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.

Multimodal Machine Learning Integrating Systemic and Imaging Data for ROP Surgical Triage

Jiaxin Cai

caijiaxin@xmut.edu.cn

Xiamen University of Technology

Lin Lin

Xiamen Maternal and Child Health Hospital

Bin Wang

Xiamen University of Technology

Lei Huang

Xiamen University of Technology

Research Article

Keywords: Retinopathy of prematurity (ROP), Multimodal machine learning, Surgical decision-making, Prognostic prediction, Fundus images, Cardiovascular and metabolic indicators

Posted Date: August 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10331369/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Retinopathy of prematurity (ROP) is the leading cause of irreversible childhood blindness worldwide. The selection of surgical intervention timing and prognosis assessment for ROP are highly dependent on clinicians' experience and subjective interpretation of imaging findings. At present, there is a lack of studies that can systematically integrate multimodal data (including cardiovascular system indicators, perinatal metabolic indicators, and temporal fundus image features) to predict the need for surgery and the risk of reoperation. This study aimed to develop and internally evaluate a multimodal machine learning model to improve the objectivity of ROP treatment decision support.

Methods

A total of 213 preterm infant cases were enrolled in this study, with data including cardiovascular indicators (congenital heart disease (CHD), patent ductus arteriosus (PDA), persistent pulmonary hypertension of the newborn (PPHN), the lowest hemoglobin level during hospitalization), metabolic indicators (gestational diabetes mellitus), and a series of fundus images. Four multimodal fusion models (Transformer_Fusion, TwoTower, EarlyFusion_Attention, and LateFusion_LSTM) were trained to predict the requirement for surgical or interventional treatment and prognosis among patients requiring treatment. Model performance was evaluated using indicators including accuracy, F1-score, and area under the curve (AUC), and the optimal model was selected for final evaluation on the test set.

Results

Among the 213 infants, 179 did not require surgery and 34 required surgical intervention, of whom 4 required secondary surgery. The EarlyFusion_Attention model was selected as the best-performing model for detailed test-set analysis. In the test set, the model achieved a surgical decision accuracy of 62.79% and a weighted F1 score of 0.6829. The surgical decision ROC AUC was 0.80. The model correctly identified all 6 patients who required surgery, with a sensitivity of 1.000 and a specificity of 0.568. Among the 6 patients who actually required surgery, the prognosis prediction accuracy was 83.3%, the weighted prognosis F1 score was 0.8519, and the ROC AUC was 1.00. However, the prognosis result should be interpreted cautiously because the surgical-patient subgroup contained only one Secondary case.

Conclusion

The EarlyFusion_Attention model showed the best overall performance among the four multimodal fusion models and achieved high sensitivity for identifying patients requiring surgery. However, its

specificity remained moderate, and the prognosis prediction results were limited by the small number of surgical and secondary-surgery cases. These findings suggest that multimodal machine learning may provide supportive information for ROP treatment decision-making, but external validation in larger cohorts is required.

Research Insights

What is currently known about this topic?

Artificial intelligence, especially deep learning, has shown great potential in the automatic screening, disease staging and Plus disease detection of ROP. Most studies have focused on unimodal image analysis, while multimodal studies integrating clinical indicators (such as birth weight and gestational age) with imaging data are still in the exploratory stage, and few of them involve the direct prediction of surgical intervention.

What is the key research question?

Do cardiovascular and perinatal metabolism-related clinical indicators (e.g., congenital heart disease, gestational diabetes mellitus) contain additional information for predicting the surgical need of ROP? How can temporal fundus image features be effectively fused with heterogeneous clinical data to construct a machine learning model for predicting both treatment requirement and reoperation risk?

What is new?

This study integrated cardiovascular-metabolic clinical indicators with serial fundus images to construct and compare four patient-level multimodal fusion models for ROP surgical decision-making and prognosis prediction. Among the tested architectures, EarlyFusion_Attention achieved the best overall performance. It showed high sensitivity for identifying patients requiring surgery, while the prognosis prediction result should be interpreted as exploratory because only a small number of patients required surgery and only one Secondary case was present in the test subset.

How might this study influence clinical practice?

The proposed model may serve as an auxiliary decision-support tool for ROP management by integrating clinical indicators and serial fundus images at the patient level. Its high sensitivity may help flag infants who potentially require closer monitoring or surgical evaluation. However, because specificity was moderate and prognosis prediction was limited by the small number of Secondary cases, the model should not be used as an independent clinical decision tool without further external validation.

Introduction

Retinopathy of prematurity (ROP) is a retinal vascular proliferative disease occurring in preterm and low-birth-weight infants[1]. In recent years, with the rapid advancement of global neonatal care technology, the survival rates of extremely preterm infants and extremely low-birth-weight infants have increased significantly, and the incidence of ROP has also gradually risen. Between 1990 and 2021, a total of 7.9 million cases of ROP-related visual loss were recorded (4.57 million in males and 4.22 million in females) [2], making it the leading cause of childhood blindness in both developed and developing countries worldwide[3]. Although ROP is classified as a neovascular disease, its underlying molecular mechanisms remain unclear. Current hypotheses suggest that multiple pathological risk factors,

including low birth weight, gestational age, postnatal fluctuations in blood oxygen concentration, and inflammation, promote the progression of ROP by inducing retinal vascular occlusion and neovascularization.

Current treatments for ROP mainly include intravitreal pharmacotherapy, retinal laser photocoagulation, cryotherapy, scleral buckling, and vitrectomy, depending on the stage of the disease. The effective therapeutic window for ROP is relatively narrow; once the disease progresses to an advanced stage or tractional retinal detachment occurs, even if anatomical structural damage can be partially or completely restored through methods such as vitrectomy, normal visual function cannot be recovered[4]. Therefore, accurate diagnosis of ROP and timely intervention within the therapeutic window are crucial for salvaging the visual function of infants. As an avoidable blinding eye disease, ROP is prioritized in the World Health Organization's "Vision 2020: The Right to Sight" initiative[5]. China's "14th Five-Year Plan" for Eye Health also emphasizes strengthening the screening and treatment of neonatal eye diseases, particularly ROP, to reduce the incidence and blindness rate of ROP.

Currently, ophthalmologists mainly assess the affected areas, stages, and the presence or absence of Plus disease based on the International Classification of Retinopathy of Prematurity (ICROP3) using fundus images collected by a binocular indirect ophthalmoscope or a wide-field digital retinal imaging system[6]. However, due to the complexity and variability of ROP lesions, the long training cycle for professional pediatric ophthalmologists, and the inconsistent quality of fundus images obtained during screening, diagnostic consistency among intra-observers or between different observers is affected. Additionally, the limited ability to predict disease progression may lead to delayed intervention timing or over-treatment.

With the rapid development of artificial intelligence (AI) in medical imaging, machine learning—especially deep learning—has opened new avenues for automated analysis and intelligent diagnosis of ROP[7]. Recent studies have confirmed that models based on convolutional neural networks can achieve or even exceed the level of human experts in ROP screening and severity grading[8–11]. However, most AI models rely only on single or single-modal fundus images and fail to fully utilize the abundant clinical information of infants.

To date, only a few studies have used multimodal data to predict the occurrence, progression, and prognosis of ROP. Wu Q et al. fused 46 risk factors of preterm infants with feature maps of each fundus image extracted by ResNet50, processed them through fully connected layers, and designed a two-stage deep learning framework to predict whether ROP would occur and its severity[12]. Although this study comprehensively revealed medical clues by combining risk factors and fundus images, it only used fundus images from the first screening for evaluation, which cannot fully reflect the dynamic structural changes of the retina during disease progression. Huang P et al. designed a novel bimodal deep learning framework named ROPRNet, which integrated 33 risk factors (including corrected gestational age at treatment) and fundus images to comprehensively assess the risk of recurrence after ROP treatment,

effectively improving the prediction accuracy of post-treatment recurrence. Nevertheless, this study also only adopted fundus images from a single screening[13].

The onset and progression of ROP are not isolated events; they are closely linked to the perinatal and postnatal systemic conditions of preterm infants. Multiple studies have demonstrated that gestational diabetes mellitus can disrupt the normal metabolic environment of the fetus during the perinatal period, representing an independent risk factor for the development and progression of ROP[14]. Cardiovascular disorders such as congenital heart disease (CHD)and patent ductus arteriosus（PDA）may alter hemodynamic status both systemically and within the ocular circulation[15]. Most investigations have reported a positive association between CHD and the occurrence of ROP[15, 16]. In preterm infants with gestational age below 27 weeks, hemodynamically significant patent ductus arteriosus (hsPDA) is associated with an elevated risk of moderate-to-severe ROP[17]. This is likely due to higher preductal cardiac output and increased blood oxygen saturation, which may lead to excessive oxygen delivery to the developing eye.

Hemoglobin (Hb) is responsible for transporting oxygen from the lungs to peripheral tissues, and reduced Hb levels directly impair oxygen-carrying capacity. More than 70% of placental iron transfer from mother to fetus occurs during the third trimester, a critical period that most high-risk preterm infants for ROP miss[18]. Anemia, defined by low blood Hb concentration, compromises tissue oxygenation and induces systemic and retinal hypoxia. Hypoxia can damage the function of multiple tissues, including the retina. Infants with anemia cannot meet the heightened oxygen demand of the maturing retina, resulting in delayed retinal vascularization. This delay subsequently upregulates vascular endothelial growth factor (VEGF) expression and promotes pathological neovascularization[19]. Previous studies have also confirmed that hypoxia contributes to ROP pathogenesis partly through the generation of reactive oxygen species (ROS)[20, 21]. Collectively, these clinical variables may contain critical information for predicting ROP progression and therapeutic response.

Multimodal machine learning, by integrating data from diverse sources and formats (such as images, numerical indicators, and categorical variables), is expected to achieve more comprehensive and accurate predictive capabilities than any single modality[22]. In the field of ROP, preliminary studies have attempted to combine basic indicators like gestational age and birth weight with image features[23]. However, research focusing on more specific cardiovascular-metabolic indicators and their in-depth fusion with sequential follow-up temporal image features remains scarce. Temporal images can dynamically reflect the evolutionary process of retinal blood vessels, which is crucial for predicting disease outcomes.

Therefore, this study aims to fill this research gap. We collected cardiovascular-related indicators, perinatal metabolic indicators, and a series of fundus images of preterm infants, then constructed and systematically compared four different multimodal fusion models. This study has dual objectives: first, to predict whether infants will eventually require ROP treatment or surgical intervention; second, among infants requiring intervention, to further predict whether they will require secondary treatment due to

disease recurrence or persistent progression. The results of this study are expected to provide preliminary evidence for developing patient-level multimodal decision-support models for ROP management.

Methods

A. Study Design and Prediction Tasks

This was a retrospective patient-level study based on clinical data and fundus images from preterm infants with retinopathy of prematurity. The study protocol was reviewed and approved by the Ethics Committee of Xiamen Maternal and Child Health Hospital (approval number: KY-2026-026-K02). The requirement for written informed consent was waived by the ethics committee because of the retrospective nature of the study and the use of de-identified data. All procedures were conducted in accordance with the Declaration of Helsinki.

This study developed a patient-level multimodal deep learning framework for predicting surgical decision and surgical prognosis in patients with retinopathy of prematurity (ROP). The model integrated structured clinical variables and multiple ROP fundus images from each patient. Two binary prediction tasks were defined.

The first task was surgical decision prediction, which classified each patient as either Not Needed or Needed according to whether surgical treatment was required. The second task was surgical prognosis prediction, which classified patients into Primary or Secondary categories according to whether secondary surgery was required. Because prognosis prediction is clinically meaningful only among patients who actually required surgery, the detailed prognosis evaluation was performed only in this surgical-patient subset. All model development and evaluation procedures were conducted at the patient level. Images from the same patient were kept within the same data subset to prevent patient-level data leakage across the training, validation, and test sets.

B. Clinical Variables and Outcome Definitions

Structured clinical variables were selected based on clinical prior knowledge and their relevance to ROP progression and surgical prognosis. Six clinical features were used as model inputs. Cardiovascular-related variables included congenital heart disease, patent ductus arteriosus, persistent pulmonary hypertension of the newborn, and the lowest hemoglobin value during hospitalization. Diabetes-related variables included maternal gestational diabetes mellitus and infant of diabetic mother syndrome recorded in the discharge diagnosis.

Binary clinical variables were encoded as numerical values. In the implemented pipeline, the presence of congenital heart disease, patent ductus arteriosus, persistent pulmonary hypertension of the newborn, maternal gestational diabetes mellitus, and infant of diabetic mother syndrome was encoded as 1, whereas absence was encoded as 0. The surgical decision label was defined according to the availability

of the surgical-treatment record. Patients without a surgical-treatment record were labeled as Not Needed, whereas those with a surgical-treatment record were labeled as Needed. Among patients labeled as Needed, the prognosis label was defined according to the record of secondary surgery. Patients without a secondary-surgery record were labeled as Primary, whereas patients with a secondary-surgery record were labeled as Secondary.

C. Fundus Image Preprocessing and Patient-Level Image Sampling

Each patient had a variable number of ROP fundus images. To balance computational feasibility and patient-level information preservation, a maximum of 200 images was used for each patient. If a patient had more than 200 images, 200 images were randomly selected without replacement. If a patient had 200 or fewer images, all available images were used. All fundus images were converted to RGB format and resized to 224 × 224 pixels. During training, image augmentation included random horizontal flipping with a probability of 0.3, random rotation within 5 degrees, and color jittering with brightness and contrast factors of 0.1. During validation and testing, only resizing and normalization were applied. Images were normalized using the ImageNet mean and standard deviation values: mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225].

Because different patients had different numbers of fundus images, a dynamic batching strategy was used. Within each batch, patients with fewer images than the maximum image count in that batch were padded with zero-valued image tensors. This enabled batch-wise training while preserving the variable-length patient-level image structure.

D. Data Splitting Strategy

A patient-level stratified splitting strategy was used to construct the training, validation, and test sets. Because secondary surgery cases were rare, a purely random split could lead to the absence of secondary cases in one or more subsets. Therefore, patients with secondary surgery were first identified and allocated preferentially to the training, validation, and test sets when sufficient cases were available. The remaining patients, including those who did not require surgery and those who required only primary surgery, were then split into training, validation, and test subsets. The nominal test-set proportion was 20%, and the validation set was generated from the remaining training-validation pool using a validation proportion of 25%. The number of fundus images per patient was also calculated and grouped into image-count strata: <50, 50–99, 100–149, 150–199, and ≥ 200 images. These strata were used to monitor and preserve the distribution of image counts across the data subsets. All splitting procedures were performed at the patient level. No patient contributed images to more than one data subset.

E. Multimodal Deep Learning Models

Four multimodal architectures were designed and compared: EarlyFusion_Attention, LateFusion_LSTM, Transformer_Fusion, and TwoTower. All models used the same clinical variables and fundus image inputs but differed in their multimodal fusion mechanisms.

The EarlyFusion_Attention model used feature-level fusion with an attention mechanism. Each fundus image was processed by a convolutional image encoder consisting of three convolutional blocks followed by adaptive average pooling, producing a 128-dimensional image feature. The six clinical variables were processed by a fully connected clinical encoder and projected into a 128-dimensional clinical representation. For each image, the image feature was concatenated with the clinical representation to form an image-clinical joint feature. An attention module was then used to assign a learnable weight to each image-clinical feature. The final patient-level representation was obtained by the weighted summation of these joint features. This representation was passed into two separate classification heads for surgical decision prediction and prognosis prediction.

The LateFusion_LSTM model treated multiple fundus images from the same patient as an image sequence. Each image was first encoded into a 128-dimensional feature vector using a convolutional image encoder. The sequence of image features was then processed by a bidirectional long short-term memory network. The LSTM output was averaged across the image dimension to obtain a patient-level image representation. Clinical variables were encoded independently using a multilayer perceptron. The image-sequence representation and the clinical representation were concatenated and passed through a fusion layer. Two classification heads were then used to predict surgical decision and surgical prognosis.

The Transformer_Fusion model used a self-attention-based multimodal fusion strategy. Each fundus image was encoded into a 128-dimensional feature vector. The clinical feature vector was projected into the same 128-dimensional space and used as a special clinical token. The clinical token and all image tokens were concatenated into a single sequence and processed by a two-layer Transformer encoder with four attention heads. A sinusoidal positional encoding was applied before the Transformer encoder. The output corresponding to the clinical token was used as the global patient-level multimodal representation. This representation was then passed into two task-specific classifiers.

The TwoTower model used separate encoders for image and clinical modalities. The image tower encoded each fundus image into a 128-dimensional feature vector, and the patient-level image representation was obtained by average pooling across all image features. The clinical tower encoded the six clinical variables into a 64-dimensional clinical representation. The image and clinical representations were concatenated and then used by two separate classification heads for surgical decision and prognosis prediction.

F. Model Training and Evaluation Metrics

All models were trained using a multitask learning strategy. The surgical decision task and prognosis prediction task were both optimized using weighted cross-entropy loss. Class weights were computed from the training set to reduce the influence of class imbalance. In addition, a weighted random sampler was used during training, with higher sampling weights assigned to patients requiring surgery and to patients with secondary surgery. The total training loss was defined as:

$$L = L_{decision} + 0.5 \times L_{prognosis}$$

where $L_{decision}$ denotes the weighted cross-entropy loss for surgical decision prediction and $L_{prognosis}$ denotes the weighted cross-entropy loss for prognosis prediction.

All models were trained using the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. A cosine annealing learning-rate scheduler was used with a maximum of 20 epochs. The batch size was 4. Gradient clipping was applied with a maximum norm of 1.0. Early stopping was used to reduce overfitting. The validation composite score was calculated as the mean of validation surgical decision accuracy, validation prognosis accuracy, validation surgical decision weighted F1 score, and validation prognosis weighted F1 score. If the validation composite score did not improve for five consecutive epochs, training was stopped. The model with the highest validation composite score was saved and used for final test-set evaluation.

The four models were first compared on the validation set, and the best-performing model was selected according to the validation composite score. Final performance was then evaluated on the independent test set. For the surgical decision task, the following metrics were calculated: accuracy, weighted F1 score, confusion matrix, receiver operating characteristic curve, and area under the ROC curve. The Needed class was treated as the positive class. Sensitivity, specificity, positive predictive value, and negative predictive value were derived from the confusion matrix. For model comparison, prognosis accuracy and weighted prognosis F1 score were also recalculated only among patients who actually required surgery, because prognosis classification is clinically meaningful only in this subgroup. For the prognosis task, the detailed clinical evaluation was restricted to patients who actually required surgery. The prognosis confusion matrix, accuracy, and ROC AUC were calculated in this surgical-patient subset. This restriction was used because prognosis classification is clinically meaningful only for patients who required surgical treatment.

G. Subgroup Analysis

Subgroup analyses were conducted to evaluate the stability of the best-performing model across clinically relevant subgroups. Patients were stratified according to gestational age and birth weight. Gestational age was divided into three groups: <28 weeks, 28–32 weeks, and ≥ 32 weeks. Birth weight was divided into three groups: <1000 g, 1000–1499 g, and ≥ 1500 g.

For each subgroup, surgical decision AUC was calculated using all test-set patients within that subgroup. Prognosis AUC was calculated only among patients who actually required surgery. The 95% confidence

intervals of AUC were estimated using 1000 bootstrap resampling iterations. If a subgroup did not contain both outcome classes required for AUC calculation, the corresponding AUC was reported as not available.

Results

A. Cohort Distribution and Data Splitting

A total of 213 infants were included in this study. Among them, 179 did not require surgery and 34 required surgical intervention. Of the 34 patients requiring surgery, 4 required secondary surgery. The independent test set included 43 patients, including 37 Not Needed patients and 6 Needed patients. The dataset was divided into training, validation, and test sets at the patient level. The splitting strategy was designed to preserve the distribution of surgical decision labels, prognosis labels, and patient-level image counts as much as possible. Because secondary surgery cases were rare, patients with secondary surgery were preferentially allocated to different subsets when possible. This strategy reduced the risk that the validation or test set would lack secondary surgery cases.

The number of images varied across patients. To address this heterogeneity, each patient contributed a maximum of 200 fundus images to the model. Patients with more than 200 images were randomly sampled to 200 images, whereas those with fewer images used all available images. This allowed patient-level multimodal learning while limiting the computational burden.

B. Overall Performance Comparison of the Four Multimodal Models

The four multimodal fusion models were compared on the test set. The test-set performance is summarized in Table 1.

Table 1
Test-set performance comparison of the four multimodal fusion models for surgical decision prediction and surgical-subset prognosis prediction.

Model	Decision Accuracy	Surgical-subset Prognosis Accuracy	Decision F1	Surgical-subset Prognosis F1
EarlyFusion_Attention	0.6279	0.8333	0.6829	0.8519
Transformer_Fusion	0.6279	0.6667	0.6829	0.7083
LateFusion_LSTM	0.5814	0.5000	0.6397	0.5429
TwoTower	0.2093	0.1667	0.1655	0.2381

Note

Decision metrics were calculated using all patients in the test set. Surgical-subset prognosis metrics were calculated only among patients who actually required surgery. The surgical-patient subset contained 6 patients, including 5 Primary cases and 1 Secondary case.

EarlyFusion_Attention achieved the best overall performance among the evaluated models. For surgical decision prediction, EarlyFusion_Attention achieved an accuracy of 0.6279 and a weighted F1 score of 0.6829, which were equal to those of Transformer_Fusion and higher than those of LateFusion_LSTM and TwoTower. For prognosis prediction among patients who actually required surgery, EarlyFusion_Attention achieved the highest prognosis accuracy of 0.8333 and the highest weighted prognosis F1 score of 0.8519. Transformer_Fusion ranked second in prognosis prediction, with an accuracy of 0.6667 and a weighted F1 score of 0.7083. LateFusion_LSTM and TwoTower showed lower prognosis performance in the surgical-patient subset. EarlyFusion_Attention was selected as the best-performing model based on the validation composite score, and its final performance was further assessed on the independent test set. Decision metrics were calculated using all patients in the test set, whereas prognosis metrics were calculated only among patients who actually required surgery. Because these prognosis metrics were calculated from only 6 surgical patients, they should be interpreted as exploratory model-comparison results.

C. Surgical Decision Performance of the Best-Performing Model

The surgical decision performance of the EarlyFusion_Attention model is shown in Fig. 1A and Fig. 1B. The test set included 43 patients, of whom 37 were labeled as Not Needed and 6 were labeled as Needed. According to the confusion matrix, the model correctly classified 21 of the 37 Not Needed patients and all 6 Needed patients. No Needed patient was misclassified as Not Needed. However, 16 Not Needed patients were misclassified as Needed. The overall accuracy was 0.6279. When Needed was considered the positive class, the model achieved accuracy of 0.6279, ROC AUC of 0.80, sensitivity of 1.000, specificity of 0.568, positive predictive value of 0.273, and negative predictive value of 1.000. These results indicate that the model was sensitive in identifying patients who required surgery, with no false-negative cases in the test set. However, the model produced a relatively large number of false-positive predictions, resulting in moderate specificity. Therefore, the model showed a high-sensitivity screening tendency but still required improvement in specificity. The ROC curve showed an AUC of 0.80, indicating moderate discriminative ability for distinguishing patients who required surgery from those who did not. Given that only 6 patients in the test set belonged to the Needed class, this result should be validated in a larger and more balanced cohort.

D. Prognosis Prediction among Patients Requiring Surgery

The prognosis prediction performance of EarlyFusion_Attention among patients who actually required surgery is shown in Fig. 1C and Fig. 1D. This subgroup included 6 surgical patients, of whom 5 were labeled as Primary and 1 was labeled as Secondary. The model correctly classified 4 of the 5 Primary cases and correctly identified the only Secondary case. One Primary case was misclassified as

Secondary. The prognosis prediction accuracy in this subgroup was therefore 5/6, corresponding to 0.833. The weighted prognosis F1 score was 0.8519. The ROC curve for prognosis prediction showed an AUC of 1.00 in this surgical-patient subset. However, this finding should be interpreted cautiously because the analysis was based on only 6 surgical patients and the Secondary class contained only 1 case. Therefore, the AUC of 1.00 should be regarded as an exploratory result rather than evidence of stable perfect discrimination. Further validation with more surgical patients, particularly more Secondary cases, is required.

E. Subgroup Analysis

Subgroup analysis was performed to assess the stability of the EarlyFusion_Attention model across gestational age and birth weight groups. The results are summarized in Table 2.

Table 2
Subgroup performance of the EarlyFusion_Attention model.

Subgroup	No. of patients	Surgical decision AUC (95% CI)	Prognosis AUC (95% CI)
Overall	43	0.800 (0.65–0.95)	1.000 (1.00–1.00)
Gestational age < 28 weeks	11	0.542 (0.20–0.86)	1.000 (1.00–1.00)
Gestational age 28–32 weeks	16	0.964 (0.80-1.00)	NA
Gestational age ≥ 32 weeks	16	0.933 (0.80-1.00)	NA
Birth weight < 1000 g	8	0.583 (0.16-1.00)	1.000 (1.00–1.00)
Birth weight 1000–1499 g	15	0.750 (0.36-1.00)	NA
Birth weight ≥ 1500 g	19	0.889 (0.72-1.00)	NA

Note

Surgical decision AUC was calculated using all test-set patients within each subgroup. Prognosis AUC was calculated only among patients who actually required surgery. NA indicates that AUC could not be calculated because the subgroup did not contain both prognosis classes. Prognosis AUC estimates should be interpreted cautiously because of the small number of surgical and secondary-surgery cases.

In the overall test set, the surgical decision AUC was 0.800, with a 95% confidence interval of 0.65–0.95. When stratified by gestational age, the model achieved higher AUCs in the 28–32 weeks subgroup and the ≥ 32 weeks subgroup, with AUC values of 0.964 and 0.933, respectively. In contrast, the AUC in the < 28 weeks subgroup was 0.542, indicating limited discrimination in extremely premature infants. When stratified by birth weight, the surgical decision AUC was 0.889 in the ≥ 1500 g subgroup, 0.750 in the

1000–1499 g subgroup, and 0.583 in the < 1000 g subgroup. These results suggest that the model performed more reliably in patients with relatively higher birth weight, whereas its performance was less stable in extremely low birth weight infants. For prognosis prediction, several subgroup-level AUC values were not available because the corresponding subgroups did not contain both prognosis classes among surgical patients. Although an AUC of 1.00 was observed in some subgroups, these estimates were based on very limited numbers of surgical patients and should be interpreted as exploratory findings.

Discussion

This study developed and compared four patient-level multimodal machine learning models integrating cardiovascular-metabolic clinical indicators and serial fundus images for ROP surgical decision-making and prognosis prediction. The EarlyFusion_Attention model showed the best overall performance among the evaluated architectures. In the test set, this model achieved high sensitivity for identifying patients requiring surgery, with no false-negative cases among the Needed patients. However, its specificity was moderate because of false-positive predictions. For prognosis prediction among patients who actually required surgery, the model achieved an accuracy of 0.8333, a weighted F1 score of 0.8519, and an AUC of 1.00. However, these results should be interpreted cautiously because the analysis included only 6 surgical patients and only 1 Secondary case. Therefore, the present findings should be regarded as preliminary evidence supporting the feasibility of multimodal learning for ROP decision support rather than definitive proof of stable prognosis prediction.

The present findings are broadly consistent with the current understanding that ROP progression and treatment requirement are influenced by both ocular manifestations and systemic conditions of preterm infants. Previous studies have identified gestational age and birth weight as important predictors of ROP occurrence and severity. In our subgroup analysis, the model showed lower discriminative performance in the < 28 weeks subgroup and the < 1000 g subgroup, suggesting that prediction remains challenging in extremely premature and extremely low birth weight infants. By incorporating cardiovascular and metabolic variables into the multimodal framework, this study explored whether these systemic indicators could provide complementary information for patient-level prediction. However, the present analysis did not independently quantify the association between each clinical variable and surgical demand. Future studies should include formal statistical analyses, such as univariate and multivariable regression or feature-importance analysis, to clarify the independent contribution of each clinical indicator.

The inclusion of cardiovascular-related variables in the present model is supported by previous evidence linking neonatal cardiovascular disorders with ROP development and severity. Although numerous studies have been conducted on risk factors for ROP, few models have explored the correlation between cardiovascular diseases in preterm infants and ROP. Small-scale observational studies have reported a general association between congenital heart disease and retinal vascular dilation as well as tortuosity—both manifestations that may resemble Plus disease[24, 25]. A multicenter prospective study screened 3,980 infants for ROP, among whom 528 cases (13.3%) were diagnosed with neonatal cardiovascular

diseases (including altered pulmonary blood flow, altered systemic blood flow, pulmonary hypertension, and arrhythmia). The results indicated that preterm infants with cardiovascular diseases leading to increased pulmonary blood flow had an elevated risk of developing ROP, and those complicated with persistent pulmonary hypertension of the newborn (PPHN) had a higher risk of severe ROP[26]. However, two studies from Hong Kong both suggested that congenital heart disease acts as a protective factor against ROP[27, 28]. It is hypothesized that for cyanotic heart disease with venoarterial shunting, the blood oxygen concentration in the body may decrease, which may explain its protective mechanism against ROP. Notably, a key feature of some severe ROP requiring treatment is Plus disease, characterized by dilation of small veins and tortuosity of small arteries in the posterior pole of the retina. This feature is thought to be associated with increased blood flow and potential resistance in the fibrovascular or neovascular lesions of ROP. PPHN may lead to systemic venous congestion, increased venous pressure, or vascular dilation caused by increased blood flow, thereby exacerbating Plus disease or pre-Plus disease and resulting in misdiagnosis. Another factor to consider is whether drugs used to treat PPHN and PDA may also induce retinal vascular dilation and worsen ROP. Case reports have shown that sildenafil treatment for persistent pulmonary hypertension may cause recurrence or deterioration of ROP[29]. Currently, inhaled nitric oxide (iNO) is the first-line treatment for neonatal persistent pulmonary hypertension in preterm infants. In a national study in the United States, Cho et al. pointed out that neonatal persistent pulmonary hypertension is an independent risk factor for increased ROP incidence, while the use of inhaled nitric oxide in such infants is associated with a reduced incidence of severe ROP[30]. Meanwhile, a Japanese cohort study further confirmed the above conclusion, finding that even after adjusting for severe ROP, neonatal persistent pulmonary hypertension remains an independent predictor of long-term visual impairment; moreover, when this disease coexists with bronchopulmonary dysplasia, the risk of ocular disease is further increased. However, the use of inhaled nitric oxide for post-acute phase treatment in neonates after the first week of life did not improve their neurodevelopmental or visual prognosis[31]. These findings provide biological and clinical rationale for including cardiovascular indicators in multimodal prediction models, although the present study did not separately estimate the effect size of each cardiovascular variable.

Most previous studies have focused on common ROP risk factors, such as birth weight, gestational age, and oxygen exposure, whereas maternal metabolic conditions have been less frequently incorporated into predictive models. In the present study, maternal gestational diabetes mellitus was included as one of the structured clinical variables because previous studies have reported its association with ROP development and severity. However, its independent contribution to the model prediction was not separately quantified in this study. The research team led by Tunay found that infants with a birth weight over 1500 g born to diabetic mothers had a 25-fold increased risk of developing ROP and a 6-fold increased risk of type 1 ROP, respectively[32]. Another study specifically investigated whether maternal diabetes constitutes an independent risk factor for severe ROP (beyond stage II) in neonates with a birth weight of ≤ 1500 g, confirming that maternal diabetes raises the risk of severe ROP in preterm infants by 3.47 times, with the strength of this association increasing in tandem with the severity of ROP[33]. Numerous studies have established a link between neonatal hyperglycemia and severe ROP as well as

all stages of the disease[34]. Almeida's research revealed that hyperglycemia is a major risk factor for ROP, exerting a more significant impact than other known risk factors[35]. Mechanistically, hyperglycemia promotes the development of ROP primarily by modulating the synergistic interaction between VEGF and insulin-like growth factor-1 (IGF-1)[36]. A hyperglycemic environment induces excessive VEGF protein expression in retinal Müller cells and further stimulates VEGF production by activating the protein kinase C pathway[37]. Conversely, low IGF-1 levels not only inhibit the normal development of retinal blood vessels but also promote the formation of pathological neovascularization, which leads to retinal ischemia and subsequent massive VEGF generation, ultimately triggering excessive vascular proliferation[38].

Anemia is nearly ubiquitous among infants at high risk of ROP, with 90% of extremely low birth weight (ELBW) infants—defined as those with a birth weight below 1000 grams—requiring at least one blood transfusion[39]. Blood transfusion itself may also elevate the risk of ROP development in preterm infants. A retrospective study conducted by Pai et al. examined the association between the incidence and severity of ROP in preterm infants born before 34 weeks of gestation and relevant risk factors, including Hb levels. The study identified a significant correlation between Hb levels and ROP severity: during the third week after birth, 95.8% of infants with Hb levels below 8 mg/dL, 78.9% of those with Hb levels between 8 and 10 mg/dL, and 28.1% of those with Hb levels above 10 mg/dL developed ROP of stage II or higher. In contrast, only 8.8% of infants with Hb levels above 10 mg/dL developed stage I ROP[40]. Another retrospective cohort study by Lundgren et al. evaluated the risk of ROP in extremely preterm infants born before 28 weeks of gestation. Anemia was defined as an Hb level below 110 g/L, and severe anemia as an Hb level below 80 g/L. After adjusting for birth gestational age and corrected gestational age, the analysis revealed that the duration of anemia during the first week of life was a significant risk factor for requiring ROP treatment[41]. This finding is consistent with research by Akyüz et al., who performed a retrospective analysis of blood routine examinations in 150 preterm infants. Their results indicated that the mean corpuscular hemoglobin (MCH) exhibited the most significant predictive value among all measured indicators. The association between MCH and ROP may be linked to the nitric oxide (NO) pathway: excessive NO can induce vasodilation, capillary leakage, and edema, while hemoglobin can scavenge NO through complex formation[42]. These studies support the clinical relevance of including the lowest hemoglobin value during hospitalization as a candidate predictor. Nevertheless, the present study used hemoglobin as one component of a multimodal model and did not evaluate its independent predictive effect.

From a clinical perspective, the high sensitivity observed in the surgical decision task suggests that the model may help flag infants who require closer monitoring or further specialist evaluation. However, the moderate specificity indicates that the model may also generate false-positive alerts. Therefore, the model should be positioned as an auxiliary decision-support tool rather than a replacement for ophthalmologist assessment. Its potential value may lie in supporting risk stratification, prioritizing follow-up, and prompting careful review of high-risk patients.

The present study has several strengths. First, the model was developed at the patient level by integrating structured clinical indicators with serial fundus images, which better reflects the multimodal nature of real-world ROP assessment. Second, this study addressed two clinically relevant tasks: identifying patients who may require surgical or interventional treatment and exploring prognosis prediction among those requiring surgery. Third, four different multimodal fusion architectures were implemented and compared under the same data-splitting and evaluation framework, providing a structured comparison of fusion strategies for this task.

Several limitations should be noted. First, this was a single-cohort retrospective study with a limited sample size, especially for secondary surgery. Only 4 patients in the overall cohort required secondary surgery, and the detailed prognosis evaluation in the test set included only 6 surgical patients, of whom only 1 was labeled as Secondary. Therefore, the prognosis AUC of 1.00 should be interpreted as exploratory rather than conclusive. Second, although the EarlyFusion_Attention model achieved high sensitivity for identifying patients requiring surgery, its specificity was moderate because of false-positive predictions. This suggests that the model is more suitable as a high-sensitivity screening aid than as a standalone tool for determining surgical treatment. Third, the number of clinical indicators was limited. Future studies could incorporate additional variables, such as delivery mode, assisted reproductive technology, duration of oxygen therapy, transfusion history, sepsis, postnatal weight gain, and treatment-related information. In addition, the current implementation used zero padding to handle variable numbers of fundus images, but did not incorporate an explicit padding mask in all fusion modules. Future models should include masking strategies to prevent padded image slots from contributing to patient-level representations. Fourth, the current model lacks formal interpretability analysis. Future work should incorporate explainable AI methods, such as Grad-CAM, image-level attention visualization, and clinical feature-importance analysis, to clarify the basis of model predictions. Finally, external validation in larger multicenter cohorts and prospective clinical evaluation are required before clinical application.

Conclusion

In conclusion, this study developed and compared four patient-level multimodal machine learning models integrating cardiovascular-metabolic clinical indicators and serial fundus images for ROP treatment decision support and prognosis prediction. The EarlyFusion_Attention model showed the best overall performance among the evaluated architectures and achieved high sensitivity for identifying patients requiring surgery or intervention. However, the model showed moderate specificity, and prognosis prediction was limited by the small number of surgical and secondary-surgery cases. These findings support the feasibility of multimodal learning for ROP decision support, but larger multicenter studies and external validation are required before clinical application.

Declarations

Data availability

All datasets supporting the conclusions of this work are publicly available on Baidu Netdisk. The access link is provided as follows: https://pan.baidu.com/s/1tAdf-43wEj9ixhjGbMBHgA?pwd= 75c5

Code availability

The code that supports the findings of this study is openly accessible in a GitHub repository at the following URL: https://github.com/wbin2002/rop.git.

Ethics declaration

The study protocol was reviewed and approved by the Ethics Committee of Xiamen Maternal and Child Health Hospital (approval number: KY-2026-026-K02). The requirement for written informed consent was waived by the ethics committee because of the retrospective nature of the study and the use of de-identified data. All procedures were conducted in accordance with the Declaration of Helsinki.

Competing interests

The authors declare no competing interests.

Funding

This work is supported by the Natural Science Foundation of Fujian Province (2023J05083).

Author Contribution

JC: Conceptualization, Investigation, Supervision, Writing – Original Draft, Writing – Review & Editing. LL: Resources. BW: Data Curation, Visualization. LH: Data Curation, Formal Analysis,Writing – Original Draft

Acknowledgements

N/A

References

1. Zhang Y, Chai X, Fan Z, et al. Research hotspots and trends in retinopathy of prematurity from 2003 to 2022: a bibliometric analysis. Front Pediatr. 2023;11:1273413. 10.3389/fped.2023.1273413.
2. Wong ES, Choy RW, Zhang Y, et al. Global and Regional Trends in Retinopathy of Prematurity. JAMA Ophthalmol. 2026;144(2):139–47. 10.1001/jamaophthalmol.2025.5103.
3. Sabri K, Ells AL, Lee EY, et al. Retinopathy of Prematurity: A Global Perspective and Recent Developments. Pediatrics. 2022;150(3). 10.1542/peds.2021-053924.

4. Sen P, Wu WC, Chandra P, et al. Retinopathy of prematurity treatment: Asian perspectives. Eye (Lond). 2020;34(4):632–42. 10.1038/s41433-019-0643-4.
5. Gilbert C. A Foster Childhood blindness in the context of VISION 2020–the right to sight. Bull World Health Organ. 2001;79(3):227–32.
6. Chiang MF, Quinn GE, Fielder AR, et al. International Classification of Retinopathy of Prematurity, Third Edition. Ophthalmology. 2021;128(10):e51–68. 10.1016/j.ophtha.2021.05.031.
7. Derks LA, Tekin YS, Loudon SE, et al. External validation of an artificial intelligence-based model for retinopathy of prematurity screening using Phoenix ICON retinal images. J aapos. 2025;29(6):104696. 10.1016/j.jaapos.2025.104696.
8. Gao J, Fang N, Xu Y. Application of Artificial Intelligence in Retinopathy of Prematurity From 2010 to 2023: A Bibliometric Analysis. Health Sci Rep. 2025;8(4):e70718. 10.1002/hsr2.70718.
9. Shoeibi N, Abrishami M, Hosseini SM, et al. Development and validation of machine learning classifiers for predicting treatment-needed retinopathy of prematurity. BMC Med Inf Decis Mak. 2025;25(1):221. 10.1186/s12911-025-03057-w.
10. Wang J, Ji J, Zhang M, et al. Automated Explainable Multidimensional Deep Learning Platform of Retinal Images for Retinopathy of Prematurity Screening. JAMA Netw Open. 2021;4(5):e218758. 10.1001/jamanetworkopen.2021.8758.
11. Belenje A, Pandya D, Jalali S, et al. Accuracy of Artificial Intelligence Versus Clinicians in Real-Life Case Scenarios of Retinopathy of Prematurity. Cureus. 2025;17(2):e78597. 10.7759/cureus.78597.
12. Wu Q, Hu Y, Mo Z, et al. Development and Validation of a Deep Learning Model to Predict the Occurrence and Severity of Retinopathy of Prematurity. JAMA Netw Open. 2022;5(6):e2217447. 10.1001/jamanetworkopen.2022.17447.
13. Huang P, Xie Y, Wu R, et al. ROPRNet: Deep learning-assisted recurrence prediction for retinopathy of prematurity. Biomed Signal Process Control. 2025;100:107135. https://doi.org/10.1016/j.bspc.2024.107135.
14. Rashidian P, Karami S, Salehi SA. A review on retinopathy of prematurity. Med Hypothesis Discov Innov Ophthalmol. 2024;13(4):201–12. 10.51329/mehdiophthal1511.
15. Wu T, Zhang L, Tong Y, et al. Retinopathy of Prematurity Among Very Low-Birth-Weight Infants in China: Incidence and Perinatal Risk Factors. Invest Ophthalmol Vis Sci. 2018;59(2):757–63. 10.1167/iovs.17-23158.
16. Chesi E, Rossi K, Ancora G, et al. Patent ductus arteriosus (also non-hemodynamically significant) correlates with poor outcomes in very low birth weight infants. A multicenter cohort study. PLoS ONE. 2024;19(7):e0306769. 10.1371/journal.pone.0306769.
17. Ford A, Beauchene M, Stanford AH, et al. Exposure to persistent hemodynamically significant patent ductus arteriosus is associated with retinopathy of prematurity. J aapos. 2024;28(3):103923. 10.1016/j.jaapos.2024.103923.
18. O'Brien KO. Maternal, fetal and placental regulation of placental iron trafficking. Placenta. 2022;125:47–53. 10.1016/j.placenta.2021.12.018.

19. Tang W, Zhang Y, Zhang H, et al. Progress in the study of association between hematological indicators and retinopathy of prematurity (Review). Biomed Rep. 2024;21(2):111. 10.3892/br.2024.1799.
20. Tandon M, Ranjan R, Muralidharan U, et al. Influence of Anaemia on Multifactorial Disease Retinopathy of Prematurity: A Prospective Observational Study. Cureus. 2022;14(8):e27877. 10.7759/cureus.27877.
21. Gudu RK, Sahoo S, Jena P, et al. Lower Hemoglobin Levels as a Risk Factor for the Development of Retinopathy of Prematurity. Cureus. 2024;16(7):e64264. 10.7759/cureus.64264.
22. Ma R, Cheng Q, Yao J, et al. Multimodal machine learning enables AI chatbot to diagnose ophthalmic diseases and provide high-quality medical responses. NPJ Digit Med. 2025;8(1):64. 10.1038/s41746-025-01461-0.
23. Durmaz Engin C, Ozturk T, Ozkan O, et al. Prediction of retinopathy of prematurity development and treatment need with machine learning models. BMC Ophthalmol. 2025;25(1):194. 10.1186/s12886-025-04025-8.
24. Cordina R, Leaney J, Golzan M, et al. Ophthalmological consequences of cyanotic congenital heart disease: vascular parameters and nerve fibre layer. Clin Exp Ophthalmol. 2015;43(2):115–23. 10.1111/ceo.12401.
25. Arriola-Lopez AE, Martinez-Perez ME, Martinez-Castellanos MA. Retinal vascular changes in preterm infants: heart and lung diseases and plus disease. J aapos. 2017;21(6):488–e4911. 10.1016/j.jaapos.2017.08.004.
26. Bhatti F, Yu Y, Ying GS, et al. Association of Cardiovascular Disease with Retinopathy of Prematurity. Ophthalmic Epidemiol. 2023;30(1):95–102. 10.1080/09286586.2022.2036766.
27. Yau GS, Lee JW, Tam VT, et al. Incidence and risk factors for retinopathy of prematurity in extreme low birth weight Chinese infants. Int Ophthalmol. 2015;35(3):365–73. 10.1007/s10792-014-9956-2.
28. Yau GSK, Lee JWY, Tam VTY, et al. Risk factors for retinopathy of prematurity in extremely preterm Chinese infants. Med (Baltim). 2014;93(28):e314. 10.1097/md.0000000000000314.
29. Jayadev C, Ramasastry P, Gul A, et al. Possible Role of Sildenafil Citrate in the Recurrence of Neovascularization in Laser-regressed Aggressive Posterior ROP. Indian Pediatr. 2016;53(Suppl 2):S155–6.
30. Cho S, Farghaly MAA, Ramey S, et al. Pulmonary hypertension, inhaled nitric oxide, and retinopathy of prematurity: evidence from the U.S. national database. Pediatr Res. 2025. 10.1038/s41390-025-04323-3.
31. Nakanishi H. From pulmonary to ocular protection: rethinking inhaled nitric oxide in preterm infants with pulmonary hypertension. Pediatr Res. 2025. 10.1038/s41390-025-04532-w.
32. Tunay Z, Özdemir Ö, Acar DE, et al. Maternal Diabetes as an Independent Risk Factor for Retinopathy of Prematurity in Infants With Birth Weight of 1500 g or More. Am J Ophthalmol. 2016;168:201–6. 10.1016/j.ajo.2016.05.022.

33. Opara CN, Akintorin M, Byrd A, et al. Maternal diabetes mellitus as an independent risk factor for clinically significant retinopathy of prematurity severity in neonates less than 1500g. PLoS ONE. 2020;15(8):e0236639. 10.1371/journal.pone.0236639.
34. Lei C, Duan J, Ge G, et al. Association between neonatal hyperglycemia and retinopathy of prematurity: a meta-analysis. Eur J Pediatr. 2021;180(12):3433–42. 10.1007/s00431-021-04140-w.
35. Almeida AC, Silva GA, Santini G, et al. Correlation between hyperglycemia and glycated albumin with retinopathy of prematurity. Sci Rep. 2021;11(1):22321. 10.1038/s41598-021-01861-8.
36. Esmail J, Sakaria RP, Dhanireddy Early R. Hyperglycemia Is Associated with Increased Incidence of Severe Retinopathy of Prematurity in Extremely Low Birth Weight Infants. Am J Perinatol. 2024;41:01. 10.1055/a-2173-8360.
37. Kim NH, Jung HH, Cha DR, et al. Expression of vascular endothelial growth factor in response to high glucose in rat mesangial cells. J Endocrinol. 2000;165(3):617–24. 10.1677/joe.0.1650617.
38. Cakir B, Hellström W, Tomita Y, et al. IGF1, serum glucose, and retinopathy of prematurity in extremely preterm infants. JCI Insight. 2020;5(19). 10.1172/jci.insight.140363.
39. Maier RF, Sonntag J, Walka MM, et al. Changing practices of red blood cell transfusions in infants with birth weights less than 1000 g. J Pediatr. 2000;136(2):220–4. 10.1016/s0022-3476(00)70105-3.
40. Pai HS, Joy R, Cherian V, et al. Anemia in relation to severity of retinopathy of prematurity in preterm babies born in tertiary care centre in South India. Int J Contemp Pediatr. 2020;7(10):2005–9. 10.18203/2349-3291.ijcp20204043.
41. Lundgren P, Athikarisamy SE, Patole S, et al. Duration of anaemia during the first week of life is an independent risk factor for retinopathy of prematurity. Acta Paediatr. 2018;107(5):759–66. 10.1111/apa.14187.
42. Akyüz Ünsal A, Key Ö, Güler D, et al. Can Complete Blood Count Parameters Predict Retinopathy of Prematurity? Turk J Ophthalmol. 2020;50(2):87–93. 10.4274/tjo.galenos.2019.45313.

Figures

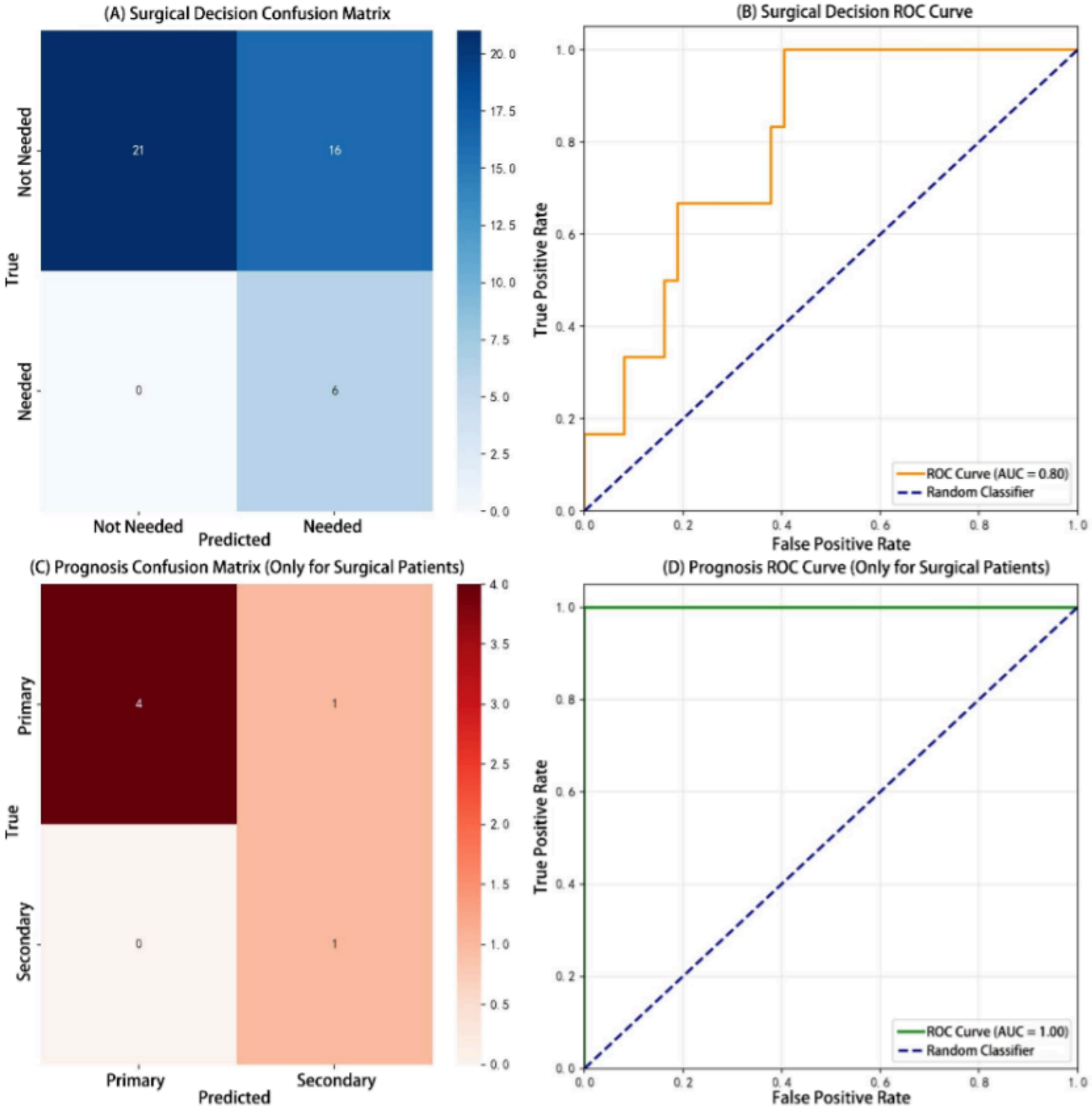


Figure 1

Test-set performance of the EarlyFusion_Attention model. (A) Confusion matrix for surgical decision prediction using the full test set. (B) Receiver operating characteristic curve for surgical decision prediction using the full test set. (C) Confusion matrix for prognosis prediction among patients who actually required surgery. (D) Receiver operating characteristic curve for prognosis prediction among patients who actually required surgery.