value during phototherapy was 0.61 for non-rebound courses and 0.43 for rebound courses. Infants with intensive phototherapy and strong bilirubin decline had high gate values. Infants with hemolysis indicators often had lower gate values even when bilirubin decreased, implying persistent production pressure. Late-preterm infants had slightly lower gate values after adjustment for bilirubin level, consistent with slower clearance or feeding immaturity. These interpretations are cautious because the model is statistical, but the patterns were stable across bootstrap samples.

A two-dimensional projection of final latent states showed four recurring regions [20]. One region contained term infants with low final bilirubin and no rebound. A second contained infants with high final bilirubin near threshold but adequate treatment response; many received ambiguous conformal sets. A third contained hemolysis-associated rebound cases with early cessation age. A fourth contained feeding-associated rebound cases with weight loss and intake concerns but less strong hemolysis evidence. The model's errors were concentrated near boundaries between the second, third, and fourth regions.

Missed rebound cases were reviewed. Among 118 rebound cases missed by the binary proposed model in the internal test, 31 had incomplete follow-up data until return, 27 had bilirubin values more than 2.0 mg/dL below threshold at cessation but rose rapidly, 22 had feeding deterioration documented only after discharge, 17 had missing hemolysis testing, 11 had transcutaneous-to-serum measurement disagreement, and 10 had record-timing inconsistencies. High-risk missed rebound was rare but clinically important. Those cases often involved rapid post-cessation rise in infants whose treatment response appeared adequate. This indicates that even time-resolved models may fail when a latent production process changes after cessation.

False alerts were more common than misses. Among 420

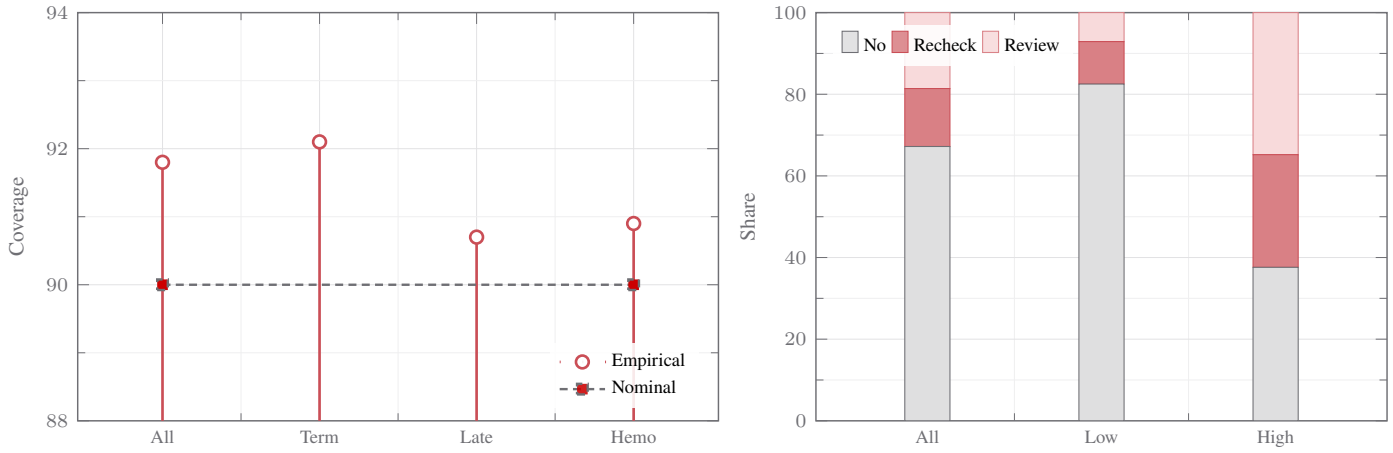


Figure 9 Conformal sets.

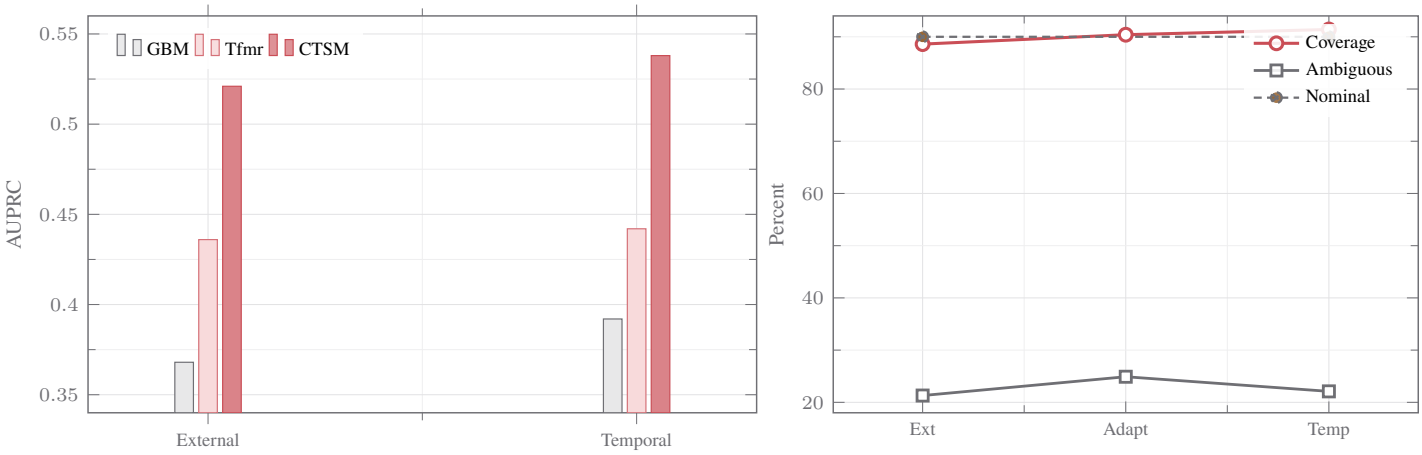


Figure 10 Shift response.

false alert cases reviewed, 143 involved bilirubin near threshold but stable or declining after discharge, 91 involved late-preterm infants kept in an ambiguous set because of calibration uncertainty, 76 involved feeding concerns that improved quickly, 54 involved hemolysis indicators without rebound, 32 involved missing recent serum values, and 24 involved documentation inconsistencies. These false alerts are not all equivalent. Some would be reasonable candidates for follow-up despite no observed rebound. Others represent unnecessary caution. The model cannot fully distinguish these without more precise post-discharge feeding and follow-up data.

Counterfactual perturbations produced expected score movement. Increasing the final bilirubin distance below threshold by 1.0 mg/dL reduced predicted rebound probability by a median of 0.071. Changing hemolysis indicator from present to absent reduced probability by 0.049. Increasing weight loss from 5.0% to 9.0% increased probability by 0.038. Removing the final pre-cessation bilirubin value increased conformal ambiguity rather than consistently increasing rebound probability. This behavior is useful because missing a recent value should make the model less certain, not automatically classify the infant as rebound positive.

The low-rank transition improved temporal smoothness. The full transition ablation produced sharper latent jumps and higher apparent training discrimination but worse external performance. In external nursery testing, its calibration slope was 0.71, compared with 0.91 for the low-rank model. This indicates overconfident transport of internal timing patterns. The low-rank model constrained temporal interactions enough to generalize better across measurement practices. The result supports the use of structured linear algebraic constraints when clinical sequences are irregular and moderate in length.

The contrastive term shaped neighborhoods in latent space [21]. Without it, infants with similar outcomes but different measurement frequency separated by nursery practice. With it, nearest neighbors more often shared rebound status and gestational-age category rather than nursery setting [22]. In the internal test set, the proportion of the ten nearest latent neighbors sharing outcome status increased from 71.2% without contrastive training to 78.5% with it. The proportion sharing nursery setting decreased from 62.9% to 51.4%. This shift did not remove all nursery signal, but it reduced representation clustering by documentation pattern.

The observation model revealed selective follow-up. Infants

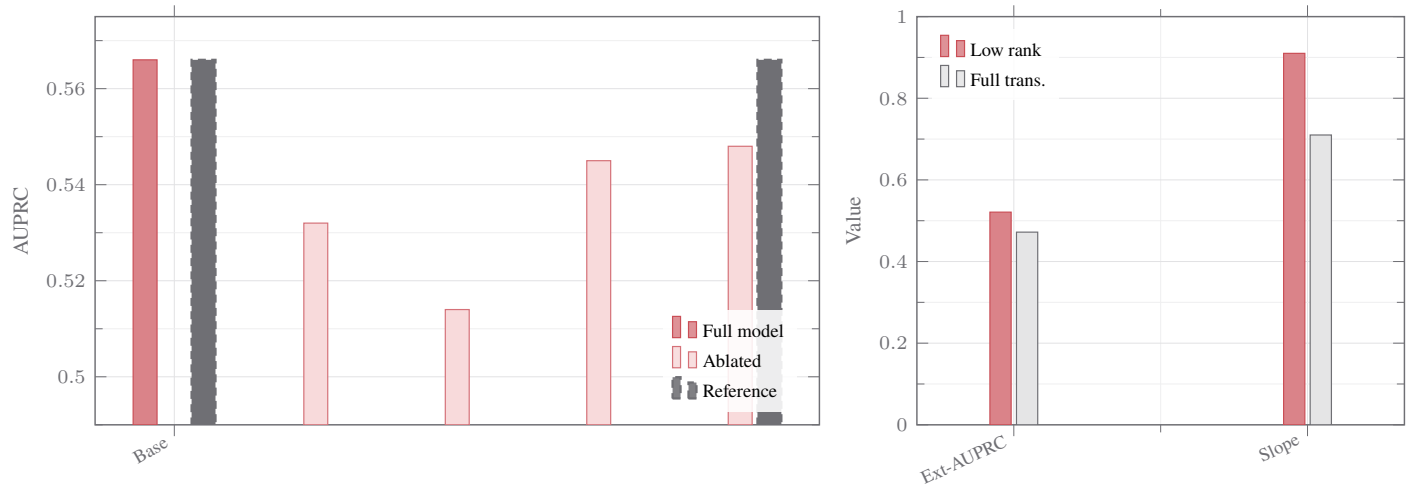


Figure 11 Ablation response.

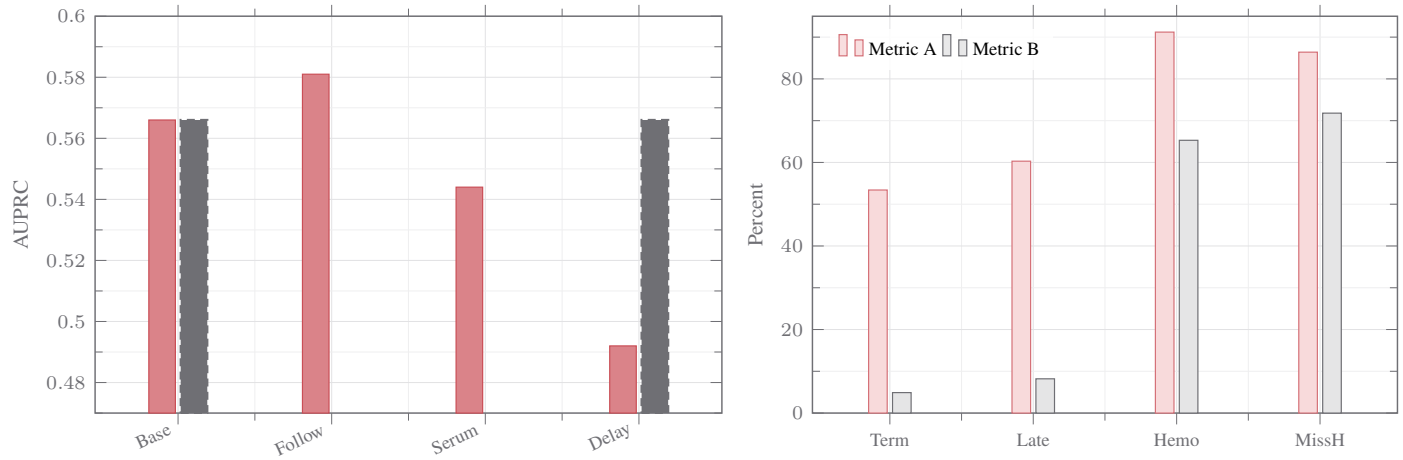


Figure 12 Robustness profile.

with final bilirubin within 1.0 mg/dL of threshold had 2.8 times higher odds of observed follow-up than infants more than 3.0 mg/dL below threshold. Late-preterm infants had 1.9 times higher odds of follow-up [23]. Nursery setting changed follow-up odds even after adjustment for bilirubin distance and gestational age. Inverse observation weighting therefore addressed a real source of bias. Removing the weights made the model more optimistic for low-follow-up nurseries and less calibrated after monitoring guidance changed.

The conformal layer behaved differently from ordinary thresholding. Some infants with predicted rebound probability around 0.30 received singleton rebound sets because their calibration stratum had high uncertainty. Others with probability around 0.35 received ambiguous sets rather than singleton rebound because the non-rebound score also passed the conformal threshold. This set-valued behavior may appear unfamiliar compared with binary thresholds, but it better reflects uncertainty in borderline clinical contexts. The model is not only saying whether rebound is probable; it is saying whether the available evidence supports excluding rebound at the desired error level.

Ambiguous conformal sets were concentrated in clinically recognizable situations. The ambiguous rate was 11.2% for term infants with no hemolysis indicators, 29.4% for late-preterm infants, 33.8% for infants with hemolysis indicators, and 31.1% when the final bilirubin value was missing or transcutaneous only. This concentration is desirable if ambiguity corresponds to cases where caution is reasonable. However, high ambiguous rates can strain workflow. When the nominal miscoverage level was relaxed from 10.0% to 15.0%, the ambiguous rate fell from 18.6% to 12.9%, but missed high-risk rebound among singleton no-rebound recommendations increased from 0.3% to 0.7%. The choice of α should therefore be local.

External nursery errors showed the effect of measurement modality. The external nursery used transcutaneous screening more heavily after phototherapy cessation. Transcutaneous values near thresholds introduced uncertainty because their relationship to serum values after phototherapy is imperfect. The model often assigned ambiguous sets in these cases. When analysis was restricted to external nursery courses with serum values near cessation, area under the precision-recall curve increased from 0.521 to 0.557 and ambiguous rate decreased from 24.9% to 19.8%. This does not mean serum testing should

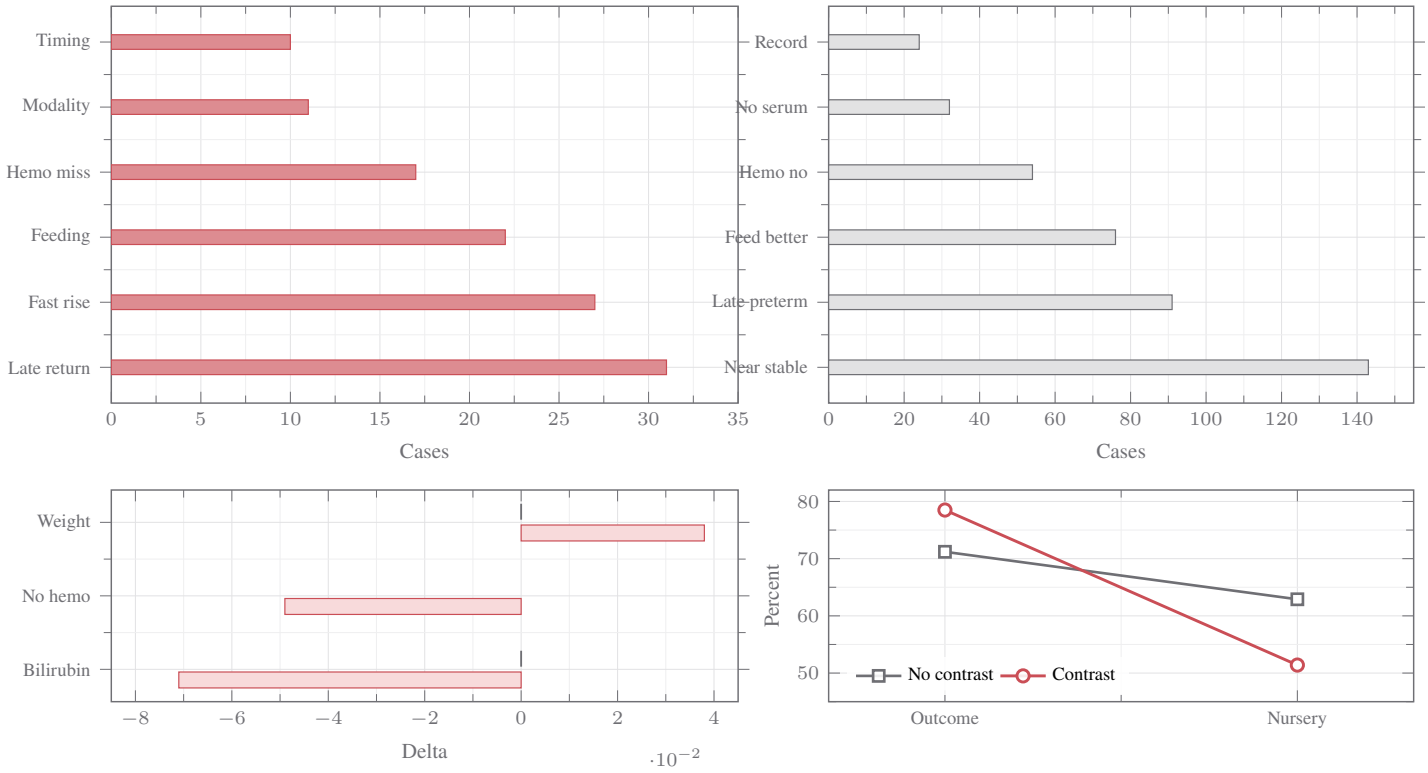


Figure 13 Error topology.

be universal; it shows that the model appropriately reflects uncertainty from modality differences.

The temporal monitoring change produced a label-observation problem. Fewer low-risk infants had routine rebound bilirubin checks, so some unobserved mild rebounds may have been missed. The inverse weighting model estimated lower observation probability after the guidance change and increased weights for low-risk infants with follow-up. This helped calibration but could not fully recover unmeasured outcomes. A sensitivity analysis assuming that 2.0% of unobserved low-risk infants had unrecorded rebound reduced estimated specificity slightly but did not change the ordering of models. Stronger assumptions changed absolute performance but still favored the tensor state model.

The model was least reliable for infants discharged very early after cessation. When discharge occurred within six hours after phototherapy stopped, the missed rebound rate among low-risk recommendations was 2.0%, compared with 0.9% when observation continued longer than six hours. This is expected because early discharge reduces the information available about immediate rebound. The conformal layer partially compensated by increasing ambiguous recommendations in early-discharge cases. Still, the finding suggests that models should be linked to actual follow-up feasibility. A low-risk prediction is less reassuring if the infant cannot obtain timely outpatient reassessment.

7. Conclusion

This paper presented a conformal tensor state model for predicting neonatal bilirubin rebound within 24 hours after phototherapy cessation. The method combines time-indexed clinical tensors, low-rank latent dynamics, treatment-response gating, inverse observation weighting, and split conformal prediction. In a multi-nursery cohort, it improved rebound discrimination over clinical rules, gradient boosting, recurrent models, and transformer encoders. It also reduced missed high-risk rebound while maintaining conformal coverage under internal, external nursery, and temporal testing.

The main result is not that rebound can be perfectly predicted. It cannot be, given irregular follow-up, changing feeding status, measurement variability, and incomplete hemolysis information. The more useful result is that structured time modeling improved risk separation and allowed uncertainty to be expressed as a set-valued recommendation [24]. The conformal layer made the model more cautious in groups where errors were more likely, especially late-preterm infants, hemolysis-associated cases, and courses without recent serum measurements. This caution increased observation recommendations but reduced missed high-risk events.

The latent state analysis suggested that the model learned clinically recognizable patterns. A bilirubin burden direction reflected more than the final measured value. Treatment-response gates were lower in rebound courses and in infants with continuing production risk. Ambiguous sets clustered in settings where clinicians would often hesitate. These findings support the plausibility of the model, although they do not prove biological mechanism. The model remains a statistical decision-support tool trained on historical records.

Several limitations should guide interpretation. The outcome depends partly on whether follow-up bilirubin values were ob-

tained, and observation practices varied by nursery and calendar period. Inverse observation weighting reduces this bias but cannot remove it completely. Some variables, including feeding adequacy and bruising, were extracted from documentation and may be inconsistently recorded. The external nursery test showed lower efficiency, indicating that local calibration matters. The model also performed worse when recent bilirubin values were unavailable, which may limit usefulness in settings with delayed laboratory turnaround. The proposed policy would identify more infants at risk for rebound but would recommend additional review for some infants who would not rebound. Whether this is acceptable depends on nursery capacity, outpatient follow-up access, parental travel burden, and local tolerance for missed rebound. The model should not be used as a standalone discharge authority. It is better suited to highlighting infants for closer clinical review and supporting standardized follow-up planning.

Future work should test the approach prospectively, with clear measurement protocols after phototherapy cessation. Additional data on feeding volume, stooling, lactation support, outpatient follow-up completion, and serum-transcutaneous differences could improve performance. Local conformal calibration should be monitored after guideline or workflow changes. A useful next step is a silent trial in which model recommendations are recorded but do not affect care, followed by review of missed rebound and alert burden. Only after such evaluation could the model's role in discharge planning be judged responsibly.

Acknowledgments

We would like to thank the reviewers of this research for their helpful comments and suggestions.

References

[1] N. Chetla, S. S. Samayamanthula, J. H. Chang, A. Y. Leigh, S. Akosman, M. Tandon, T. R. Hage, and M. Cusick, "Assessing the diagnostic capabilities of chatgpt-4 omni in grading diabetic retinopathy fundoscopy using color fundus photographs.," *Clinical ophthalmology (Auckland, N.Z.)*, vol. 19, pp. 3103–3112, 8 2025.

[2] B. Sadanandan and V. Behzadan, "Psf-med: Measuring and explaining paraphrase sensitivity in medical vision language models," *arXiv preprint arXiv:2602.21428*, 2026.

[3] W. Zhou, Z. Shi, B. Xie, F. Li, J. Yin, Y. Zhang, L. Hu, L. Li, Y. Yan, X. Wei, Z. Hu, Z. Luo, W. Peng, X. Xie, and X. Long, "Smf-net: semantic-guided multimodal fusion network for precise pancreatic tumor segmentation in medical ct image.," *Frontiers in oncology*, vol. 15, pp. 1622426–1622426, 7 2025.

[4] D. Badar, J. Abbas, R. Alsini, T. Abbas, W. ChengLiang, and A. Daud, "Transformer attention fusion for fine grained medical image classification.," *Scientific reports*, vol. 15, pp. 20655–20655, 7 2025.

[5] A. Ahmed, G. Sun, A. Bilal, Y. Li, and S. A. Ebad, "Precision and efficiency in skin cancer segmentation through a dual encoder deep learning model.," *Scientific reports*, vol. 15, pp. 4815–4815, 2 2025.

[6] S.-E. Choi, S. Mukherjee, L. E. Gibbons, E. H. Trittschuh, M. Lee, P. Scollard, R. E. Sanders, B. E. Snitz, C. E. Shaaban, O. L. Lopez, J. Mez, A. J. Saykin, T. J. Hohman, and P. K. Crane, "Beyond nacc's uniform data set for cognition: the impact of additional items on measurement precision.," *Alzheimer's & dementia : the journal of the Alzheimer's Association*, vol. 21, pp. e70694–e70694, 9 2025.

[7] N. D. Modi, B. D. Menz, A. A. Awaty, C. A. Alex, J. M. Logan, R. A. McKinnon, A. Rowland, S. Bacchi, K. Gradon, M. J. Sorich, and A. M. Hopkins, "Assessing the system-instruction vulnerabilities of large language models to malicious conversion into health disinformation chatbots.," *Annals of internal medicine*, vol. 178, pp. 1172–1180, 6 2025.

[8] F. Zhu, Z. Liu, J. Chang, Y. Qin, and L. Wang, "Deep learning for scene understanding in mitochondrial dysregulation and blood cancer diagnosis.," *Frontiers in oncology*, vol. 15, pp. 1609851–1609851, 10 2025.

[9] T. Girolamo, A. Escobedo, L. Butler, C. A. Larson, I. Campos, and K. Greene-Pendelton, "The role of social drivers of health in communication abilities of autistic adolescents and young adults.," *Autism : the international journal of research and practice*, vol. 30, pp. 329–345, 11 2025.

[10] I. N. Chan, P. K. Wong, T. Yan, Y.-Y. Hu, C. I. Chan, Y.-Y. Qin, C. H. Wong, I. W. Chan, I. H. Lam, S. H. Wong, Z. Li, S. Gao, H. H. Yu, L. Yao, B.-L. Zhao, and Y. Hu, "Assessing deep learning models for multi-class upper endoscopic disease segmentation: A comprehensive comparative study.," *World journal of gastroenterology*, vol. 31, pp. 111184–111184, 11 2025.

[11] M. Farsi, H. ZainEldin, H. A. Sayed, R. F. El-Agamy, E.-S. Atlam, S. A. Alsaedi, M. Alwateer, H. M. Balaha, M. Badawy, and M. A. Elhosseini, "Deep learning for pathology: Yolov8 with eigencam for reliable colorectal cancer diagnostics.," *Bioengineering (Basel, Switzerland)*, vol. 12, pp. 1203–1203, 11 2025.

[12] M. Omar, R. Agbareia, M. E. Naffaa, A. Watad, B. S. Glicksberg, G. N. Nadkarni, and E. Klang, "Applications of artificial intelligence in vasculitides: A systematic review.," *ACR open rheumatology*, vol. 7, pp. e70016–e70016, 3 2025.

[13] K. H. Miao, J. H. Miao, M. Finkelstein, A. Chatterjee, and A. Oto, "Next-generation advances in prostate cancer imaging and artificial intelligence applications.," *Journal of imaging*, vol. 11, pp. 390–390, 11 2025.

[14] F. Guo, R. Guan, Y. Li, Q. Liu, X. Wang, C. Yang, and J. Wang, "Foundation models in bioinformatics.," *National science review*, vol. 12, pp. nwaf028–nwaf028, 1 2025.

[15] Y. Duan, Q. Zhang, Y. Yang, Y. Weng, T. Cai, and C. Yuan, "Functional profiling of authoritative breast cancer-related key opinion leaders based on topic distributions: A comparative cluster analysis across social media platforms.," *Digital health*, vol. 11, pp. 20552076251410043–20552076251410043, 12 2025.

[16] S. J. Price, J. G. Hughes, S. Jain, C. Kelly, I. Sederias, F. M. Cozzi, J. Fares, Y. Li, J. C. Kennedy, R. Mayrand, Q. H. W. Wong, Y. Wan, and C. Li, "Precision surgery for glioblastomas.," *Journal of personalized medicine*, vol. 15, pp. 96–96, 2 2025.

[17] B. Straub, K. R. Mollan, J. R. Keys, C. Lee, R. Swanstrom, L. Premkumar, T. P. Sheahan, S. Zhou, W. A. Fischer, and D. A. Wohl, "Research protocol design and implementation of a hybrid no-touch and low-touch prospective observational study during the covid-19 public health emergency: the vision study.," *BMJ public health*, vol. 3, pp. e002796–e002796, 8 2025.

[18] J. L. Brenner, J. T. Anibal, L. A. Hazen, M. J. Song, H. B. Huth, D. Xu, S. Xu, and B. J. Wood, "Ir-gpt: Ai foundation models to optimize interventional radiology.," *Cardiovascular and interventional radiology*, vol. 48, pp. 585–592, 3 2025.

[19] K. M. Kostick-Quenet, "A caution against customized ai in healthcare.," *NPJ digital medicine*, vol. 8, pp. 13–13, 1 2025.

[20] C. Li, W. Huang, H. Yang, J. Liu, Y. Liang, and S. Wang, "Medical vision-language modeling with semantic interaction and adaptive refinement prompting for bias mitigation.," *IEEE journal of biomedical and health informatics*, vol. PP, pp. 1–9, 11 2025.

[21] V. V. R. K. Reddy, M. Villordon, Q. N. Do, Y. Xi, M. A. Lewis, C. L. Herrera, D. Owen, C. Y. Spong, D. M. Twickler, and B. Fei, "Ensemble of fine-tuned machine learning models for hysterectomy prediction in pregnant women using magnetic resonance images.," *Journal of Medical Imaging*, vol. 12, pp. 024502–024502, 3 2025.

[22] J. Chen, H. Wang, Z. Zhang, K. Xia, F. Gao, X. Xu, and G. Wang, "Development and validation of a multi-task artificial intelligence-assisted system for small bowel capsule endoscopy.," *International journal of general medicine*, vol. 18, pp. 2521–2536, 5 2025.

[23] I. L. Alter, K. Chan, K. Andreadis, and A. Rameau, "Generative artificial intelligence methodology reporting in otolaryngology: A scoping review.," *The Laryngoscope*, vol. 136, pp. 1109–1117, 9 2025.

[24] B. Peng, J. Mu, F. Xu, W. Guo, C. Sun, and W. Fan, "Artificial intelligence in ophthalmology: a bibliometric analysis of the 5-year trends in literature.," *Frontiers in medicine*, vol. 12, pp. 1580583–1580583, 7 2025.