

Smartphone Imaging and Artificial Intelligence For Neonatal Jaundice Screening: From Colorimetry To Clinical Validation

Dr. M. Indra Devi^{*1}, Dr. Annalakshmi Govindaraj², Mrs. S. Ummugulthum Natchiar³

¹Department of Computer Science and Engineering (AIML), SRM Madurai College of Engineering and Technology, Tamil Nadu, India, indradevi20@gmail.com

²Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana, India, annalakshmi@klh.edu.in

³Department of Computer Science and Engineering (AIML), NPR College of Engineering and Technology, Tamil Nadu, India, summugulthum@gmail.com

Corresponding Author: Dr. M. Indra Devi

Abstract

Neonatal jaundice is common during the newborn period, but clinically significant hyperbilirubinemia requires timely detection and appropriate follow-up. Total serum bilirubin (TSB) remains an important reference measurement, while transcutaneous bilirubinometry (TcB) provides a non-invasive screening approach. Smartphone imaging combined with artificial intelligence (AI) has emerged as a potential complementary method for bilirubin assessment.

This article synthesizes published evidence on smartphone-based neonatal jaundice screening, covering colourimetric approaches, machine learning, deep learning, calibration, population diversity and clinical implementation. Findings from studies of BiliCam, Biliscan, Picterus, BiliSG and other smartphone-based systems indicate that image-derived measurements can be associated with TSB. However, reported performance varies with the study population, imaging device, acquisition conditions, bilirubin range and validation method. Important unresolved issues include underestimation at high bilirubin concentrations, performance in preterm infants and populations with darker skin pigmentation, cross-device generalizability, image-quality failures and limited prospective external validation.

The evidence also highlights the need to distinguish bilirubin estimation from threshold classification, referral decisions and demonstrated clinical benefit. Future investigations should prioritize multicentre datasets, standardized acquisition and calibration, subgroup-specific performance assessment, uncertainty-aware prediction and prospective evaluation of screening pathways. Smartphone-based systems should currently be considered potential screening and decision-support tools rather than replacements for established bilirubin measurement.

Keywords: Neonatal jaundice; hyperbilirubinemia; smartphone imaging; bilirubin estimation; artificial intelligence; machine learning.

1. Introduction

Neonatal jaundice is one of the most common clinical conditions during the newborn period. Although most cases are transient, significant unconjugated hyperbilirubinemia can progress to acute bilirubin encephalopathy and kernicterus. Early recognition and appropriate bilirubin assessment are important, particularly because bilirubin levels may change rapidly after birth and opportunities for follow-up can be limited after early discharge.

Clinical assessment currently relies on visual examination together with transcutaneous bilirubinometry (TcB) and total serum bilirubin (TSB) measurement. TSB remains an important reference measurement for clinical decision-making, but it requires blood collection and access to appropriate laboratory or point-of-care facilities. TcB provides a non-invasive alternative for screening, while visual assessment is inexpensive and readily available but can be affected by observer experience, illumination, anatomical site and skin pigmentation. These practical and measurement-related limitations have encouraged investigation of portable optical methods that could complement existing approaches.

Smartphone imaging represents one such approach, using a consumer device as a potential optical sensing platform. In a typical workflow, an image of neonatal skin or another anatomical region is acquired, colour information is calibrated and processed, and statistical or artificial-intelligence (AI) methods are applied to estimate bilirubin levels or classify screening risk. The resulting pathway can be represented as image acquisition → calibration → region-of-interest/feature extraction → AI analysis → bilirubin estimation or classification → screening/referral. This approach has the potential to combine low-cost image acquisition with computational analysis while reducing dependence on specialized equipment.

Early clinical investigations provided evidence that smartphone images can capture information associated with bilirubin levels. BiliCam, evaluated in 530 newborns across seven sites in the United States, reported a correlation

of 0.91 with TSB and high sensitivity under the study's screening criteria [1]. Subsequent work explored colour calibration, machine-learning methods, multi-site imaging, optical accessories and deep-learning models. These studies, however, have not produced uniform performance across populations or clinical settings. A meta-analysis of 14 studies reported a pooled correlation of 0.77 (95% CI 0.69–0.83) with TSB, with substantial variation in reported sensitivity and specificity [2]. Such variation highlights the importance of considering not only the association between image-derived measurements and TSB, but also the conditions under which these systems are developed and validated.

The clinical relevance of smartphone-based bilirubin assessment depends on more than its correlation with total serum bilirubin (TSB). The studies reviewed used different imaging methods, calibration procedures, patient populations and validation designs, which limits direct comparison of their reported performance. For example, a model evaluated in term newborns under controlled lighting conditions cannot be assumed to perform similarly in preterm infants or when images are captured at home. Likewise, a high correlation does not establish that the method will reliably identify infants who require further assessment. These differences make it necessary to examine measurement agreement, threshold-specific sensitivity and specificity, and the extent of validation beyond the original study setting. This review therefore considers both the technical development of smartphone-based assessment and the evidence needed to establish its suitability for clinical screening.

This article examines the development of smartphone-based neonatal jaundice assessment, from early colourimetric techniques to machine-learning and deep-learning systems. It focuses on four related questions: how smartphone-derived bilirubin estimates are obtained and evaluated; how performance varies across populations, devices and clinical settings; which methodological and implementation challenges limit generalizability; and what evidence is needed before these technologies can be integrated safely into screening and referral pathways. By distinguishing measurement performance from clinical utility, the article identifies priorities for further technical development and prospective evaluation.

1.1 Literature Selection Approach

A focused literature search was conducted on 10 October 2026 to identify published research on smartphone-based neonatal jaundice assessment. PubMed-indexed records were examined using combinations of terms related to neonatal jaundice and hyperbilirubinemia, smartphone imaging, mobile health, colourimetry, image analysis, machine learning, deep learning, bilirubin estimation and screening. The search was intended to identify studies evaluating smartphone-derived bilirubin measurements against total serum bilirubin (TSB), as well as studies reporting threshold-based screening performance, validation across populations or devices, and clinical implementation.

Studies were considered relevant when they investigated smartphone-based or mobile-image approaches to neonatal jaundice assessment and reported findings related to measurement performance, screening classification, calibration, population-specific performance or implementation. Particular attention was given to studies involving colour-calibrated imaging, machine-learning and deep-learning models, preterm infants, populations with diverse skin pigmentation, and use in clinical or home settings. Published systematic reviews and meta-analyses were considered where they provided broader evidence on the performance of smartphone-based screening.

The selected literature was examined in relation to the clinical task evaluated, study population, imaging method, reference measurement and validation design. Quantitative findings were interpreted according to the measures reported in the original studies, including correlation, agreement, sensitivity, specificity and area under the receiver operating characteristic curve (AUC). These measures were not treated as interchangeable indicators of performance. Particular consideration was given to reported underestimation at higher bilirubin concentrations, differences between term and preterm populations, and the extent to which findings had been evaluated beyond the original study setting.

The evidence was synthesized narratively to compare the development of smartphone-based approaches, identify areas of agreement and inconsistency, and examine limitations affecting generalizability and clinical implementation. Quantitative estimates from published systematic reviews were reported as findings of those reviews; no new pooled estimates were calculated for this article.

This review is a focused narrative synthesis rather than a systematic review or diagnostic-test meta-analysis. It does not claim to provide an exhaustive search of all relevant publications, a formal study-selection flow diagram or a comprehensive risk-of-bias assessment. The findings should therefore be interpreted in the context of the selected literature and the methodological differences between the studies discussed.

The search was restricted to PubMed-indexed records; engineering-oriented sources (e.g., IEEE Xplore, ACM Digital Library, Scopus) and conference proceedings were not searched, so relevant image-processing and AI studies may have been missed.

2. Neonatal Jaundice and the Need for Accessible Screening

Clinical assessment of neonatal jaundice requires more than image recognition. Decisions about bilirubin depend on the infant's age in hours, gestational age and relevant clinical risk factors [3]. Estimating a continuous bilirubin value and classifying an infant as being above a clinically important threshold are related, but distinct, tasks.

TSB remains an important reference measurement for treatment decisions, although blood collection and access to laboratory or point-of-care facilities limit its use for decentralized screening. TcB avoids blood sampling and supports non-invasive assessment, but requires dedicated equipment. Smartphone imaging may provide an additional screening option where such equipment is not readily available.

The World Health Organization recommends universal transcutaneous bilirubin screening at health-facility discharge, followed by serum bilirubin measurement, appropriate treatment and follow-up when indicated [4]. This recommendation provides a clinical context for evaluating smartphone-based methods. Any proposed role for these systems should depend on demonstrated measurement reliability, safe referral performance and compatibility with established care pathways.

The same guideline made no recommendation for or against universal TSB screening at discharge, citing insufficient evidence.

Visual assessment is inexpensive and widely available, yet it does not provide a quantitative bilirubin measurement. Perceived yellowness varies with skin pigmentation and illumination. Smartphone imaging has additional sources of variation, including camera characteristics, lighting and computational photography. Findings from populations with darker pigmentation make subgroup-specific evaluation particularly important.

The intended clinical task should be explicit. Smartphone systems may estimate bilirubin, classify values against a threshold, predict the need for referral or support a clinical workflow. These tasks are not interchangeable. Correlation with TSB establishes an association between an image-derived estimate and the reference measurement; it does not by itself establish safe referral decisions or improved outcomes. The value of smartphone screening therefore depends on reliable measurement, accessibility and a clear route to confirmatory testing.

3. Smartphone-Based Optical Imaging for Bilirubin Assessment

Smartphone-based jaundice assessment is an optical measurement task. Recorded pixel values reflect tissue characteristics as well as illumination, tissue reflectance, camera spectral response, exposure, white balance, imaging distance, anatomical site and proprietary image-processing operations. RGB values should consequently not be interpreted as direct measurements of bilirubin concentration.

A simplified representation of the recorded signal is given by Equation (1):

$$I_c(p) = \int E(\lambda) \cdot R_p(\lambda) \cdot S_c(\lambda) d\lambda \quad (1)$$

In Equation (1), $I_c(p)$ is the recorded signal for colour channel c at pixel p ; $E(\lambda)$ is the illumination spectrum; $R_p(\lambda)$ is the tissue's spectral reflectance at pixel p ; $S_c(\lambda)$ is the spectral sensitivity of channel c ; and λ is wavelength. The relationship helps explain why two smartphones may record different RGB values from the same skin. Automatic exposure, white balance, tone mapping, denoising and JPEG processing can add further variation.

Most systems analyse a region of interest (ROI) from the sternum, chest, forehead or abdomen; the sclera has also been investigated [5]. The recorded RGB values may be transformed into HSV, CIE Lab or normalized chromaticity coordinates. The normalized chromaticity coordinates are defined in Equation (2):

$$r = R/(R + G + B); \quad g = G/(R + G + B); \quad b = B/(R + G + B) \quad (2)$$

In Equation (2), R , G and B denote the red, green and blue channel values, while r , g and b are their normalized chromaticity coordinates. Normalization reduces dependence on overall brightness, but it does not remove the effect of illuminant colour. The blue–yellow component of CIE Lab, b^* , may be relevant to jaundice and is calculated from the CIE XYZ tristimulus values of the linearized pixel colour, as shown in Equation (3):

$$b^* = 200 [f(Y/Y_n) - f(Z/Z_n)] \quad (3)$$

In Equation (3), Y and Z are the pixel's CIE XYZ tristimulus values, and Y_n and Z_n are the corresponding tristimulus values of the reference white. The function $f(t)$ is the standard CIE Lab nonlinear transformation, $f(t) = t^{1/3}$ for $t > (6/29)^3$, and $f(t) = t/[3(6/29)^2] + 4/29$ otherwise. The XYZ values are obtained from gamma-decoded (linearized)

sRGB values using the standard sRGB-to-XYZ matrix. Larger b^* values indicate greater yellowness. No colour representation fully removes variation caused by illumination and device characteristics.

Calibration is one response to these optical limitations. A reference colour card or controlled optical accessory can help characterize transformations introduced during image acquisition. In a Norwegian study of 302 newborns, colour-calibrated estimates correlated with TSB at $r = 0.84$; sensitivity was 100% and specificity 69% for severe jaundice above $250 \mu\text{mol/L}$ ($\approx 14.6 \text{ mg/dL}$; $1 \text{ mg/dL} \approx 17.1 \mu\text{mol/L}$) [6]. These results support the use of calibration, but do not show that it eliminates all measurement variability.

Image quality and ROI selection also affect the reliability of a prediction. Motion blur, poor focus, shadows, reflections, incorrect positioning and inadequate exposure can degrade the input image. A cautious workflow is therefore:

Image $\rightarrow$ Quality assessment $\rightarrow$ ROI $\rightarrow$ Calibration $\rightarrow$ Prediction $\rightarrow$ Uncertainty $\rightarrow$ Screening decision

Images that fail quality checks should be reacquired rather than passed directly to the prediction model.

Studies should report the proportion of images rejected or requiring repeat acquisition. If difficult images are excluded without reporting how often this occurs, practical performance may be overstated, particularly when parents or community health workers capture the images.

The choice between simple hardware and tighter optical control involves a practical trade-off. Ordinary smartphone photography is convenient and requires little additional equipment; dedicated illumination or clip-on accessories may improve consistency but add hardware. The robustness of either approach needs to be assessed in routine clinical settings as well as semi-controlled community and home environments.

4. Evolution of Smartphone-Based Jaundice Assessment

Table 1 summarizes selected studies. Across this literature, development has moved from basic smartphone colourimetry to calibrated imaging, machine-learning regression, optical instrumentation, multi-site and multi-country evaluation, and deep-learning methods. These developments address two related challenges: variability in smartphone photography and the limits of simpler statistical models.

BiliCam provided early evidence that smartphone photographs contain information associated with bilirubin levels. Later colour-calibrated approaches sought to reduce acquisition-related variation. Biliscan combines sternal skin imaging, a standard colour card and machine-learning regression. In a multi-ethnic neonatal population, its evaluation reported moderate correlation and limited agreement with TSB [7]. A separate evaluation in preterm infants also underscores that performance in this group should be assessed independently rather than inferred from studies of term newborns [8].

Picterus has been evaluated across more than one country. Among 416 infants in Mexico, Nepal and the Philippines, the overall correlation with TSB was 0.76, although site-level performance differed and limits of agreement were wide [9]. In a separate Mexican study of 177 term newborns, the Spearman correlation was $\rho = 0.68$; sensitivity was 85.7% and specificity 80.1% for TSB $\geq 250 \mu\text{mol/L}$. The authors also reported a tendency to underestimate higher bilirubin values [10]. The results should be interpreted in light of the populations and evaluation conditions rather than treated as directly interchangeable estimates.

BiliSG illustrates the use of feature-based machine learning across multiple anatomical sites. It combined forehead, sternum and abdomen yellowness features using an ensemble of gradient-boosted trees. The development cohort included 352 neonates, and temporal validation involved another 194. In that validation cohort, Pearson correlation was $r = 0.84$, 82% of paired measurements were within $\pm 3 \text{ mg/dL}$, sensitivity was 100%, specificity was 70% (predicted $\geq 13 \text{ mg/dL}$ used to detect TSB $\geq 17 \text{ mg/dL}$) and the area under the receiver operating characteristic curve (AUC) was 0.89 [11]. These measures describe different aspects of performance and should be considered together. The findings in Table 1 illustrate why the reported correlation coefficients should not be used as a stand-alone basis for comparing systems. BiliCam reported a correlation of 0.91 with TSB, whereas the multi-country Picterus study reported 0.76 and the multi-ethnic Biliscan study reported 0.54 [1,7,9]. These values were obtained using different devices, populations and evaluation procedures; they therefore do not establish a simple ranking of the technologies. The BiliSG results provide additional information because the study reported both correlation and the proportion of estimates within a specified error range, together with threshold-based diagnostic performance [11]. Taken together, the studies suggest that comparative evaluation should consider agreement and the consequences of classification errors alongside correlation.

A smartphone-based transcutaneous bilirubinometer incorporating a snap-on adapter, smartphone flash and diffuse-reflectance measurement was evaluated in a 37-patient pilot study. The reported coefficient of determination was $R^2 = 0.42$, with a mean difference of -0.28 mg/dL and 95% limits of agreement from -7.0 to $+6.4$ mg/dL [12]. The results indicate the need for further evaluation before the approach can be considered reliable for clinical screening.

In a study of 341 newborns, a Taiwanese smartphone model reported $R = 0.90$, a root-mean-square error (RMSE) of 2.44 mg/dL and $AUC = 0.963$ [13]. These figures arise from different study designs and should not be used alone to rank the systems.

Table 1. Selected smartphone-based neonatal jaundice assessment studies

System or study	Population	Imaging site, calibration and device	Reference standard	Validation design	Main finding (95% CI where reported)	Important qualification
BiliCam [1]	530 newborns, <7 days old, ≥ 35 weeks; seven US sites	Sternum; colour calibration card; iPhone 5s; images with and without flash	TSB	Prospective, multi-site	$r = 0.91$ (95% CI 0.89–0.92). Sensitivity/specificity 84.6%/75.1% (Bhutani high-risk zone) and 100%/76.4% (≥ 13 mg/dL to predict TSB ≥ 17 mg/dL)	Efficacy study by trained investigators; the authors did not consider it sufficient as a standalone diagnostic tool.
Colour-calibrated imaging [6]	302 healthy term infants aged up to 15 days; two Norwegian hospitals	Colour calibration card; standardised illumination at one hospital, non-standardised at the other	TSB (TcB also measured)	Prospective, cross-sectional	$r = 0.84$ with TSB; $r = 0.81$ with TcB. TSB > 250 $\mu\text{mol/L}$: sensitivity 100%, specificity 69%	Healthy term infants only; calibration does not remove every source of optical variation.
Biliscan, multi-ethnic [7]	61 paired Biliscan–TSB measurements; 85 paired Biliscan–TcB measurements. Infants born at gestational age ≥ 35 weeks, with clinical jaundice or recommended for screening, within 21 days after birth	Sternal skin images captured using the Biliscan smartphone application with a standard colour card; colour features were extracted and processed using machine-learning regression.	TSB	Prospective cross-sectional study at a tertiary teaching hospital, Singapore; August 2020–October 2021.	$r = 0.54$; sensitivity 76.92%, specificity 70.83% for phototherapy-requiring jaundice	Moderate correlation and limited agreement with TSB.

System or study	Population	Imaging site, calibration and device	Reference standard	Validation design	Main finding (95% CI where reported)	Important qualification
Biliscan, preterm [8]	Preterm infants, n = 129, born at gestational age below 35 weeks and admitted to the neonatal intensive care unit	Chest/sternum; colour card	TSB	Clinical feasibility and reliability study in hospitalized preterm infants. Infants requiring phototherapy during the preceding 12 hours were excluded.	ICC = 0.512 (95% CI 0.353–0.638) for Biliscan versus TSB	Moderate agreement; preterm performance should be assessed separately from term infants.
Picterus, multi-country [9]	n=416; Mexico, Nepal and the Philippines	Chest; Picterus colour calibration card	TSB	Cross-sectional, three countries (2018–2022)	Overall correlation $r=0.76$; reported 95% limits of agreement approximately -98.8 to +79.8 $\mu\text{mol/L}$.	Site-level performance and agreement require attention.
Picterus, Mexico [10]	177 term newborns; Mexico	Chest; Picterus colour calibration card	TSB	Cross-sectional	Spearman $\rho = 0.68$; sensitivity 85.7%, specificity 80.1% (TSB $\geq 250 \mu\text{mol/L}$)	Tendency to underestimate higher bilirubin values.
BiliSG [11]	352 neonates (development, cross-validation) and 194 (temporal validation); 546 in total; multi-ethnic, Singapore; median gestational age 38 weeks	Forehead, sternum and abdomen; colour calibration stickers; gradient-boosted trees	TSB	Temporal validation (November 2023–June 2024)	$r = 0.84$ (95% CI 0.79–0.88); 82% within $\pm 3 \text{ mg/dL}$; sensitivity 100%, specificity 70%, AUC 0.89 (predicted $\geq 13 \text{ mg/dL}$ used to detect TSB $\geq 17 \text{ mg/dL}$)	Positive predictive value 10%; tested on one smartphone model so far.
Harrison-Smith et al. [12]	37 patients; multi-ethnic pilot	Snap-on adapter, smartphone flash, diffuse reflectance	TSB	Pilot study	$R^2 = 0.42$; mean difference = -0.28 mg/dL ; 95% limits of agreement = -7.0 to +6.4 mg/dL	Very small pilot; requires dedicated hardware.

System or study	Population	Imaging site, calibration and device	Reference standard	Validation design	Main finding (95% CI where reported)	Important qualification
Taiwan smartphone model [13]	341 healthy newborns aged under 28 days; gestational age ≥ 35 weeks. Infants who had received phototherapy were excluded.	Facial and chest images; smartphone camera with a custom-made colour-calibration card; colour normalization, skin-region extraction and colour-feature engineering.	Total serum bilirubin (TSB).	Model development and evaluation in the enrolled cohort; several regression models compared. No separate external-validation cohort reported	$R = 0.90$; RMSE = 2.44 mg/dL; AUC = 0.963 sensitivity = 90.1%; specificity = 89.0%	Healthy newborns born at ≥ 35 weeks' gestation; further validation is required before wider clinical use.
Picterus, Botswana [16]	123 term newborns aged 1–14 days; darker skin pigmentation	Chest; Picterus colour calibration card	Laboratory TSB (TcB device as comparator)	Prospective	$r = 0.66$ (95% CI 0.55–0.75) for Picterus vs 0.84 (0.78–0.89) for TcB; TSB ≥ 250 $\mu\text{mol/L}$: sensitivity 64% vs 95%	Picterus underestimated high TSB, reducing sensitivity; optimisation needed for darkly pigmented newborns.

TSB, total serum bilirubin; TcB, transcutaneous bilirubin; r , Pearson correlation unless stated; ρ , Spearman correlation; ICC, intraclass correlation coefficient; AUC, area under the receiver operating characteristic curve; RMSE, root-mean-square error; CI, confidence interval. Sensitivity and specificity depend on each study's own threshold or decision rule.

Recent work has explored multilayer perceptrons (MLPs), convolutional neural networks (CNNs), transfer learning and transformer-based models [13–15]. Two recent deep-learning studies illustrate both the promise and the limits of this approach. Althnani et al. [14] compared VGG-16 transfer learning with four conventional classifiers (MLP, SVM, decision tree and random forest) using forehead-skin, scleral and fused-feature images from neonates of 35–42 weeks' gestation and 0–5 days of age, photographed with a smartphone and a colour card at a single hospital in Saudi Arabia. Of 100 enrolled neonates, 68 samples remained after face-detection failures. Infants were labelled healthy or jaundiced from transcutaneous bilirubin readings rather than TSB, and models were evaluated by five-fold cross-validation. Transfer learning performed best with skin images, whereas the conventional models performed best with eye or fused features, and the best transfer-learning model did not differ significantly from the best MLP model. Lotfi et al. [15] trained a Tokens-to-Token Vision Transformer (T2T-ViT) on 792 smartphone skin images from 500 neonates at a single centre in Iran, acquired with one device under controlled lighting and a colour card, and classified infants as jaundiced or not using a fixed TSB threshold of 15 mg/dL. On a held-out set of 158 images, the T2T-ViT reported accuracy of 0.994 (95% CI 0.965–0.999), sensitivity of 0.984 (0.912–1.000) and AUC of 0.99, compared with accuracy of 0.854 and sensitivity of 0.787 for ResNet-50, although the authors' own validation curves plateaued near 80% accuracy. Both studies produced binary classifications rather than continuous bilirubin

estimates, and neither included independent external validation. In the second study, patient-level separation of training and test images was not stated; because 792 images derived from 500 infants, images of the same infant may have appeared in both sets. The authors also noted the single centre and device, a single 80/20 split without cross-validation, manual region-of-interest selection and a threshold not adjusted for postnatal age. These findings demonstrate feasibility, but with small samples, binary labels and single-site designs they do not establish clinical performance. Greater model complexity does not in itself establish stronger clinical validity. These methods still depend on sufficient numbers of unique patients, patient-level separation of development and test data, external validation and representative study populations.

Taken together, these approaches have developed in parallel rather than as a strict progression, spanning digital photography, calibrated colourimetry, machine learning, optical instrumentation, multi-site AI and deep learning, with prospective clinical evaluation still limited.

5. Artificial Intelligence Methods for Smartphone-Based Neonatal Jaundice Screening

Early systems generally extracted hand-crafted colour features and then applied statistical regression. Features included mean RGB values, colour ratios, HSV and Lab components, yellowness measures and differences from calibration references. Reported modelling approaches include linear and polynomial regression, support vector regression, random forests and gradient-boosted trees. [1][11]

Hand-crafted features can be inspected relatively easily, and the associated models may require modest computational resources. Their reliability nevertheless depends on whether the extracted colours remain stable. Illumination, camera characteristics, anatomical site and acquisition conditions can all alter the relationship between these features and bilirubin.

The choice of modelling method should reflect the amount and structure of the available data, the characteristics of the images and the intended clinical task. Feature-based approaches, such as those used in smartphone colourimetry, allow investigators to examine how selected colour measurements contribute to the estimate. Ensemble methods can combine information from several anatomical sites, as demonstrated by BiliSG [11]. Deep-learning methods, including convolutional neural networks and transfer-learning models, offer an alternative by learning image representations directly from the training data [13–15]. However, their performance depends on the diversity and quality of the training images and on how the evaluation data are separated from the development data. A more complex architecture should therefore not be regarded as evidence of better clinical performance unless the improvement is demonstrated in an appropriate independent evaluation.

Data separation is particularly important. Images from one infant should not be distributed across both training and test sets, because patient-level information leakage can inflate apparent performance. Likewise, performance on images acquired with one smartphone or at one clinical site cannot be assumed to transfer to another device or population without testing.

The move from conventional machine learning to deep learning should not be read as a direct progression toward greater clinical reliability. Model architecture is only one determinant of performance; data quality, validation design and population diversity remain central.

Comparisons between AI methods therefore require more than a ranking of correlation coefficients or AUC values. The training population, patient-level data separation, reference standard, evaluation threshold and external validation design also matter. When these features differ substantially, reported performance is best compared descriptively rather than used to rank algorithms directly.

6. Validation and Performance Across Populations

Performance estimates vary across smartphone-based jaundice studies, and direct comparison is complicated by differences in population, device, anatomical site, bilirubin distribution, reference measurement and evaluation criteria. Correlation with TSB measures association, whereas threshold sensitivity, specificity, agreement and referral performance address different questions. Bilirubin concentrations are reported in mg/dL and $\mu\text{mol/L}$; the conversion is provided in Section 3.

BiliCam, evaluated in 530 newborns across seven US sites, reported a correlation of 0.91 with TSB [1]. The colour-calibrated study of 302 newborns reported $r = 0.84$, with 100% sensitivity and 69% specificity for values above 250

$\mu\text{mol/L}$ [6]. In a multi-ethnic neonatal population, Biliscan showed $r = 0.54$, sensitivity of 76.92% and specificity of 70.83% for detecting jaundice requiring phototherapy [7]. In a separate preterm cohort of infants born before 35 weeks, the ICC between Biliscan and TSB was 0.512 [8]. These results do not support assuming that performance in term newborns will carry over to preterm infants.

Picterus results also vary by evaluation setting. In 416 infants from Mexico, Nepal and the Philippines, correlation with TSB was 0.76 [9]. Among 177 term newborns in Mexico, the Spearman correlation was $p = 0.68$, with sensitivity of 85.7% and specificity of 80.1% at the specified threshold [10]. The reported tendency to underestimate higher bilirubin values matters clinically because underestimation near a treatment threshold may result in missed referrals. During temporal validation, BiliSG reported Pearson $r = 0.84$, $\text{AUC} = 0.89$, sensitivity of 100% and specificity of 70% (predicted $\geq 13 \text{ mg/dL}$ used to detect TSB $\geq 17 \text{ mg/dL}$) [11]. A Taiwanese study of 341 newborns reported $R = 0.90$, $\text{RMSE} = 2.44 \text{ mg/dL}$ and $\text{AUC} = 0.963$ [13]. These findings show what was achieved in the respective study settings; they do not replace evaluation in independent populations and clinical environments.

Evidence from populations with darker skin pigmentation provides another reason to avoid assuming uniform performance. In Botswana, among 123 term newborns, Picterus had a correlation of 0.66 with TSB compared with 0.84 for TcB. At TSB $\geq 250 \mu\text{mol/L}$, sensitivity was 64% for Picterus and 95% for TcB [16]. This comparison highlights the need to assess screening performance in the populations in which a system is intended to be used.

Population characteristics alone cannot explain all differences between studies. Recruitment and study design, bilirubin distribution, image-acquisition protocols, device characteristics, timing of reference measurements and definitions of a positive screen may also influence reported results. Cross-study interpretation should account for both the performance measure and the conditions under which it was obtained. Subgroup results deserve particular attention when underestimation has different consequences for different clinical risk groups.

Overall, the studies support the feasibility of extracting bilirubin-related information from smartphone images, but they do not establish uniform screening performance across settings. Differences in population, bilirubin distribution, image acquisition and decision thresholds must be considered, particularly where false-negative results could delay confirmatory assessment.

7. Image Quality, Calibration and Acquisition Conditions

Reliable smartphone-based bilirubin assessment begins with image acquisition. Images are affected by ambient illumination, camera hardware, exposure, white balance, computational photography, distance and anatomical positioning. Motion blur, shadows, reflections and poor focus can further compromise the input to a prediction model.

Calibration can reduce some of the variation introduced during image acquisition, particularly when colour-reference cards or controlled optical accessories are used. In the study of 302 newborns, colour-calibrated smartphone estimates correlated with TSB at $r = 0.84$, with 100% sensitivity and 69% specificity for TSB above $250 \mu\text{mol/L}$ [6]. These findings support further investigation of calibrated imaging, but they do not establish that calibration alone ensures reliable measurements across different smartphones and lighting conditions. Its contribution should be evaluated alongside image quality, device characteristics and performance in independent populations.

ROI selection also matters. Studies have used the sternum, forehead, chest and abdomen, and the anatomical site can affect the observed relationship with bilirubin. A system intended for routine use therefore needs clear acquisition instructions and a plan for cases in which the preferred site cannot be captured consistently.

Image quality is best treated as part of the measurement system, not merely as a preprocessing step. Automated quality checks, rejection of unsuitable images and repeat acquisition can help prevent unreliable images from reaching the bilirubin prediction stage.

8. Skin Pigmentation, Preterm Infants and Population Diversity

Population diversity is central to evaluating smartphone-based jaundice systems. Skin pigmentation may affect the relationship between recorded colour and bilirubin-related optical characteristics. Gestational age and clinical status also affect bilirubin distributions and may influence whether a model developed in term newborns is appropriate for other groups.

The Botswana evaluation, in which Picterus showed lower correlation and sensitivity than TcB among 123 term newborns with darker pigmentation [16], illustrates the value of subgroup-specific assessment. Similarly, the ICC of

0.512 between Biliscan and TSB in infants born before 35 weeks [8] indicates that performance in preterm infants requires dedicated study.

In contrast, BiliCam [1] reported similar correlations across ethnic groups (white 0.92, African American 0.90, Hispanic 0.91, Asian American 0.88); such findings are specific to the population and device studied and do not establish equivalent performance elsewhere.

Future evaluations should define relevant subgroups in advance and report sample sizes, measurement errors and threshold-specific performance where the data allow. Overall estimates should be accompanied by uncertainty measures, because aggregate performance can conceal clinically important subgroup differences. If subgroup samples are too small to support reliable conclusions, that limitation should be stated rather than interpreted as evidence of equivalent performance.

9. From Measurement Accuracy to Clinical Screening

Across the literature, measurement accuracy and clinical usefulness emerge as separate questions. A model can correlate strongly with TSB yet provide insufficient information for safe referral. A classifier can also perform well in its development population without evidence that its use improves clinical outcomes.

Four evaluation levels are therefore useful to distinguish: measurement, screening, referral and clinical impact. Measurement studies assess agreement with a reference bilirubin value; screening studies test identification of clinically relevant thresholds; referral studies examine whether infants are directed appropriately to confirmatory assessment; and clinical-impact studies assess whether the technology changes care processes or outcomes.

This distinction is particularly consequential when high bilirubin values are underestimated. Even reasonable average agreement may be inadequate if a model misses infants near a clinically important threshold. Evaluation should include sensitivity, specificity, false-negative rates and threshold-focused analyses, rather than relying on correlation or overall accuracy alone.

For systems intended to guide referral, evaluation should also describe false-negative results around prespecified thresholds, the number of infants sent for confirmatory testing and the consequences of incorrect classification. Predictive values depend on the prevalence of clinically significant hyperbilirubinemia in the evaluated population. These measures characterize practical screening performance that cannot be inferred from correlation alone.

10. Datasets and Benchmarking Challenges

Dataset quality shapes the reliability of AI models. Studies vary in the number of unique infants, number of images, participating sites, smartphone models, reference methods and validation procedures. Image counts can overstate the effective dataset size when many images come from relatively few infants.

A useful dataset should link each image with patient characteristics, device and acquisition metadata, a temporally matched TSB or other reference measurement, anatomical site, image-quality information and relevant clinical outcomes. Gestational age, pigmentation-related characteristics and bilirubin distribution should also be represented.

Patient-level data separation is essential: images from the same infant should not occur in both training and test sets. Temporal and external-site validation provide stronger evidence of generalization, and cross-device testing is relevant because smartphone cameras and computational-photography pipelines differ.

The reference outcome also needs to be defined carefully. TSB and TcB are not interchangeable standards, and the interval between image capture and bilirubin measurement should be reported. Clinical thresholds should be specified in advance and interpreted in relation to age and gestational age.

A meaningful benchmark should report mean absolute error (MAE), RMSE, correlation, Bland–Altman agreement, threshold sensitivity and specificity, AUC where appropriate, subgroup performance and image-failure rate. Correlation or accuracy alone gives an incomplete account of model performance.

For n infants with smartphone-estimated bilirubin $TS\hat{B}_i$ and reference TSB_i , the mean absolute error (MAE) and root-mean-square error (RMSE) are given by Equations (4) and (5), respectively:

$$MAE = (1/n) \sum_{i=1}^n |TS\hat{B}_i - TSB_i| \quad (4)$$

$$RMSE = \sqrt{(1/n) \sum_{i=1}^n (TS\hat{B}_i - TSB_i)^2} \quad (5)$$

For threshold classification, sensitivity and specificity are defined by Equations (6) and (7):

$$\text{Sensitivity} = TP/(TP + FN) \quad (6)$$

$$\text{Specificity} = TN/(TN + FP) \quad (7)$$

Here, TP, FN, TN and FP denote true positives, false negatives, true negatives and false positives relative to a prespecified TSB threshold. Because underestimation near a treatment threshold is clinically dangerous, the false-negative rate, $FN/(TP + FN)$, should be reported explicitly. Agreement is usually summarized by Bland–Altman analysis, in which the bias and 95% limits of agreement are given by Equation (8):

$$\text{Bias} = \text{mean}(d_i); \text{ 95\% limits of agreement} = \text{bias} \pm 1.96 \times \text{SD}(d_i) \quad (8)$$

Here, $d_i = \text{TS}\hat{B}_i - \text{TSB}_i$ is the difference between the smartphone estimate and the reference TSB for infant i , $\bar{d}$ is the mean difference (the bias), and $\text{SD}(d_i)$ is the standard deviation of the differences.

A practical minimum dataset structure can be represented as Equation (9):

$$D = \{I_i, P_i, M_i, \text{TSB}_i, A_i, V_i, O_i\} \quad (9)$$

In Equation (9), D represents the dataset record; I_i denotes image data, P_i patient information, M_i acquisition metadata, TSB_i the reference bilirubin measurement, A_i annotations, V_i validation information and O_i clinical outcomes. Standardized reporting using such a structure could improve reproducibility and facilitate comparison between systems.

To support meaningful benchmarking, publications should distinguish the number of unique infants from the number of images and observations. They should also report missing data, image exclusions, repeated measurements, the timing of reference measurements and the method used to separate development and evaluation datasets. This information helps readers assess whether reported performance reflects generalization or may be influenced by repeated observations or selective exclusion of images.

11. Technical, Clinical and Implementation Challenges

The limitations reported across the literature are linked and can be grouped into five areas: optical variability, clinical variability, AI generalization, workflow integration and governance.

Optical variability arises from illumination, camera hardware, exposure and white balance, computational photography, imaging distance, anatomical site and image quality. Calibration and controlled illumination may reduce some of these effects, but should not be assumed to eliminate them.

Clinical variability includes skin pigmentation, gestational age, bilirubin distribution, preterm status and treatment conditions such as phototherapy. Underestimation at high bilirubin concentrations is especially concerning because a false-negative screening result may delay confirmatory assessment.

The reliability of an AI model may decline when the data encountered in practice differ from those used during development. Differences in smartphone hardware, image-processing pipelines, patient characteristics and acquisition conditions can create a shift between the development and deployment settings. Small numbers of unique infants and the use of images from the same infant in both training and test sets can further inflate apparent performance. Evaluation should therefore use patient-level data separation and, where feasible, independent-site or prospective validation. The frequency of rejected images and uncertain predictions should also be reported because these affect the proportion of infants for whom the system can provide a usable result.

Clinical deployment requires more than a prediction algorithm. A practical pathway is:

image acquisition → quality check → prediction → uncertainty → referral → confirmatory bilirubin measurement → management.

Connectivity, user training, privacy, parental acceptance and access to referral services can influence whether technical performance translates into clinical usefulness. Feasibility, qualitative and parent-use studies have begun to examine these workflow issues [17–19].

Safe deployment requires fail-safe procedures. Poor-quality images should be rejected, uncertain or out-of-distribution cases escalated, and high-risk threshold crossings followed by confirmatory assessment. A screening estimate should not be presented as a definitive diagnosis.

Implementation studies should assess the complete screening pathway, not just the prediction. Relevant outcomes include successful image acquisition, repeat-image frequency, referral completion, access to confirmatory testing, user acceptability and the time needed to obtain a usable result. These measures help establish whether technical performance is maintained in routine clinical or community settings.

Post-deployment monitoring also matters because performance may change as devices, populations and acquisition environments change. Systems intended for routine use require controlled updates and clear clinical governance. Regulatory status also differs between systems; for example, the Picterus system was developed iteratively and underwent conformity assessment for CE certification after a clinical validation study of healthy term newborns aged

1–15 days [20]. Systems intended for clinical use require the clearance appropriate to their intended jurisdiction and claims.

12. Research Gaps and Future Directions

The available evidence points to a shift from proof-of-concept image analysis toward more carefully validated digital screening. Several questions remain unresolved across the studies reviewed.

The next stage of research should address three priorities. First, smartphone-based systems need evaluation using multicentre datasets that represent different smartphone models, skin pigmentation groups, gestational ages and bilirubin distributions. The number of unique infants, image exclusions and timing of reference TSB measurements should be reported so that the effective sample size and measurement conditions can be assessed. Second, validation should examine performance around prespecified clinical thresholds, particularly the frequency of false-negative results and the extent of underestimation at higher bilirubin concentrations. Independent-site and cross-device testing would help determine whether performance is maintained beyond the development setting. Third, prospective studies should evaluate the complete screening pathway, including image acquisition, repeat measurements, referral completion and access to confirmatory testing. These outcomes are necessary to determine whether an apparently promising image-based estimate can be used safely in practice.

Finally, decentralized use raises questions about edge AI, privacy-preserving learning and secure data governance. A staged development pathway is therefore appropriate:

standardized data development → robust AI development → external and prospective validation → workflow/home validation → health-system impact evaluation.

13. Limitations of the Evidence and Review

The evidence base has important limitations. Many studies are single-centre investigations or include relatively small samples. Their populations differ in gestational age, bilirubin distribution, skin pigmentation, imaging conditions and clinical thresholds. Selected participants or controlled acquisition conditions may yield higher performance estimates than routine screening would produce. Differences in reference standards, anatomical sites, smartphone models, calibration and outcome definitions also limit direct numerical comparison.

This review has limitations as well. It is a focused narrative synthesis, not a PRISMA-compliant systematic review, and it does not claim an exhaustive study count or a formal risk-of-bias assessment for every publication. Quantitative results are taken from individual studies or published systematic reviews and should be interpreted within their original clinical and methodological settings. Although studies from 2025–2026 broaden the review's currency, some newer reports have limited follow-up or implementation experience.

The available evidence is therefore more useful for identifying current capabilities and unresolved challenges than for concluding that smartphone systems can replace established bilirubin measurement. Prospective, multicentre and externally validated studies are still needed to establish how reported performance translates to clinical workflows.

14. Conclusion

Smartphone imaging offers a practical approach to investigating non-invasive neonatal jaundice screening, and published studies demonstrate associations between image-derived features and TSB. However, the reported results vary across devices, populations and evaluation methods. The evidence reviewed does not yet justify treating smartphone estimates as interchangeable with established bilirubin measurements, particularly where high bilirubin concentrations may be underestimated or where performance in preterm infants and diverse pigmentation groups remains uncertain.

Further progress depends on independent validation, reliable image acquisition, threshold-specific evaluation and prospective assessment of referral pathways. Studies should report not only measurement error and correlation but also false-negative results, image failures and the proportion of infants who complete confirmatory assessment. Until these outcomes are established in the populations and settings for which the technology is intended, smartphone-based assessment should remain a supplementary screening approach rather than a replacement for established bilirubin measurement and clinical decision-making.

References

1. Taylor JA, Stout JW, de Greef L, et al. Use of a Smartphone App to Assess Neonatal Jaundice. *Pediatrics*. 2017;140(3):e20170312. doi:10.1542/peds.2017-0312. PMID:28842403.
2. Hegde S, Rath C, Amarasekara S, Saraswati C, Patole S, Rao S. Performance of smartphone application to accurately quantify hyperbilirubinemia in neonates: a systematic review with meta-analysis. *Eur J Pediatr*. 2023;182(9):3957-3971. doi:10.1007/s00431-023-05073-2. PMID:37368007.
3. Kemper AR, Newman TB, Slaughter JL, Maisels MJ, Watchko JF, Downs SM, et al. Clinical Practice Guideline Revision: Management of Hyperbilirubinemia in the Newborn Infant 35 or More Weeks of Gestation. *Pediatrics*. 2022;150(3):e2022058859. doi:10.1542/peds.2022-058859.
4. World Health Organization. WHO recommendations on maternal and newborn care for a positive postnatal experience. Geneva: World Health Organization; 2022. ISBN: 978-92-4-004598-9. Available from: <https://www.who.int/publications/i/item/9789240045989>.
5. Outlaw F, Nixon M, Odeyemi O, MacDonald LW, Meek J, Leung TS. Smartphone screening for neonatal jaundice via ambient-subtracted sclera chromaticity. *PLoS One*. 2020;15(3):e0216970. doi:10.1371/journal.pone.0216970. PMID:32119664.
6. Aune A, Vartdal G, Bergseng H, Randeberg LL, Darj E. Bilirubin estimates from smartphone images of newborn infants' skin correlated highly to serum bilirubin levels. *Acta Paediatr*. 2020;109(12):2532-2538. doi:10.1111/apa.15287. PMID:32267569.
7. Ngeow AJH, Tan MG, Dong X, Tagamolila V, Ereno I, Tay YY, Xin X, Poon WB, Yeo CL. Validation of a smartphone-based screening tool (Biliscan) for neonatal jaundice in a multi-ethnic neonatal population. *Journal of Paediatrics and Child Health*. 2023;59(2):288-297. doi:10.1111/jpc.16287.
8. Ustun H, Oncel MY, Kefeli M, Akar M, Engür D. Bilirubin Measurement Through a Smartphone Application in Preterm Infants. *Klin Padiatr*. 2023;235(5):265-269. doi:10.1055/a-1816-5878. PMID:36539195.
9. Aune A, Jiménez-Díaz G, Gierman LM, Vartdal G, Reyes Berlanga M, Tusoy J, Bergseng H, Shakya S, Darj E. Smartphone-based screening of neonatal jaundice in three populations in low and middle-income countries: a cross-sectional study. *BMJ Paediatrics Open*. 2025;9(1):e002242. doi:10.1136/bmjpo-2023-002242.
10. Jiménez-Díaz G, Gierman LM, Keitsch M, Elizarrarás-Rivas J, Silva-Méndez RM, Infanti JJ, Marcuzzi A. Validation of a Mobile Health Device for Neonatal Jaundice Screening: A Cross-Sectional Study in a Resource-Limited Setting in Mexico. *SAGE Open Pediatr*. 2025;12:30502225251320544. doi:10.1177/30502225251320544. PMID:40612198.
11. Ngeow AJH, Moosa AS, Tan MG, Zou L, Goh MMR, Lim GH, et al.; BiliSG Study Group. Development and Validation of a Smartphone Application for Neonatal Jaundice Screening. *JAMA Netw Open*. 2024;7(12):e2450260. doi:10.1001/jamanetworkopen.2024.50260. PMID:39661385.
12. Harrison-Smith B, Dumont AP, Arefin MS, Sun Y, Lawal N, Dobson D, Nwaba A, Grossarth S, Paed AM, Farouk ZL, Weitkamp JH, Patil CA. Development of a mobile phone camera-based transcutaneous bilirubinometer for low-resource settings. *Biomedical Optics Express*. 2022;13(5):2797-2809. doi:10.1364/BOE.449625.
13. Shen GT, Liu YH, Lin JY, Hsu KC, Chou HH. Development and validation of neonatal jaundice detection using smartphone application. *Comput Biol Med*. 2025;196(Pt C):110845. doi:10.1016/j.compbimed.2025.110845. PMID:40784079.
14. Althnian A, Almanea N, Aloboud N. Neonatal Jaundice Diagnosis Using a Smartphone Camera Based on Eye, Skin, and Fused Features with Transfer Learning. *Sensors (Basel)*. 2021;21(21):7038. doi:10.3390/s21217038. PMID:34770345.
15. Lotfi M, Rabiee M, Hagbini Nazarpak M, et al. Neonatal jaundice detection using a vision transformer-based deep learning model. *Sci Rep*. 2026;16:9243. doi:10.1038/s41598-026-40515-5. PMID:41699004.
16. Zimmer J, Hansen BM, Peswa L, Ditiwira A, Otukile N, Bratlid D, Gierman LM, Nakstad B; Botswana Picterus Study Group. Noninvasive Assessment of Bilirubin Levels in Newborn Infants With Dark Skin Pigmentation. *Acta Paediatr*. 2026;115(5):1092-1098. doi:10.1111/apa.70454. PMID:41589509.
17. Jiménez-Díaz G, Elizarrarás-Rivas J, Keitsch M, Gierman LM, Marcuzzi A, Infanti JJ. Advancing neonatal jaundice screening with mHealth: a mixed-methods randomised controlled feasibility study in a resource-limited setting in Mexico. *BMJ Paediatr Open*. 2026;10(1):e003624. doi:10.1136/bmjpo-2025-003624. PMID:41927322.
18. Terpstra FJ, Hazan A, Kohn E, Berkovitch M, de Vries TW, Been JV, Dijk PH, Hulzebosch CV. Evaluating the Picterus Jaundice Pro App for Neonatal Jaundice Screening by Parents: Prospective Mixed Methods Study. *JMIR Pediatr Parent*. 2026;9:e83077. doi:10.2196/83077. PMID:42508005.

19. Jiménez-Díaz G, Aune A, Elizarrarás-Rivas J, Gierman LM, Keitsch M, Marcuzzi A, Infanti JJ. Neonatal jaundice detection in low-resource Mexican settings: possibilities and barriers for innovation with mobile health. BMC Health Serv Res. 2024;24:671. doi:10.1186/s12913-024-11141-6. PMID:38807158.
20. Aune A, Vartdal G, Jimenez Diaz G, Gierman LM, Bergseng H, Darj E. Iterative Development, Validation, and Certification of a Smartphone System to Assess Neonatal Jaundice: Development and Usability Study. JMIR Pediatr Parent. 2023;6:e40463. doi:10.2196/40463. PMID:36853753.