BMC Ophthalmology

RESEARCH

Open Access

Prediction of retinopathy of prematurity development and treatment need with machine learning models

Ceren Durmaz Engin[1,2*], Taylan Ozturk[3], Ozlem Ozkan[4], Ali Oztas[5], Mustafa Alper Selver[2,6] and Funda Tuzun[7]

Abstract

Background To evaluate the effectiveness of machine learning (ML) models in predicting the occurrence of retinopathy of prematurity (ROP) and treatment need.

Methods Four ML models were created using 49 parameters known within the first 24 h post-birth and obtained during the initial screening examination, encompassing demographic, maternal, clinical, and neonatal intensive care unit-related data. The models' performances were assessed using five machine learning (ML) classifier algorithms: logistic regression (LR), decision tree (DT), support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost). Performance metrics were calculated, and the top ten parameters with the highest predictive value were identified.

Results In the cohort of 355 preterm infants, Model I, predicting ROP development using birth data, achieved a balanced accuracy of 80%, with gestational age (GA), birth weight (BW) and mean corpuscular volume (MCV) as the top predictive parameters. Model II, predicting treatment-requiring ROP using birth data, exhibited a balanced accuracy of 81%. Key predictive parameters included low GA, BW, 1-minute and 5-minute APGAR scores, and low erythrocyte counts. For Model III, predicting ROP using the first screening examination data, and Model IV, predicting treatment-requiring ROP using the same data, the accuracy values were 80% and 66%, respectively, with BW, daily weight gain, total O2 support duration, and platelet/lymphocyte ratio emerged as the most significant predictive parameters in both models.

Conclusion This study demonstrates the potential of ML models to predict ROP development and treatment need. Incorporating clinical and intensive care-related parameters can enhance ROP screening and clinical decision-making.

Keywords Artificial intelligence, Treatment, Machine learning, Retinopathy of prematurity

*Correspondence:
Ceren Durmaz Engin
cerendurmaz@gmail.com
[1]Department of Ophthalmology, Izmir Democracy University Buca Seyfi Demirsoy Education and Research Hospital, Kozagac Mah, Ozmen Cad No:147 Buca, Izmir, Turkey
[2]Department of Biomedical Technologies, Dokuz Eylul University, Izmir, Turkey
[3]Department of Ophthalmology, Tinaztepe University, Izmir, Turkey
[4]Department of Ophthalmology, Dokuz Eylul University School of Medicine, Izmir, Turkey
[5]EPAM Systems, Izmir, Turkey
[6]Department of Electrical and Electronics Engineering, Dokuz Eylul University, Izmir, Turkey
[7]Department of Neonatology, Dokuz Eylul University School of Medicine, Izmir, Turkey

Background

Retinopathy of Prematurity (ROP) is a prevalent condition among preterm infants and remains a leading preventable cause of childhood blindness worldwide [1, 2]. In 2010 alone, approximately 184,700 preterm infants were diagnosed with ROP, with an estimated 20,000 cases resulting in blindness [3]. The incidence of severe ROP varies significantly based on birth weight (BW) and gestational age (GA). Recent data from the United States indicate that 2.4% of infants weighing above 2,500 g and 30.2% of those between 750 and 999 g exhibit severe ROP [4]. Among neonates weighing less than 1,500 g or born before 32 weeks, approximately 12.5% develop severe ROP. As neonatal care advances and survival rates increase, particularly for extremely preterm infants, the need for efficient ROP screening and timely intervention becomes even more critical.

ROP screening remains a major challenge, even in high-income countries, as it requires frequent examinations—weekly or biweekly—until the retinal vasculature matures. This process not only demands a skilled workforce but also exposes neonates to physical discomfort, pain, and an increased risk of apnea [5]. Given these constraints, stratifying infants based on their risk of developing ROP could improve screening efficiency, prioritizing those at the highest risk while minimizing unnecessary examinations in lower-risk infants.

Artificial intelligence (AI) and machine learning (ML) have emerged as promising tools for assisting ophthalmologists in detecting and predicting ROP progression [6–8]. Previous studies have explored computer-based image analysis techniques to classify ROP using color fundus photographs (CFP) [8–10]. While these approaches have demonstrated high diagnostic accuracy, their ability to predict disease progression remains limited [11]. Moreover, the reliance on CFPs poses challenges in resource-limited settings where fundus imaging devices are not widely available. To address these limitations, researchers have developed clinical data-based prediction models such as DIGIROP-Birth, Colorado-ROP, WINROP, and G-ROP, which incorporate factors like BW, GA, postnatal weight gain, oxygen exposure, and other comorbidities [12–15]. However, recent studies suggest that additional risk factors—including complete blood count (CBC) parameters, blood gas analysis, and maternal factors—may further enhance predictive accuracy [16, 17]. Therefore, this study aims to evaluate the predictive capacity of ML models constructed using demographic and clinical data to identify infants at risk of developing ROP and those likely to require treatment. By focusing solely on clinical and demographic parameters, our study seeks to develop a widely applicable and resource-efficient model, particularly for settings where fundus imaging may not be readily accessible.

Materials and methods

Study population

In this retrospective cohort study, the charts of preterm infants screened for ROP in the Ophthalmology Department of Dokuz Eylul University Hospital from October 2015 to October 2023 were reviewed. Following the ROP screening guidelines published in Türkiye, the study included all infants born at a GA of 34 weeks or less, or with a BW of 1700 g or less [18]. Additionally, preterm infants who received cardiopulmonary support or were deemed at risk for ROP by their neonatologist, regardless of BW and GA, were also included in the screening program. The initial ROP screening was scheduled to begin at 4 weeks postnatally, but not before the infants reached a postmenstrual age of 31 weeks.

The demographic and clinical characteristics associated with the development of ROP were extracted from electronic medical records and are detailed in Table 1. Infants who were missing data on any study parameters or those diagnosed with familial exudative vitreoretinopathy (FEVR) were excluded from the study. Moreover, infants with known metabolic diseases or chromosome abnormalities as well as those with other congenital ocular abnormalities were also excluded. Treatments, including laser photocoagulation and/or anti-vascular endothelial growth factor (VEGF) therapy, were administered according to the latest criteria outlined in the Turkish ROP Screening Guidelines [18–20]. Daily weight gain (DWG) was calculated by subtracting the birth weight from the weight at the first examination and then dividing this difference by the infant's age in days. The study also included initial CBC results, blood gas analysis -ensuring no hemolysis-, and C-reactive protein (CRP) levels within the first 24 h. The total duration of oxygen support used in the ML models was determined as follows: if an infant was discharged from the Neonatal Intensive Care Unit (NICU) before the first ROP screening, the total number of days with oxygen support was obtained from their medical chart. If the infant was still receiving oxygen support at the time of the first ROP screening, the number of postnatal days up to that screening was considered as the total duration of oxygen support.

Structure of the machine learning models

A total of 49 standard features, encompassing demographic information and laboratory tests conducted for any premature infant admitted to the NICU post-birth, were utilized. This set includes nominal features such as BW, GA, the number of days until NICU discharge, and total days with oxygen support, among others. Categorical features include binary variables, such as the occurrence of blood transfusion, presence of intraventricular hemorrhage, and the mode of delivery. Models I and II were developed using parameters available within the

Table 1 The demographic and clinical characteristics related with ROP development

Demographical characteristics	Maternal factors	Blood test results within first 24 h	Co-morbidities	Neonatal intensive care unit related factors
Gestational age (weeks) *	Maternal age *	CBC parameters including erythrocyte count, haemoglobin, haematocrit, MCV, MCHC, RDW, leukocyte, neutrophil, lymphocyte, platelet count, platecrit, NLR and PLR *	Respiratory distress syndrome	Daily weight gain
Birth weight (gr) *	Maternal HT *	Blood gas analysis results including pH, pCO2, pO2, HCO3, lactate and base deficit *	Sepsis	History of mechanical ventilation
Gender *	Gestational DM *	CRP *	Intraventricular hemorrhage	History of non-invasive mechanical ventilation
Multiple pregnancy *	Antenatal steroid treatment *		Hyperbilirubinemia	Total time with O2 support (days)
Type of delivery (NSVB or C/S) *			Bronchopulmonary displasia	Total NICU days
APGAR Score 1st min and 5th min *			Necrotising enterocolitis	History of transfusion
History of CPR *			Hypo/hyperglycemia	History of surfactant use
History of premature membrane rupture *				Total parenteral nutrition support
Existence of any other ocular pathology except for ROP				

All parameters represented in Table 1 are included in models with data known at the first ROP screening

* Included parameters in the machine learning models with data known at birth (first 24 h)

CBC, complete blood count; CPR, cardiopulmonary resuscitation; CRP, C-reactive protein; C/S, cesarean section; DM, diabetes mellitus; HCO3, bicarbonate; HT, hypertension; MCV, mean corpuscular volume; MCHC, mean corpuscular haemoglobin concentration; NICU, neonatal intensive care unit; NLR, neutrophil to lymphocyte ratio; NSVB, normal spontaneous vaginal birth; PLR, platelet to lymphocyte ratio; RDW, red cell distribution weight; ROP, retinopathy of prematurity

first 24 h after birth, incorporating 33 features, while Models III and IV utilized all parameters listed in Table 1.

Therefore, four models with different outputs were designed as follows:

Model I ROP prediction (ROP vs. no ROP) with data known at the first 24 h after birth.

Model II Treatment-required ROP prediction (Treat vs. follow) with data known at the first 24 h after birth.

Model III ROP prediction (ROP vs. no ROP) with data known at the first ROP screening visit.

Model IV Treatment-required ROP prediction (Treat vs. follow) with known data at the first ROP screening visit.

Jupyter Lab, a Python-based platform, was employed to construct the model architectures and assess performance metrics. Algorithm implementation was facilitated by the Sci-kit library, and numeric data preprocessing was performed. The dataset was randomly partitioned into training (85%) and test sets (15%). The performance of each of the four models was explored using five diverse classifier algorithms, including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost).

As a fundamental algorithm, DT partitions data based on attribute values to create a tree-like structure for classification or regression. As a powerful ensemble method, RF, leverages multiple DTs to enhance prediction accuracy while reducing overfitting. LR, on the other hand, is a widely-used linear classification algorithm that models the probability of an event occurring. SVM is a versatile algorithm that excels in both classification and regression tasks by finding the optimal hyperplane that maximizes the margin between different classes. XGBoost, an efficient gradient boosting algorithm, iteratively builds DTs to minimize the loss function and has gained popularity for its robustness and predictive power. Each of these algorithms brings distinct strengths, and their performance was thoroughly evaluated to ensure the reliability of our predictive models for prediction of ROP and treatment necessity.

To address potential issues related to imbalanced data, Synthetic Minority Over-sampling Technique (SMOTE) was applied. Furthermore, 5-fold cross-validation was employed to assess the generalization capabilities of the models.

Performance metrics for the machine learning models

Key performance metrics, including sensitivity, specificity, precision, recall, and F1 score, were computed for all algorithms. Additionally, balanced accuracy was calculated to address the imbalanced nature of the data, which could otherwise skew model performance.

To gain insights into the most significant features within each model, a feature importance analysis was conducted. This analysis facilitated the identification of the top ten parameters that significantly contributed to predictive accuracy.

Results

Study population

A total of 1,850 preterm infants were screened for ROP at least once in our tertiary care center. Infants who were followed as outpatients, those hospitalized in our NICU but with missing data for any study parameter, or those lost to follow-up were excluded from the study. After applying these exclusion criteria, 355 infants remained in the study who were hospitalized in NICU of Dokuz Eylul University Hospital, of whom 114 (32.1%) developed any stage of ROP, and 45 (12.7%) required treatment.

The mean GA at birth for the entire study group was 30.11 ± 3.8 weeks, the mean BW was 1453.19 g, and 50.7% of the participants were male. ROP of any stage was found in 114 (32.1%) infants, among whom 54 (15.2%) infants had stage 1, 34 (9.5%) infants had stage 2, 26 (7.3%) infants had stage 3, 9 (2.5%) had aggressive ROP, and none presented ROP at stage 4 or 5.

Machine learning models

Model I: prediction of ROP with data known at birth

In Model I, 33 features known at birth or during the first 24 h after birth, as detailed in Table 1, were included. Among the tested ML algorithms, the XGBoost classifier demonstrated the highest performance metrics, achieving a sensitivity of 68%, specificity of 92%, balanced accuracy of 0.80, and an F1 score of 0.73. In contrast, the DT classifier exhibited the lowest specifity at 76%, while both RF and LR models shared the lowest sensitivity at 62%. The performance metrics of Model I, using five different classifier algorithms, are presented in Table 2.

The feature importance analysis revealed that BW, GA, and mean corpuscular volume (MCV) in CBC were the most significant parameters contributing to the predictive accuracy of the model. The feature importance graphic for Model I is depicted in Fig. 1.

Model II: prediction of treatment need for ROP with data known at birth

In Model II, which aimed to predict the need for treatment in cases of ROP, 33 features known at birth or within the first 24 h post-birth were employed, as outlined in Table 1. The LR classifier demonstrated the highest sensitivity at 66%, along with a specificity of 95% and a balanced accuracy of 0.81. However, both the DT and RF models exhibited inadequate performance in

Table 2 Performance metrics of various machine learning models for predicting ROP and treatment need using data at different time points

	Sensitivity	Specificity	Balanced Accuracy	Precision	F1 Score
Model I: Prediction of ROP with data known at birth					
Decision Tree	0.68	0.76	0.72	0.55	0.61
Random Forest	0.62	0.86	0.74	0.67	0.65
Logistic Regression	0.62	0.78	0.70	0.75	0.60
Support Vector Machine	0.68	0.86	0.77	0.69	0.69
XGBoost	0.68	0.92	0.80	0.79	0.73
Model II: Prediction of treatment need for ROP with data known at birth					
Decision Tree	0.32	0.82	0.56	0.26	0.25
Random Forest	0.33	0.95	0.64	0.60	0.43
Logistic Regression	0.66	0.95	0.81	0.75	0.71
Support Vector Machine	0.44	0.95	0.70	0.67	0.53
XGBoost	0.44	0.95	0.70	0.67	0.53
Model III: Prediction of ROP with data known at the first ROP screening visit					
Decision Tree	0.56	0.78	0.67	0.53	0.55
Random Forest	0.75	0.76	0.75	0.57	0.65
Logistic Regression	0.50	0.89	0.69	0.67	0.57
Support Vector Machine	0.50	0.86	0.68	0.62	0.56
XGBoost	0.75	0.86	0.80	0.71	0.73
Model IV: Prediction of treatment need for ROP using data known at the first ROP screening visit					
Decision Tree	0.22	0.91	0.56	0.33	0.27
Random Forest	0.33	0.97	0.65	0.75	0.46
Logistic Regression	0.33	0.97	0.65	0.75	0.46
Support Vector Machine	0.22	0.91	0.56	0.33	0.27
XGBoost	0.44	0.88	0.66	0.44	0.44

differentiating cases of ROP requiring treatment. The highest F1 score was achieved by the LR model as 0.71. The performance metrics for Model II are presented in Table 2.

BW and GA, similar to Model I, as well as APGAR score were identified as the top three contributing features to the accuracy of Model II. The feature importance graphic for Model II is presented in Fig. 2.

Model III: prediction of ROP with data known at the first ROP screening visit

In Model III, 49 features identified during the initial screening visit for ROP, as detailed in Table 1, were included. Both XGBoost and RF classifier demonstrated the highest sensitivities at 75%, while the LR classifier achieved the highest specificity of 89%. XGBoost classifier exhibited the highest balanced accuracy and F1 scores of 0.80 and 0.73, respectively. The performance metrics of Model III with five different classifier algorithms are presented in Table 2.

BW, DWG and total NICU days showed the top three contributing features for the accuracy of Model III. Feature importance graphic of Model III was given in Fig. 3.

Model IV: Prediction of treatment need for ROP using data known at the first ROP screening visit

The XGBoost classifier exhibited the highest sensitivity at 44%, while the RF, LR and SVM classifiers demonstrated the highest specificity at 97% for predicting treatment needs based on data from the initial ROP screening visit. The XGBoost algorithm achieved the highest balanced accuracy of 0.66 among all five algorithms. The performance metrics of Model IV with the five classifiers are presented in Table 2.

Gestational age, BW and total time with O2 support showed the top three contributing features for the accuracy of Model IV. Feature importance graphic of Model IV was given in Fig. 4.

Discussion

The findings of our study reveal the significant potential of ML models in predicting ROP among preterm infants, using a combination of demographical and clinical data. The balanced accuracy of our ML models in forecasting both the development of ROP and the subsequent need for treatment has attained a commendable threshold of 80%. This performance is particularly noteworthy given the relatively low prevalence of ROP, akin to conditions such as diabetes or hypertension, resulting in a dataset derived from a limited patient cohort. Furthermore, our study did not rely solely on known standard screening

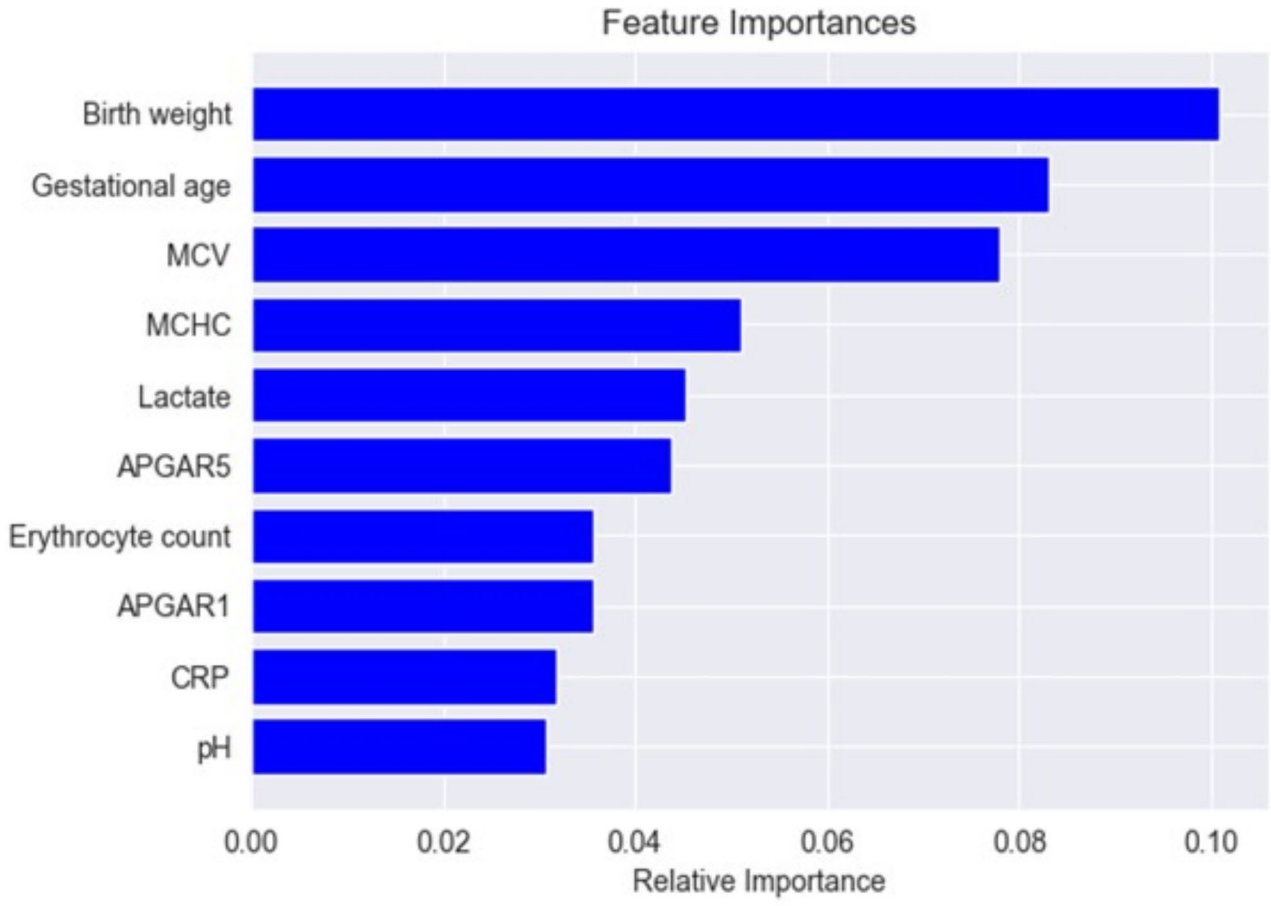

Fig. 1 Feature importance graphic of Model I showing the top ten parameters contributing to predictive accuracy. APGAR, Appearance (skin color), Pulse (heart rate), Grimace (reflex irritability), Activity (muscle tone), and Respiration score; CRP, C-reactive protein; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume

parameters; instead, a large number of variables were incorporated into the model. While this approach carries the potential risk of reducing model performance, it is crucial for the comprehensive evaluation of the multiple factors involved in the disease's etiopathogenesis. Aligned with our purpose, prior studies have explored the interrelationships among various independent risk factors for ROP, revealing that the combined effect of these risks on ROP exceeds the sum of their individual effects, indicating a positive interaction on an additive scale [21].

So far, predictive models of ROP have been developed with applying advanced statistics by incorporating a few parameters besides GA and BW, the two most impornant parameters for the screening. The WINROP algorithm, one of those predictive models is based on GA, BW and weekly weight gain levels. Despite its strong predictive results in various studies, WINROP's clinical application is limited, particularly because it doesn't include infants born after 32 weeks of gestation and also conflicting results between different cohorts belonging to different countries [13, 22]. Similarly, DIGIROP-Birth, a predictive model that estimates early risk for ROP treatment based on GA, BW, and sex in infants born between 24 and 30 weeks, demonstrated strong predictive accuracy, with area under curves ranging from 0.84 to 0.94. Yet, it has not included babies born after 30th weeks. Another predictive algorithm, ROP Score is a scoring system applied at 6 weeks postnatal age, serves as a prediction model for ROP occurrence and ROP severity, including BW, GA, blood transfusion, mechanical ventilation and proportional weight gain at the sixth week postnatal age [23]. However, its use is limited as it doesn't include infants born either before 24 weeks or after 31 weeks. On the other hand, ROP could be diagnosed in more mature babies according to recent studies across various countries including ours [24]. Therefore, a model that includes these mature infants would likely be more practical and generalizable. For this reason, we did not exclude infants born after the 32nd gestational week, except for those suspected of FEVR. Therefore, this study evaluates a comprehensive range of risk factors in the prediction of ROP by ML models in a novel and detailed manner.

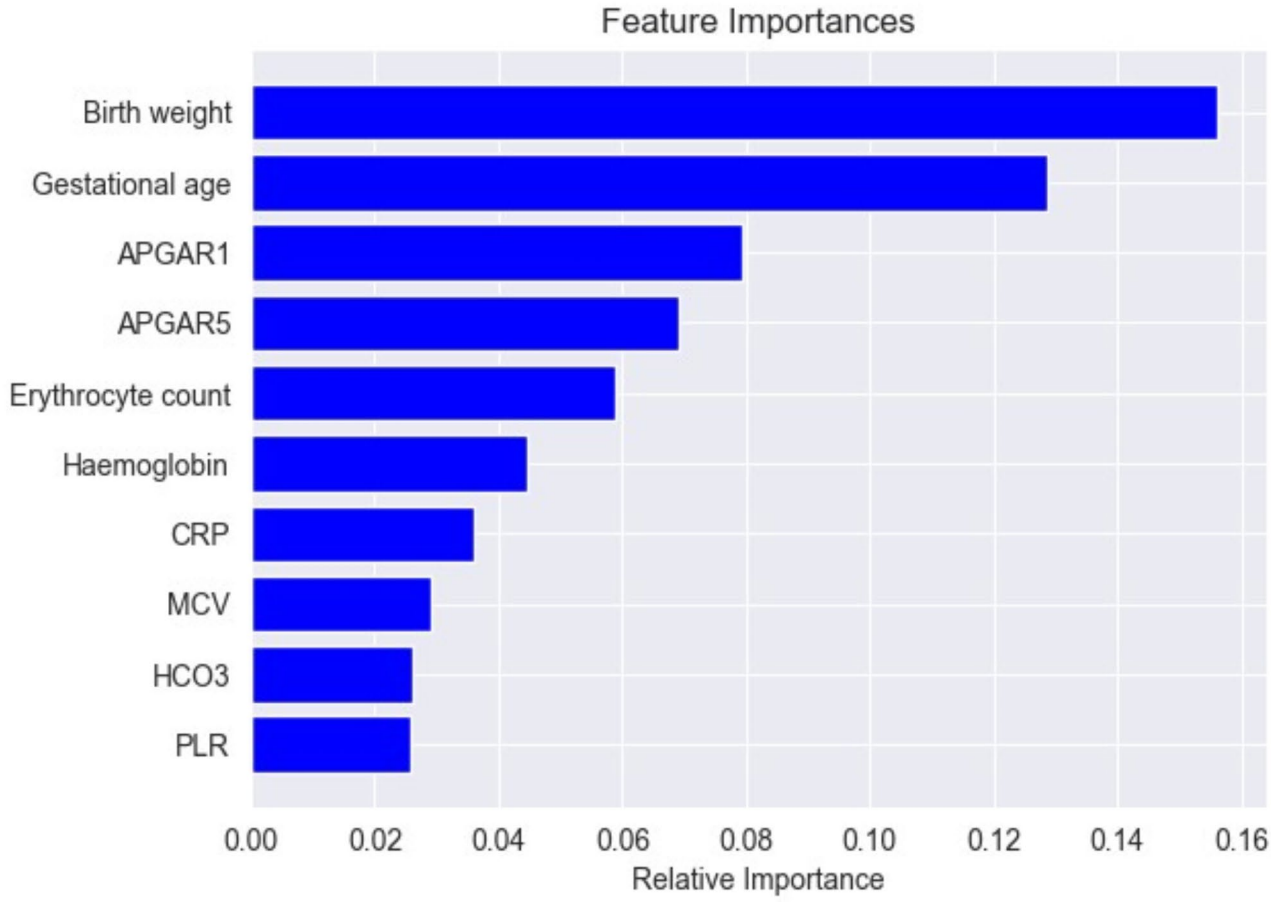

Fig. 2 Feature importance graphic for Model II, showing the top ten parameters that contribute to its predictive accuracy. APGAR, Appearance (skin color), Pulse (heart rate), Grimace (reflex irritability), Activity (muscle tone), and Respiration score; CRP, C-reactive protein; HCO3, blood bicarbonate level; MCV, mean corpuscular volume; PLR, platelet-to-lymphocyte ratio

Our models demonstrated that BW and GA are the primary predictors of ROP development. These factors consistently ranked within the top five predictive features across all four models and were among the top two predictors in three out of the four models. On the other hand, recent studies emphasizing the significance of intrauterine hypoxia and inflammatory markers in ROP development prompted us to incorporate additional parameters into our models, including CBC and blood gas analysis results, as well as systemic inflammatory conditions associated with premature birth, such as necrotizing enterocolitis and intraventricular hemorrhage [16, 25]. Notably, our findings indicate that certain CBC parameters—specifically MCV, erythrocyte count, and hemoglobin (Hgb)—are strong predictors of ROP occurrence in our models. A recent study on a Portuguese cohort revealed that an increase in erythroblasts, MCV, and basophils during the first week of life was significantly and independently associated with ROP development, suggesting that these CBC parameters could serve as early indicators of ROP [26]. Similarly, Akyuz Unsal et al. [27] reported that infants with ROP had lower values of Hgb, hematocrit, MCV, and mean corpuscular hemoglobin concentration (MCHC) during the 4th postnatal week compared to healthy infants. Given that these parameters are related to the blood's oxygen-carrying capacity, it is unsurprising that they may contribute to a hypoxia-related condition such as ROP and probably to longer oxygen administration times. Additionally, we observed that these parameters were more prominent in models (Model I and II) based on data from the first 24 h, which may be due either to the absence of NICU-related strong predictors such as total NICU stay and total days with O2 in those models or to the relative importance of CBC parameters among factors relevant to the first 24 h.

We also identified that lower APGAR scores were predictive of both the development of ROP and the subsequent need for treatment. Several studies have reported an association between lower APGAR scores, as an overall indicator of compromised neonatal health, and an increased incidence of ROP [25]. Additionally, our models highlighted the significance of total NICU days

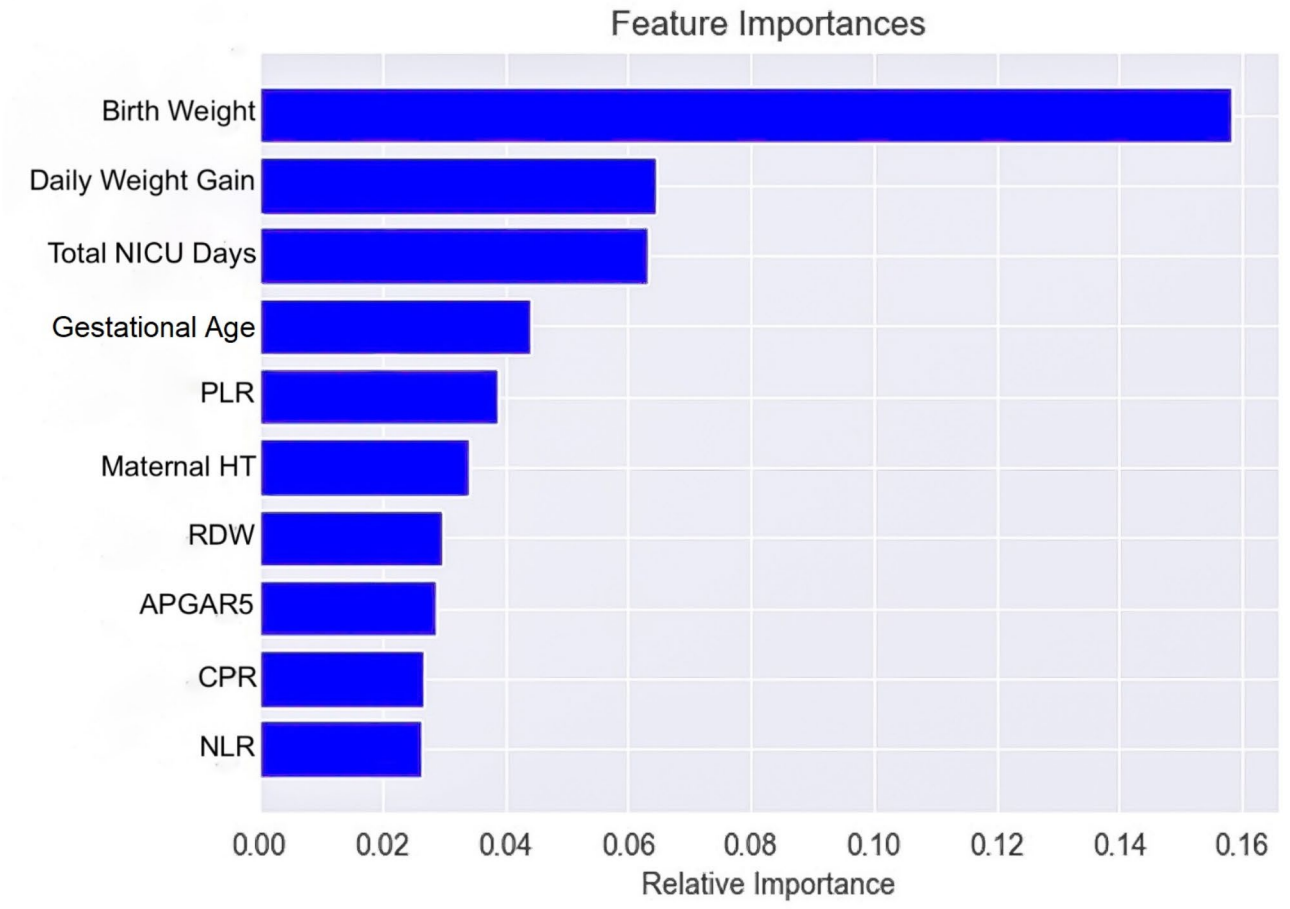

Fig. 3 Feature importance graphic of Model III showing the top ten parameters contributing to predictive accuracy. APGAR, Appearance (skin color), Pulse (heart rate), Grimace (reflex irritability), Activity (muscle tone), and Respiration score; CRP, C-reactive protein; NICU, neonatal intensive care unit; NLR, neutrophil-to-lymphocyte ratio; PLR, platelet-to-lymphocyte ratio; RDW, red cell distribution width

and the duration of oxygen support, both of which have been previously linked to the development and treatment requirement of ROP [28].

Recent studies evaluating the use of AI in ROP have predominantly utilized dilated fundus photographs of infants undergoing ROP screening, demonstrating significant potential in accurately identifying ROP and its various stages [9, 10, 29, 30]. While these advanced deep learning (DL) models are promising, they often require a substantial amount of image data for effective training, and can be computationally demanding. For instance, a study by Wu et al. [9] developed a DL system that successfully predicts the occurrence and severity of ROP before 45 weeks' postmenstrual age. This study included an extensive dataset comprising 7033 retinal photographs of 725 infants for training, and 763 photographs of 90 infants for external validation, along with 46 clinical characteristics for each infant. Remarkably, the study achieved a sensitivity of 100% but a specificity of 7.5% in the external validation set. Further advancing the field, Salih et al. [29] investigated the use of convolutional neural networks (CNNs) with transfer learning models, including VGG-19, ResNet-50, and EfficientNetB5, for ROP detection, attaining an overall accuracy of 87% for EfficientNetB5 based model. Additionally, Wang et al. [30] developed an automated ROP detection method named DeepROP using fundus pictures, which showed a sensitivity and specificity values of 84.91% and 96.90%, respectively, in clinical settings. We also acknowledge that combining clinical data with fundus images may improve model performance, as suggested by the study of Wu et al. [9] Although the authors did not report separate performance metrics for models using only fundus images versus those incorporating clinical data, their high sensitivity may reflect the advantage of this integration. However, the integration of image-based and clinical data-based machine learning models for ROP prediction presents several challenges, including disparities in data availability, standardization of imaging protocols, and model interpretability, which remain significant barriers to seamless implementation. Moreover, acquiring fundus images in infants requires specialized camera

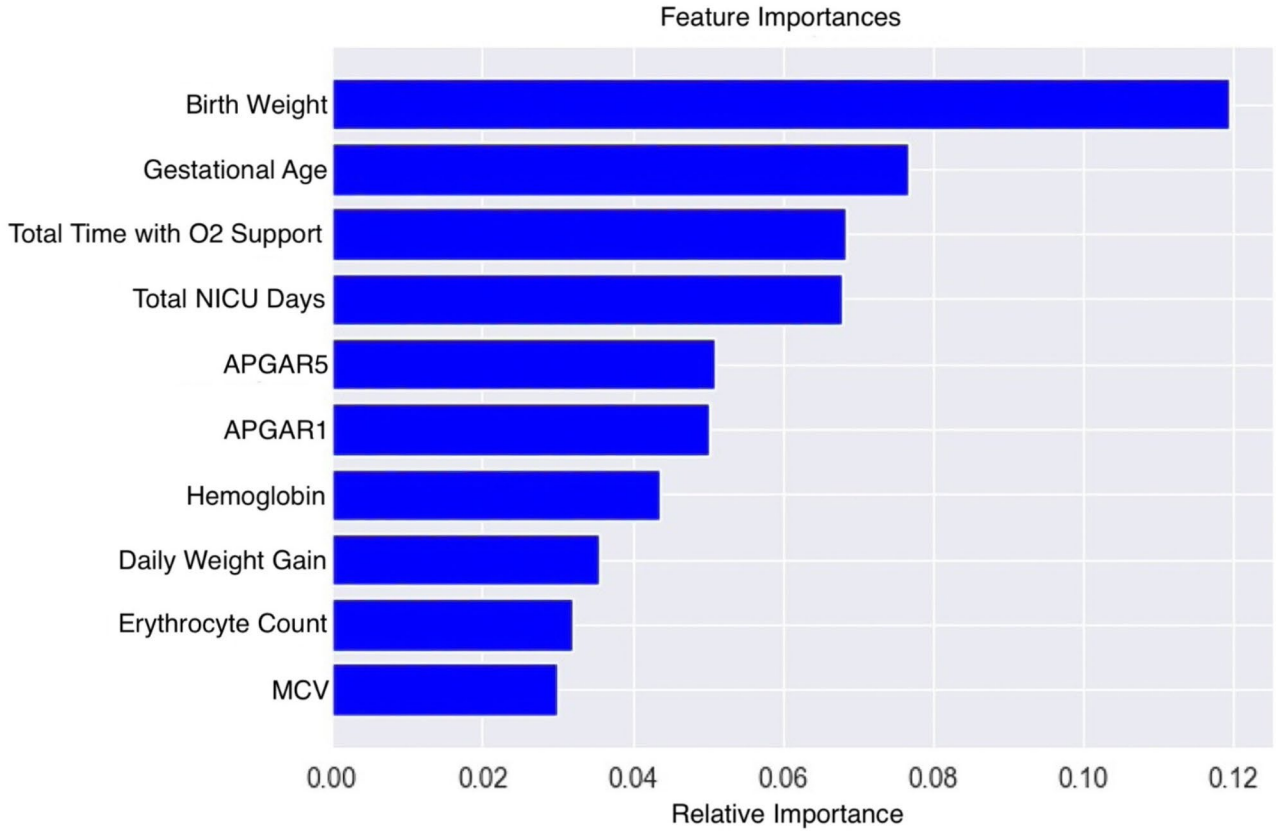

Fig. 4 Feature importance graphic of Model IV showing the top ten parameters contributing to predictive accuracy. APGAR, Appearance (skin color), Pulse (heart rate), Grimace (reflex irritability), Activity (muscle tone), and Respiration score; MCV, mean corpuscular volume; NICU, neonatal intensive care unit; O2, oxygen

systems that are not widely available. Therefore, in this study, we aimed to develop ML models using only tabular clinical data to predict both the occurrence of ROP and the need for treatment.

In a study comparable to ours, Poppe et al. [31] investigated various ML models to predict the need for laser treatment in preterm infants with ROP. The authors integrated routinely monitored physiological data, with a particular emphasis on the infants' oxygenation status, alongside demographic data. The most effective model they developed achieved a sensitivity of 0.73 and notably incorporated the SpO2/FiO2 ratio in conjunction with baseline demographic factors such as GA and BW. Their study, however, was limited by a relatively small sample size of 208 preterm infants, of whom 30 (14%) required laser treatment. To address the challenge of imbalanced data, the authors employed random undersampling to achieve balanced representation between the groups in their analysis. In our study, we utilized the SMOTE technique to address the issue of imbalanced data and incorporated a broader range of risk factors associated with ROP. A recent study by Wu et al. [21] developed a predictive model for ROP screening using back propagation neural network using 12 parameters including GA, BW, caesarean delivery, APGAR score at five minutes, bronchopulmonary displasia, respiratory distress syndrome, intraventricular hemorrhage, hypoxic ischemic encephalopathy, mechanical ventilation, blood transfusions, necrotizing enterocolitis, and patent ductus arteriosus. The area under the curve for prediction model was 0.857 in test in ROP prediction.

In the realm of ML, algorithms like RF and XGBoost have been compared for their effectiveness in various context. These algorithms are known for handling large datasets and complex feature spaces effectively, yet they may not always perform optimally with imbalanced datasets, as in our study [32]. It's a common issue in medical data analysis, where certain conditions or outcomes are rarer than others [33]. This imbalance can lead classifiers to perform better in predicting the majority class while struggling with the minority class, impacting the sensitivity and specificity of the models. Moreover, considering that only 12.7% of cases required treatment, Model II and Model IV in our study appear to be even more imbalanced in terms of sample sizes in the groups than the other two models. To address this issue, several methods were employed, and the best performance metrics were achieved using the SMOTE, a method designed

to generate synthetic data for the minority class. SMOTE works by creating synthetic samples between existing minority class instances, effectively balancing the dataset without simply duplicating data points. XGBoost classifier achieved the highest balanced accuracy of 80% and 81% in Model I and III, respectively which may be explained by its power in handling complex data with high number of features. Although, XGBoost performed the best with a balanced accuracy of 66% and a F1 score of 0.44 in Model 4, those metrics were lower compared to the other models due to the classes which are more imbalanced and model to incorporate higher number of parameters. As for Model II, LR performed the best among all classifiers and this strong performance of LR suggests that linear relationships may exist among the risk factors in this model.

Our findings indicate that the prevalence of ROP and treatment-requiring cases in our cohort aligns with the variability reported across European countries [34, 35]. While the overall treatment rate in our screened population (2.4%) is higher than in Switzerland (1.2%) and some German studies (1.42–7.5%), it remains lower than in Portugal (5.8%) and Sweden (5.7–6.1%) [36]. These discrepancies likely arise from differences in screening guidelines, neonatal care standards, and oxygen therapy protocols. Additionally, as our study was conducted in a tertiary care hospital, high-risk preterm infants were likely overrepresented, contributing to an increased proportion of treatment-requiring cases. These factors highlight the need for further multicenter studies to evaluate ROP incidence and treatment thresholds in different healthcare settings.

It is important to note that in regions experiencing the third ROP epidemic, where access to NICUs is limited and risk factors for ROP development are higher, both the number of infants developing ROP and those requiring treatment are significantly greater. This increased prevalence could lead to a larger and more balanced dataset, potentially enhancing the overall performance of XGBoost while also allowing other ML models, such as RF and LR, to perform more comparably. A more balanced class distribution in these settings might reduce the need for synthetic oversampling techniques like SMOTE and further improve model generalizability. Future studies should explore how epidemiological variations impact ML model performance and validate predictive models in diverse populations.

The primary limitation of our study stems from the imbalance in the sample sizes of our groups. Specifically, the scarcity of cases requiring treatment in our dataset may have introduced a bias in the model's performance. This bias could lead to a skewed preference for detecting healthier cases, rather than those necessitating treatment. To mitigate the impact of this imbalance, we employed SMOTE technique. Additionally, we utilized balanced accuracy as a metric to provide a more accurate evaluation of our models' performance. Another point is that Principal Component Analysis (PCA) for preprocessing, as it makes the data more manageable by transforming correlated variables into uncorrelated principal components, might have increased the performance of our models. But our priority was to assess the clinical significance and impact of each parameter individually to enhance the interpretability of the results. While PCA is an effective method for reducing the dimensionality of data, we chose not to use it in order to fully evaluate the contribution of each variable. Additionally, dimensionality reduction techniques like PCA can reduce the explainability of the parameters, and since one of the goals of this study was to clearly highlight the contribution of key clinical factors to the models' outcomes, we opted to model with raw data rather than using PCA. Also, we did not conduct a head to head comparison with previous ROP predicting algorithms in our study, as we investigated a significantly larger number of parameters in our models compared to those algorithms. Additionally, our results demonstrated that CBC parameters were particularly important for the success of the models, which had not been incorporated into previous algorithms. A potential limitation of our study may be the selection bias introduced by including only infants with complete datasets, which primarily consisted of hospitalized neonates. This may have led to an overrepresentation of infants with additional systemic comorbidities, potentially affecting the generalizability of our findings to all preterm infants. Another key limitation of this study is the lack of external validation, as the models were developed using data from a single institution. Due to the absence of publicly available ROP datasets that include both demographic and clinical parameters, external validation could not be performed. Future studies incorporating multicenter or publicly accessible datasets are necessary to assess the generalizability of our findings. Lastly, the absence of fundus photographs, which could have contributed additional imaging-based features, represents a limitation that may have impacted the model's predictive performance.

Despite the aforementioned limitations, the strength of our study lies in its comprehensive methodology. This involves integrating a broad spectrum of demographic and clinical data to predict ROP. Such an approach not only aligns with the current trends in applying ML in healthcare but also paves the way for the development of more robust and precise prediction models in the future.

Our study successfully establishes that ML models show promise in predicting both the development of ROP and the need for treatment, particularly with high specificity in identifying infants without the condition.

The balanced accuracy of our ML models being around 80% demonstrates that complex clinical data can be effectively processed using ML. Incorporating additional clinical and intensive care-related parameters including hypoxia-related parameters from CBC, the APGAR score, and respiratory data beyond traditional factors may enhance the accuracy of ROP screening and improve clinical decision-making. Despite advances in predictive modeling, fundus examination remains the gold standard for ROP diagnosis, as it enables direct visualization of retinal vascular changes. Machine learning models may serve as complementary tools to enhance screening efficiency, but final clinical decisions should be based on ophthalmologic evaluation.

Acknowledgements
None.

Author contributions
Conceptualization: Ceren Durmaz Engin; Methodology: Ceren Durmaz Engin, Ali Oztas, Mustafa Alper Selver; Formal analysis and investigation: Ceren Durmaz Engin, Ozlem Ozkan, Ali Oztas; Writing - original draft preparation: Ceren Durmaz Engin; Writing - review and editing: Taylan Ozturk, Mustafa Alper Selver; Funding acquisition: None; Resources: Ceren Durmaz Engin; Supervision: Taylan Ozturk, Mustafa Alper Selver, Funda Tuzun.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability
The dataset is available upon request from the corresponding author.

Declarations

Ethics approval and consent to participate
This study was approved by the Ethics Committee of Dokuz Eylul University (Decision Number: 2024/18 – 03, Date: 22.05.2024), which also waived the requirement for informed consent due to the retrospective nature of the study. All procedures were conducted in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 8 December 2024 / Accepted: 27 March 2025
Published online: 10 April 2025

References

1. Gilbert C. Retinopathy of prematurity: a global perspective of the epidemics, population of babies at risk and implications for control. Early Hum Dev. 2008;84(2):77–82. https://doi.org/10.1016/j.earlhumdev.2007.11.009
2. Loh L, Prem-Senthil M, Constable PA. A systematic review of the impact of childhood vision impairment on reading and literacy in education. J Optom. 2024 Apr-Jun;17(2):100495. https://doi.org/10.1016/j.optom.2023.100495. Epub 2023 Nov 1.
3. Adams GGW, Williams C, Modi N, et al. Can we reduce the burden of the current UK guidelines for retinopathy of prematurity screening? Eye (Lond). 2018;32(2):235–7. https://doi.org/10.1038/eye.2017.163
4. Ludwig CA, Chen TA, Hernandez-Boussard T, et al. The epidemiology of retinopathy of prematurity in the united States. Ophthalmic Surg Lasers Imaging Retina. 2017;48(7):553–62. https://doi.org/10.3928/23258160-20170630-06
5. Mitchell AJ, Green A, Jeffs DA, Roberson PK. Physiologic effects of retinopathy of prematurity screening examinations. Adv Neonatal Care. 2011;11(4):291–7. https://doi.org/10.1097/ANC.0b013e318225a332
6. Campbell JP, Chiang MF, Chen JS, et al. Artificial intelligence for retinopathy of prematurity: validation of a vascular severity scale against international expert diagnosis. Ophthalmology. 2022;129(7):e69–76. https://doi.org/10.1016/j.ophtha.2022.02.008
7. Morrison SL, Dukhovny D, Chan RVP, et al. Cost-effectiveness of artificial Intelligence-Based retinopathy of prematurity screening. JAMA Ophthalmol. 2022;140(4):401–9. https://doi.org/10.1001/jamaophthalmol.2022.0223
8. Scruggs BA, Chan RVP, Kalpathy-Cramer J, et al. Artificial intelligence in retinopathy of prematurity diagnosis. Transl Vis Sci Technol. 2020;9(2):5. https://doi.org/10.1167/tvst.9.2.5. Published 2020 Feb 10.
9. 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. https://doi.org/10.1001/jamanetworkopen.2022.17447. Published 2022 Jun 1.
10. Chen JS, Coyner AS, Ostmo S, et al. Deep learning for the diagnosis of stage in retinopathy of prematurity: accuracy and generalizability across populations and cameras. Ophthalmol Retina. 2021;5(10):1027–35. https://doi.org/10.1016/j.oret.2020.12.013
11. Taylor S, Brown JM, Gupta K, et al. Monitoring disease progression with a quantitative severity scale for retinopathy of prematurity using deep learning. JAMA Ophthalmol. 2019;137(9):1022–8. https://doi.org/10.1001/jamaophthalmol.2019.2433
12. Hutchinson AK, Melia M, Yang MB, et al. Clinical models and algorithms for the prediction of retinopathy of prematurity: A report by the American academy of ophthalmology. Ophthalmology. 2016;123(4):804–16. https://doi.org/10.1016/j.ophtha.2015.11.003
13. Lundgren P, Stoltz Sjöström E, Domellöf M et al. WINROP identifies severe retinopathy of prematurity at an early stage in a nation-based cohort of extremely preterm infants. PLoS One. 2013;8(9):e73256. Published 2013 Sep 12. https://doi.org/10.1371/journal.pone.0073256
14. Binenbaum G, Bell EF, Donohue P, et al. Development of modified screening criteria for retinopathy of prematurity: primary results from the postnatal growth and retinopathy of prematurity study. JAMA Ophthalmol. 2018;136(9):1034–40. https://doi.org/10.1001/jamaophthalmol.2018.2753
15. Hauspurg AK, Allred EN, Vanderveen DK, et al. Blood gases and retinopathy of prematurity: the ELGAN study. Neonatology. 2011;99(2):104–11. https://doi.org/10.1159/000308454
16. Ozturk T, Durmaz Engin C, Kaya M, Yaman A. Complete blood count parameters to predict retinopathy of prematurity: when to evaluate and what do they tell Us?? Int Ophthalmol. 2021;41(6):2009–18. https://doi.org/10.1007/s10792-021-01756-7
17. Holmström G, Thomassen P, Broberger U. Maternal risk factors for retinopathy of prematurity–a population-based study. Acta Obstet Gynecol Scand. 1996;75(7):628–35. https://doi.org/10.3109/00016349609054687
18. Koc E, Bas A, Ozdek S. Türkiye Prematüre Retinopatisi Rehberi. https://www.todnet.org/tod-rehber/rop-tedavi-rehberi-2021.pdf. Published 2021. Accessed August 14, 2024.
19. Koç E, Baş AY, Özdek Ş, Ovalı F, Başmak H, Türkiye Prematüre Retinopatisi Rehberi Komisyonu, et al. Türkiye Prematüre Retinopatisi Rehberi. Türk Neonatoloji Derneği; 2016. Available from: https://www.medikalakademi.com.tr/wp-content/uploads/2020/07/turkiye-premature-retinopatisi-rehberi-2016.pdf. Accessed: March 16, 2025.
20. Koç E, Baş AY, Özdek Ş, Ovalı F, Başmak H. Turkish neonatal and Turkish ophthalmology societies consensus guideline on the retinopathy of prematurity. Turkish Archives Pediatrics/Türk Pediatri Arşivi. 2018;53(Suppl 1):S151.
21. Wu R, Chen H, Bai Y et al. Prediction models for retinopathy of prematurity occurrence based on artificial neural network. BMC Ophthalmol. 2024;24(1):323. Published 2024 Aug 5. https://doi.org/10.1186/s12886-024-03562-y
22. Koçak N, Niyaz L, Ariturk N. Prediction of severe retinopathy of prematurity using the screening algorithm WINROP in preterm infants. J AAPOS. 2016;20(6):486–9. https://doi.org/10.1016/j.jaapos.2016.08.008
23. Eckert GU, Fortes Filho JB, Maia M, Procianoy RS. A predictive score for retinopathy of prematurity in very low birth weight preterm infants. Eye (Lond). 2012;26(3):400–6. https://doi.org/10.1038/eye.2011.334

24. Dericioğlu V, Butur S, Celiker H, Şahin Ö, Incidence. Risk factors and screening evaluation of retinopathy of prematurity in high birthweight infants: A large cohort study in Turkey. Ophthalmic Epidemiol. 2022;29(1):78–84. https://doi.org/10.1080/09286586.2021.1894582
25. de Las Rivas Ramírez N, Luque Aranda G, Rius Díaz F, et al. Risk factors associated with retinopathy of prematurity development and progression. Sci Rep. 2022;12(1):21977. https://doi.org/10.1038/s41598-022-26229-4. Published 2022 Dec 20.
26. Fevereiro-Martins M, Santos AC, Marques-Neves C, GenE-ROP Study Group, et al. Complete blood count parameters as biomarkers of retinopathy of prematurity: a Portuguese multicenter study. Graefes Arch Clin Exp Ophthalmol. 2023;261(10):2997–3006. https://doi.org/10.1007/s00417-023-06072-7
27. Akyüz Ünsal Aİ, Key Ö, Güler D, et al. Can complete blood count parameters predict retinopathy of prematurity?? Turk J Ophthalmol. 2020;50(2):87–93. https://doi.org/10.4274/tjo.galenos.2019.45313
28. Stefánsson E. Ocular oxygenation and the treatment of diabetic retinopathy. Surv Ophthalmol. 2006;51(4):364–80. https://doi.org/10.1016/j.survophthal.2006.04.005
29. Salih N, Ksantini M, Hussein N, et al. Prediction of ROP zones using deep learning algorithms and voting classifier technique. Int J Comput Intell Syst. 2023;16:86. https://doi.org/10.1007/s44196-023-00268-9
30. Wang J, Ju R, Chen Y, et al. Automated retinopathy of prematurity screening using deep neural networks. EBioMedicine. 2018;35:361–8. https://doi.org/10.1016/j.ebiom.2018.08.033
31. Poppe JA, Fitzgibbon SP, Taal HR, et al. Early prediction of severe retinopathy of prematurity requiring laser treatment using physiological data. Pediatr Res. 2023;94(2):699–706. https://doi.org/10.1038/s41390-023-02504-6
32. Chang W, Liu Y, Xiao Y, et al. A Machine-Learning-Based prediction method for hypertension outcomes based on medical data. Diagnostics (Basel). 2019;9(4):178. https://doi.org/10.3390/diagnostics9040178. Published 2019 Nov 7.
33. Mundra S, Vijay S, Mundra A, et al. Classification of imbalanced medical data: an empirical study of machine learning approaches. J Intell Fuzzy Syst. 2022;43:1933–46. https://doi.org/10.3233/JIFS-219294
34. Bas AY, Demirel N, Koc E, Isik DU, Hirfanoglu İM, Tunc T. Incidence, risk factors and severity of retinopathy of prematurity in Turkey (TR-ROP study): a prospective, multicentre study in 69 neonatal intensive care units. Br J Ophthalmol. 2018;102(12):1711–6.
35. Özdek Ş, Ozdemir HB, Ozen Tunay Z, Bayramoglu SE, Alyamac Sukgen E, Kır N et al. Clinical and Demographic Characteristics of Treatment Requiring Retinopathy of Prematurity in Big Premature Infants in Turkiye: Report No. 1 (BIG-ROP Study). Ophthalmologica. 2024:1–11.
36. Modrzejewska M, Bosy-Gąsior W. Most Up-to-Date Analysis of Epidemiological Data on the Screening Guidelines and Incidence of Retinopathy of Prematurity in Europe-A Literature Review. J Clin Med. 2023;12(11):3650. Published 2023 May 24. https://doi.org/10.3390/jcm12113650

Publisher's note