International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(6s) 2026 pp. 597-608
https://doi.org/10.70917/ijcisim-2026-2959
Article

Towards Robust Non-Invasive Neonatal Jaundice Detection: Problem Formulation and Analysis of Real-World Challenges

Saurabh Gupta[1], Karthik Kovuri[2]

[1]Department of Computer Applications, RIMT University, Mandi Gobindgarh, Punjab, India.
Email: g.saurabh.gupta@gmail.com
[2]Academic Affairs, RIMT University, Mandi Gobindgarh, Punjab, India.
Email: dean.academics@rimt.ac.in

Abstract: Neonatal jaundice (neonatal hyperbilirubinemia) appears in 60–80% of newborns around the world and is one of the leading causes of neonatal death and of neurological disorders due to a lack of treatment in many end, and even some developing, countries, especially of those in sub-Saharan Africa and South Asia. For the majority of laboratories, the gold standard for diagnosis is invasive testing of serum bilirubin. It is painful and expensive and is often not an option in developing countries. Clinical research on non-invasive testing, such as the use of transcutaneous bilirubinometry (TcB) or the use of deep learning systems and imaging-based systems, is promising. However, these techniques tend to be limited by five interrelated practical challenges: (1) the effect of melanin and the associated skin tone bias, and the resulting inequity in diagnosis across the different levels of the Fitzpatrick scale, (2) the small and homogeneous datasets, as the majority of the published deep learning solutions have been trained on fewer than 300 neonates from a specific geographic location, (3) the systems are not robust to variations in real-life imaging conditions such as different lighting, cameras or imaging strategies, (4) all systems describe a lack of research that validates the internal imaging and segmentation of the body, and (5) a lack of thorough research in different imaging and segmentation methods. This article is a systematic analysis of all records currently published. We identify the limitations of current systems, and in relation to those failures, we assign performance tiers. We have created PPC (Perceptual Pre-processing Correction) as a framework to solve the issues, and from that framework, we created new focused research questions. The framework contains several original inventions, including the design of a biased skin tone melanin color norm spotting Convolutional Neural Network (ST-CNN) method, the CZZBP (Colour-Zone-based Z-score Bilirubin Prediction) segmentation method, and the melanin-informed color normalization method.

Keywords: neonatal jaundice, hyperbilirubinemia, non-invasive bilirubin estimation, skin tone bias, deep learning, convolutional neural networks, image segmentation, transcutaneous bilirubinometry, melanin interference, algorithmic fairness, and low-resource settings.

1. Introduction

Neonatal jaundice, which is the hyperbilirubinemia seen within the first month of life in humans, is one of the most common conditions that a health practitioner in the world will encounter with the care of neonates. Neonatal jaundice becomes apparent in the skin, sclera, mucous membranes, and a neonate will present with the deposition of bilirubin causing jaundice in the skin of 60% of term neonates and 80% of pre-term neonates within the first few days of life and can be seen with the care of neonates in the first month of life. In most cases, it is benign and physiological and will spontaneously resolve in one to two weeks. However, when total serum bilirubin (TSB) exceeds critical values, especially if they cross 25mg/dl, unconjugated bilirubin will cross the blood-brain barrier, and the neonate will develop acute bilirubin encephalopathy. In severe cases, the long-term complications will be the neurological sequelae that include the development of cerebral palsy and sensorineural deafness, and the dysfunction of the ocular pathways.

Kernicterus is an avoidable condition that can cause developmental disabilities in newborns and is identified by the WHO as an emerging problem in low- and middle-income countries due to early hospital discharge, limited access to labs, and lack of neonatal medical professionals (*WHO Recommendations on Maternal and Newborn Care for a Positive Postnatal Experience*, 2022). Jaundice is the cause for 15-20% of newborns admitted to neonatal

intensive care units globally (Lin et al., 2022) and most cases are in sub-Saharan Africa and South Asia, the regions with the least coverage for diagnostic research.

The most common diagnostic technique is to measure serum bilirubin by an invasive heel-prick or venipuncture. Bilirubin can reach dangerous levels and exceed the intervention threshold if the lab work is delayed for several hours, which is not uncommon. Additionally, neonatal heel-pricks can cause a lot of pain and discomfort, especially to pre-term neonatal babies. ((Deora & Khanal, 2023); (Joshi & Singhvi, 2024)). These limitations have motivated two decades of research into non-invasive alternatives: transcutaneous bilirubinometry (TcB), digital colour photometry, and, most recently, smartphone-integrated deep learning systems.

Despite genuine technological progress - from Kramer's 1969 cephalocaudal visual assessment model through first-generation TcB devices to deep convolutional neural networks - no published system has demonstrated clinically acceptable, equitable diagnostic performance across the full spectrum of neonatal skin tones in real-world imaging environments. The field is characterised by a structural disconnect: systems evaluated under laboratory-controlled conditions consistently fail when deployed in the diverse lighting, camera, and demographic contexts of clinical and community practice.

This paper presents a rigorous problem formulation for non-invasive neonatal jaundice detection, grounded in a critical synthesis of the existing literature spanning clinical visual assessment, transcutaneous bilirubinometry, classical machine learning, and deep learning approaches. We identify and analyse five interrelated real-world challenges and systematically document the corresponding model failure modes. From this foundation, we derive five research questions, five study objectives, and five original research contributions constituting an integrated methodological framework. This work argues that progress in this domain requires not merely better algorithms, but a principled reformulation of the problem itself.

2. Background and Related Work

2.1 Clinical and Physiological Foundations

The photophysical principle underlying all non-invasive bilirubin measurement relies on the optical absorption of bilirubin, which has a strong absorption peak in the blue-green range (450–460 nm). This peak absorption gives bilirubin its yellow color, evident in jaundiced skin and scleral tissues. When polychromatic light penetrates skin tissues, it is scattered and absorbed by many chromophores present in the skin, such as melanin, oxyhemoglobin, deoxyhemoglobin, bilirubin, water and lipid. The profile of spectral reflectance contains information about the concentration of the chromophores and allows non-invasive estimation of bilirubin (Schumacher, 1990).

One of the first clinical models for assessing jaundice was (Kramer, 1969) model of cephalocaudal progression. This model posits that as serum bilirubin concentration increases, jaundice advances from the scalp to the tips of the extremities, across five distinct dermal zones. As a first approximation, Kramer's model is useful, but its limitations are well recognized. There is poor agreement among clinicians to identify Kramer's five dermal zones (kappa 0.3–0.4). Furthermore, the model is not sensitive to bilirubin concentrations in the range of 12–15 mg/dL, the clinically significant threshold, and loses reliability in jaundiced neonates with a dark pigmentation (Watchko, 2009).

The most significant challenge of photophysical measurement that most non-invasive methods of measurement (TcB) face is that melanin, which is the primary skin pigment, strongly absorbs all visible light, and the absorption increases in the shorter wavelengths. This includes the peak absorption of melanin which directly overlaps with bilirubin’s diagnostic peak at 450 nm, and creates a fundamental confounding problem. This was addressed by Hannemann et al (1978) through the use of 460 nm and 550 nm reflectance measurements, and they found bilirubin to have a correlation of $r = 0.78$. This also established the multi-wavelength correction for all subsequent Transcutaneous Bilirubin (TcB) measurement technology.

Figure 1 identifies the five research gaps and illustrates the relationships among these gaps. Each of these gaps interacts with the others, and therefore, cannot be resolved independently.

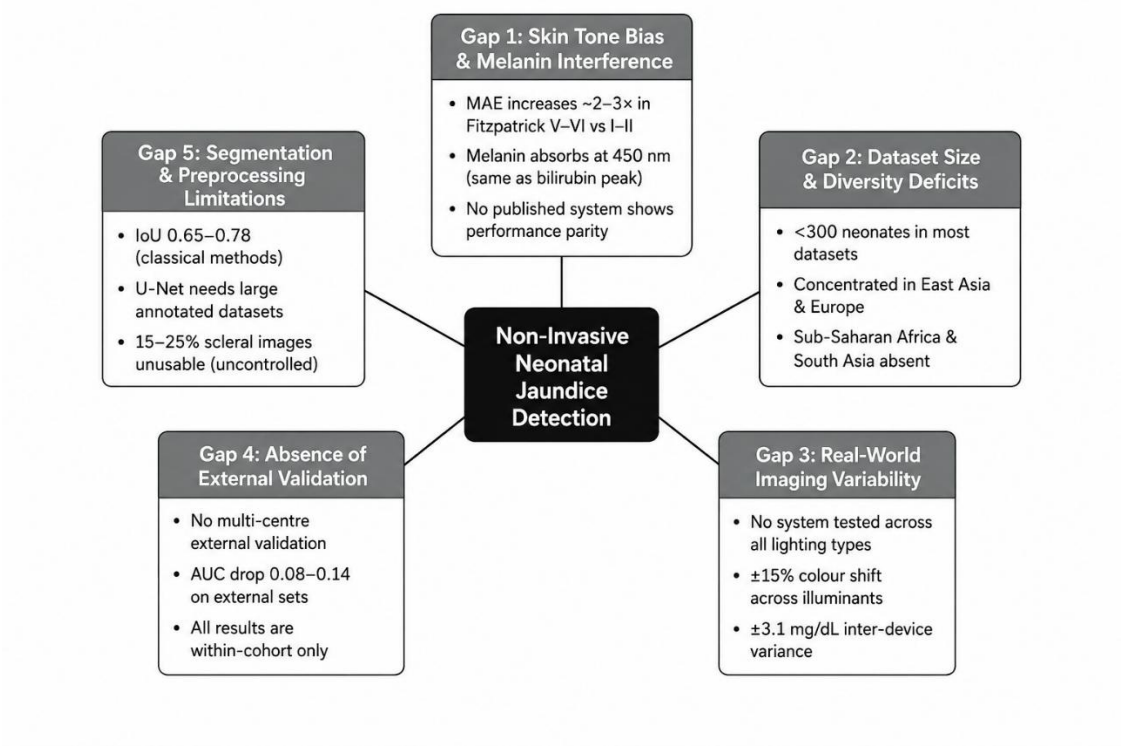

Figure 1 : Research Gap Mind Map - Five Interrelated Gaps in Non-Invasive Neonatal Jaundice Detection

2.2 Transcutaneous Bilirubinometry: Technology and Documented Skin Tone Bias

In clinical practice, the most validated non-invasive bilirubin measurement technique is transcutaneous bilirubinometry (TcB). These devices apply multi-wavelength diffuse reflectance spectroscopy post melanin and haemoglobin correction of bilirubin estimation. Techniques have made improvements of subsequent generations as seen with the correlational strength of light presence in skin of the first generation Minolta Air-Shields device (r = 0.80-0.88), the JM-102/103 at r = 0.88-0.93, and the 95% limits of agreement of the Philips BiliCheck at ±3.2 mg/dL versus ± 4.1 mg/dL of the JM-103 ((Jangaard et al., 2006). The TcB is supported by the postnatal wards and NICUs of the American Academy of Pediatrics (AAP, 2004) and UK NICE guidelines for its screening practice.

However, there is a notable and clinically meaningful exception: accuracy decreases as skin pigmentation increases. (Engle et al., 2002) conducted a large, multi-ethnic neonatal study of 604 participants. They found that the Minolta JM-103 TSB meter systematically overestimated total serum bilirubin (TSB) in lightly pigmented neonates and underestimated TSB in darkly pigmented neonates, with 95% limits of agreement widening from ±3.8 to ±5.9 mg/dL within the differing skin tone groups. (Mishra et al., 2010) found that overall, Indian neonates had reasonable TcB correlation (r = 0.87) when compared to European neonates, with significant bias and scatter when compared to the European norms. The most alarming finding was that (Olusanya et al., 2016) showed that TcB measures in sub-Saharan African neonates had mean biases ranging from 2.1 to 4.8 mg/dL, indicating a likely clinically significant underdiagnosis of severe hyperbilirubinemia in the regions of the world most affected.

2.3 Classical Image Processing and Machine Learning Approaches

Bilirubin estimation involves constructing multi-stage pipelines that integrate feature extraction and machine learning models with carefully designed stages. These models leverage the CIE 1976 L*a*b* (CIELAB) color space for classification. The b* (blue-yellow) axis of this color space is often the focus of analysis. Research suggests that b* values are positively correlated with total serum bilirubin (TSB) levels for lightly pigmented neonates and are sensitive to jaundice-color yellows, demonstrating values of b* in the range of 0.70–0.85. Constructing support vector machine (SVM) classifiers whose input features are CIELAB color values yielded promising results. Castro-Ramos et al. (2014) reported a threshold of 85% in classifying TSB levels of 12 mg/dL, based on their analysis of 100 Taiwanese neonates.

In lightly pigmented populations, Support Vector Regression (SVR) reports Marginal Absolute Error (MAE) at 1.5–2.8 mg/dL and considerably weakened performance in the darker skin strata due to melanin interference. A very significant study in this area is that of Bhutani et al. (2013). By employing an ANN architecture on the Nigerian population (N = 117), an RMSE of 4.6 mg/dL was reported, which is more than two times that reported for the Japanese population and shows that biases due to higher melanin cannot be overcome from models primarily trained on lighter-skinned populations.

2.4 Deep Learning and CNN-Based Systems

The first smartphone-based jaundice detection system is the BiliScreen system (Mariakakis et al., 2017). It consists of a CNN-based scleral mapping segmentation module (Dice similarity 0.92) combined with Random Forest regression that yielded a Pearson correlation of r = 0.89 with 89.7% sensitivity in an adult sample. (Aune et al., 2020), on the other hand, used a physics-inspired Kubelka-Munk optical model on a sample of 302 neonates and achieved a correlation of r = 0.84 with total serum bilirubin (TSB). (Gupta et al., 2024) reported an AUC of 0.90 with the EfficientNet model and an AUC of 0.87 with the Vision Transformer model, representing the best performance metrics reported to date in the specialization. However, none of these studies reported results of an external validation and skin-tone stratified analyses.

U-Net (Ronneberger et al., 2015) and its variants achieved IoU scores of 0.88–0.95 on neonatal sclerals segmentation benchmarks, considerably outperforming classical threshold-based methods with IoU scores of 0.65–0.78. However, deep learning segmentation models are data hungry, especially with appropriate annotations that are not currently available for diverse neonatal populations and performance is adversely affected in the presence of uncontrolled imaging conditions. Generative Adversarial Networks (Lyra et al., 2023) have been explored to address data scarcity concerns, but there are doubts abount the clinical veracity of synthesized jaundice images.

3. Analysis of Real-World Challenges

A lot of variability and uncertainty arises when developing non-invasive newborn jaundice detection that combines clinical imaging, photophysics, and machine learning. A systematic study has identified five categories of real-world obstacles corresponding to five defined areas of research.

3.1 Skin Tone Variation and Melanin Interference

Human skin colour is influenced by the type and amount of melanin with the two main subtypes being eumelanin (brown/black) and pheomelanin (red/yellow). Notably, melanin has the ability to absorb all the visible spectrum with bilirubin absorbes at the 450–470 nm (Schumacher, 1990) A model specifically adapted to babies with low pigmentation can use colour features to estimate bilirubin. However, when the same model is used with babies with dark skin, the model is going to attribute melanin to bilirubin therefore making systematic errors (Slusher et al., 2017).

The impact, from a quantitative perspective, is significant. The studies summarized in chapter 2 tell us that when moving from Fitzpatrick phototypes I-II (lightest) to V-VI (darkest) the absolute mean error of bilirubin estimation increases by 2-3 times. This bias is evident in TcB devices and CNN-based systems. As seen in Figure 2, in estimation error, TcB devices demonstrate systematic bias, and their mean limits of agreement are wider for darker skin tone categories than the clinically acceptable MAE of 2 mg/dL. TcB devices show systematically increasing bias and wider limits of agreement as skin pigmentation increases.

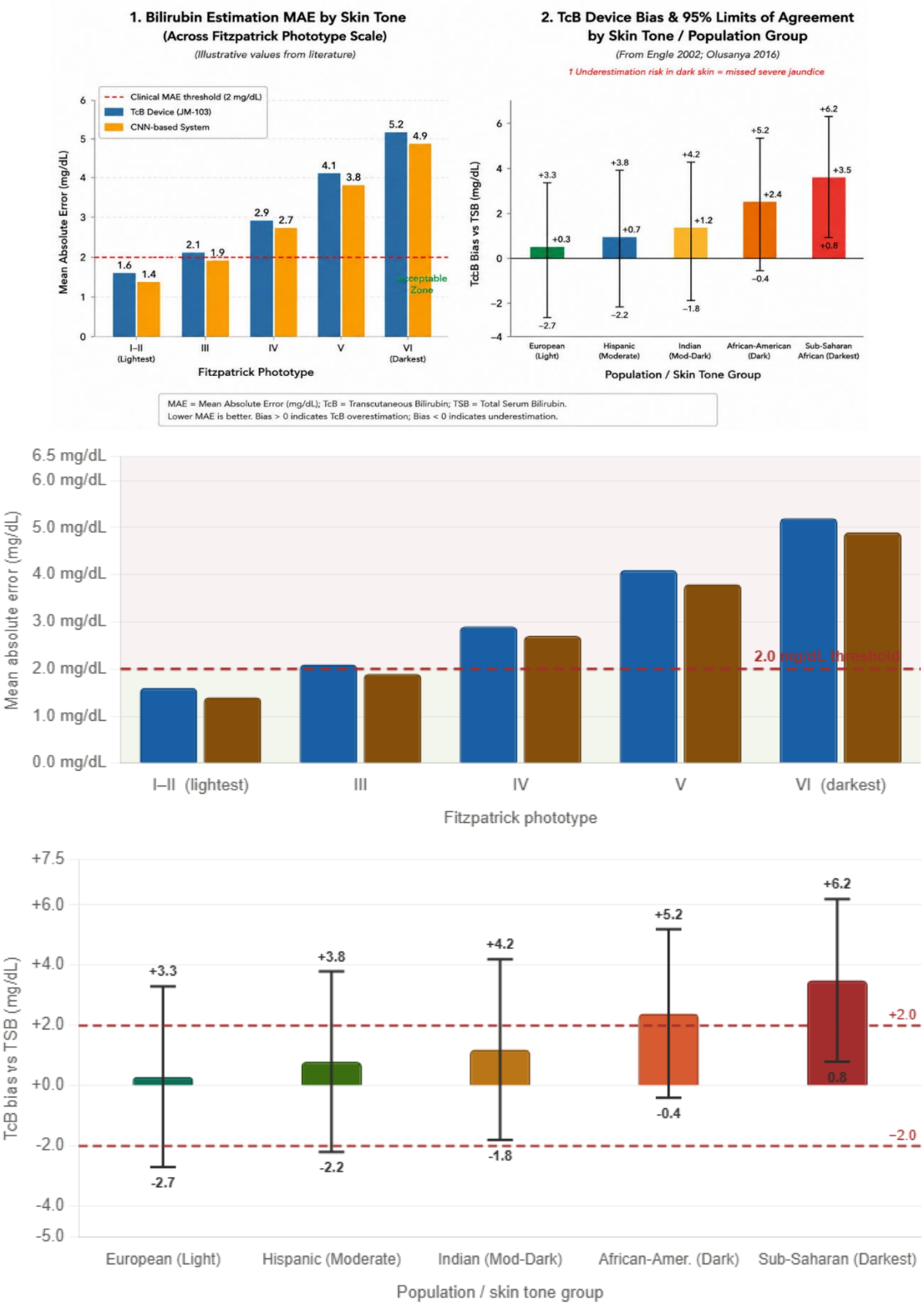

Figure 2: Skin Tone Bias in Non-Invasive Bilirubin Estimation -MAE across Fitzpatrick phototype strata for TcB and CNN-based systems; TcB device bias and 95% limits of agreement by skin tone group (based on Engle et al., 2002; Olusanya et al., 2016)

Among the available systems, no published device has demonstrated performance parity across all Fitzpatrick strata. The individual typology angle (ITA), Fitzpatrick phototyping, and the Monk skin tone scale have been developed as standardized instruments and metrics to measure bias. Their systematic integration in training and evaluation has yet to be applied to neonatal jaundice detection in the literature.

3.2 Dataset Size and Diversity Deficits

Training generalisable models has been especially difficult because of large, diverse, and well-characterised training datasets being in limited supply. Many published studies use datasets with fewer than 300 neonates. Some studies use fewer than 150 labelled image-bilirubin pairs. These are completely inadequate for training high-capacity deep learning models. This results in systematic overfitting and leads to accuracy drops to around 30% on independent external cohorts.

Most published datasets are geographically and demographically concentrated in East Asian and Northern European populations. Published training data are virtually absent in South Asian and sub-Saharan African neonates who account for 78% of the world's severe neonatal jaundice cases (*WHO Recommendations on Maternal and Newborn Care for a Positive Postnatal Experience*, 2022). Lack of representativeness in training data causes systematic inequities in clinical decision support tools (Obermeyer et al., 2019). These findings can also be used in support of neonatal jaundice detection.

3.3 Robustness to Real-World Imaging Variability

Clinical ambient lighting most often employs incandescent tungsten, fluorescent tubes, LED arrays, or even daylight. Each of these lighting options forms a different spectral power distribution. This means that articulated differences will result in images showing different shades of skin for the same infant and bilirubin concentration. Neonates with jaundice pose a specific and unique challenge to white balance algorithms due to the yellow-brown skin colors of neonates with jaundice. This problem, as observed in CNN models for bilirubin estimation, means skin-dependent features can shift by ±15% across various lighting conditions, and the variance of the bilirubin estimate can be as high as ±3.1 mg/dL due to the use of different imaging devices.

3.4 Absence of Prospective Multi-Centre External Validation

External validation should be the minimal acceptable standard for a clinical diagnostic technology to be included in clinical guidelines. Assessment of the literature shows that no published image-based neonate jaundice detection system meets this standard. For systems designed to work across different clinical settings, significant domain shifts across skin tone, imaging device, lighting, and imaging protocols adversely affect the systems' performances with as much as a decrease in AUC of 0.08–0.14.

3.5 Limitations in Image Segmentation and Preprocessing

In order to use features sensitive to bilirubin, accurate segmentation of the conjunctival sclera, skin on the face, nasal skin, as well as skin on the palms and soles, is needed. Classical thresholding methods, due to their dependence on lighting conditions and skin tone, report IoU values of 0.65–0.78.

Deep learning-based techniques (U-Net and their variants) achieve state-of-the-art scores of between 0.88–0.95 in IoU. However, they are heavily reliant on large-scale annotated datasets that are unavailable for heterogeneous neonatal populations. Under unconstrained conditions (i.e. neotates with closed eyes, with specular reflections and/or an oblique gaze), segmentation techniques can classify only 75–85% of images of the sclera.

4. Why Existing Models Fail: A Systematic Analysis

The five real-world challenges analyzed in Section 3 represent distinct, and interrelated, model failure types. Table 1 provides an additional complementary, systematic, structured mapping of the same, along with some performance data extracted from the literature that is reviewed.

Real-World Challenge	Failure Mode	Quantified Impact	Affected Models	Research Gap
Skin tone variation & melanin interference	Melanin absorption at 450 nm confounded with bilirubin; models	MAE increases 2–3× in Fitzpatrick V–VI vs I–II	TcB, SVM, ANN, CNN	Gap 1

	calibrated on light-skinned data cannot compensate in dark-skinned neonates			
Dataset size & diversity deficits	Overfitting to source population; inflated internal but poor external accuracy; African/South Asian populations absent from training	~30% accuracy drop on independent external cohorts	All ML/DL approaches	Gap 2
Ambient lighting variability	Camera white balance and ISP alter colour channel values; CNN features learned under one lighting regime fail under another	Colour features shift ±15% across lighting types; inter-device variance ±3.1 mg/dL	All image-based systems	Gap 3
Absence of multi-centre external validation	All published systems evaluated on within-cohort test sets; real-world performance under domain shift unknown	AUC drop 0.08–0.14 (analogous medical imaging domains)	All published systems	Gap 4
Segmentation & preprocessing limitations	Classical threshold methods fail under lighting/skin-tone variation; U-Net needs large annotated datasets unavailable for diverse neonatal populations	IoU 0.65–0.78 (classical); 15–25% scleral images unusable in uncontrolled settings	All image-based systems	Gap 5
Haemoglobin interference	Haemoglobin absorption at 540/577 nm overlaps colour channels; anaemia alters baseline skin tone independently of bilirubin; no published model explicitly corrects for this	Error amplified in jaundiced neonates with concurrent anaemia; impact unquantified in neonatal imaging literature	TcB, classical image processing, CNN	Gaps 1, 3

Table 1 : Mapping of Real-World Challenges to Model Failure Modes in Non-Invasive Neonatal Jaundice Detection

4.1 Ignoring Melanin and Haemoglobin as Confounding Chromophores

The primary limitation of existing image-based jaundice detection models is their treatment of skin color as a two-component system: a bilirubin signal and noise. In reality, skin color is derived from a complex multi-chromophore relationship. Models that do not explicitly account for the relationship of melanin and hemoglobin will not be able to isolate bilirubin from other chromophore contributors, which will yield estimation errors that will be directly proportional to skin pigmentation. In CNN-based models, if the training data are largely comprised of neonates with very light pigmentation, then the model will capture the skin color associated with low-melanin skin, which will lead the model to misinterpret melanin as bilirubin in darker neonates. (Engle et al., 2002).

4.2 Reliance on Single Modality Without Physiological Context

A second systematic failure is the almost complete dependence on image data originating from a single imaging site, in the absence of accompanying physiologic covariates. Gestational age, post-natal age, an infant's

weight at birth, the manner in which an infant is fed, the blood group incompatibility, and G6PD status all have a substantial effect on the significance of a given Total Serum Bilirubin (TSB) level. (Mariakakis et al., 2017)

4.3 Poor Cross-Domain Generalisation and Absence of Fairness Evaluation

Deep learning models are especially susceptible to distribution shifts when training and application data differ systematically in their statistical properties. With regard to neonatal jaundice detection, distribution shifts occur simultaneously for differences in skin tone, imaging devices, lighting, and protocols. None of the systems reviewed employed any form of domain adaptation. A more critical failure, however, is the overwhelming lack of skin tone-stratified performance evaluation: metrics that are aggregated conceal perilous disparities in subgroups based on skin tone. (Obermeyer et al., 2019) exemplified how the absence of disparities can perpetually sustain an inequitable distribution of health when AI systems are employed to support their clinical decisions.

5. Research Questions and Study Objectives

5.1 Research Questions

Drawing from the structured analysis of real-world obstacles and the systematic catalog of model failure modes, the following five research questions are posed:

RQ1 (Skin Tone Bias): How can skin tone and melanin cross interference be accounted for using a targeted Perceptual Pre-processing Correction (PPC) pipeline in order to achieve accurate and equitable neonatal bilirubin estimation for all Fitzpatrick I-VI skin tones?

RQ2 (Dataset Diversity): How can photometric and geometric data be integrated within a Skin Tone Stratified framework to develop novel and enhanced estimation of neonatal bilirubin within deep learning frameworks for darkly pigmented populations?

RQ3 (Imaging Robustness): How effectively does the combination of PPC normalisation, CZZBP segmentation, and the ST-CNN maintain clinically acceptable bilirubin estimation standards across the range of clinical and community lighting environments - LED, fluorescent, incandescent, and natural daylight - without requiring per-image colour calibration?

RQ4 (Segmentation): Can the CZZBP segmentation algorithm, by leveraging the structural optical and anatomical properties of neonatal skin and sclera, achieve accurate and robust segmentation of clinically relevant anatomical regions under varied lighting and diverse skin tones, without requiring large annotated segmentation training datasets?

RQ5 (Clinical Performance and Deployability): Can the integrated end-to-end system (PPC + CZZBP + ST-CNN) achieve clinically acceptable bilirubin estimation accuracy - MAE $\leq$ 2 mg/dL and sensitivity $\geq$ 90% for clinically significant hyperbilirubinemia - consistently across different skin tone strata, while satisfying the computational and operational constraints for deployment in resource-limited clinical and community settings?

5.2 Study Objectives

Main Objective: To design, implement, and evaluate a non-invasive neonatal jaundice detection system based on deep learning image processing that provides clinically valid and equitable bilirubin estimation accuracy across diverse neonatal skin tones and the range of imaging conditions encountered in real-world clinical and community practice.

Sub-objective 1 (Addressing Gap 1 - Skin Tone Bias): To design and evaluate the Perceptual Pre-processing Correction (PPC) module - a skin-tone-aware colour enhancement and normalisation preprocessing pipeline that estimates melanin contribution using the Individual Typology Angle (ITA) and applies melanin-corrected colour transformations in the CIELAB colour space - to reduce the confounding effect of skin pigmentation on image-derived bilirubin indicators and diminish estimation bias for darkly pigmented neonates.

Sub-objective 2 (Addressing Gap 5 - Segmentation): To design, implement, and evaluate the Colour-Zone-based Z-score Bilirubin Prediction (CZZBP) segmentation algorithm, enabling robust automatic delineation of facial skin, conjunctival sclera, and nasal bridge under varied lighting and skin tone conditions without large-scale annotated segmentation training datasets.

Sub-objective 3 (Addressing Gaps 1 and 2 - Equitable Classification): How can melanin cross interference and skin tone be incorporated in a mixed data domains conference for enhanced estimation of bilirubin in Fitzpatrick VII and below using a blended data domain CNN augmented with ImageNet transfer learning and multisite data to enable accurate bilirubin estimation for neonates across Fitzpatrick skin type I-VI?

Sub-objective 4 (Addressing Gaps 3 and 4 - Robustness and Validation): The goal was to assess the performance, fairness, and robustness of the proposed framework using skin-tone-stratified metrics (Fitzpatrick, ITA quintiles, Monk Scale), varying ambient lighting conditions, and clinically relevant performance metrics, which also include explicit fairness metrics (equal opportunity difference, demographic parity, subgroup MAE disparity).

Sub-objective 5 (Addressing Gaps 3 and 4 - Deployability): As for the proposed system's practical deployability, we considered computational efficiency, compatibility with consumer-grade smartphones, and ease of operation in resource-limited environments, with the goal of achieving an inference latency of three seconds or less when using mid-range smartphones.

6. Proposed Framework: Integrated Architecture for Equitable Neonatal Jaundice Detection

The five challenges and the related research questions provide the impetus for developing an integrated methodological framework comprising three major components. Rather than viewing the five challenges in isolation, the framework considers them as a unified system. Each stage of the system is dependent on the output of the previous stage, and each stage is optimized by the output of the previous stage. Figure 4 provides a fully constructed framework, from the input of a smartphone image to the three processing stages and the final output of bilirubin estimation and severity classification.

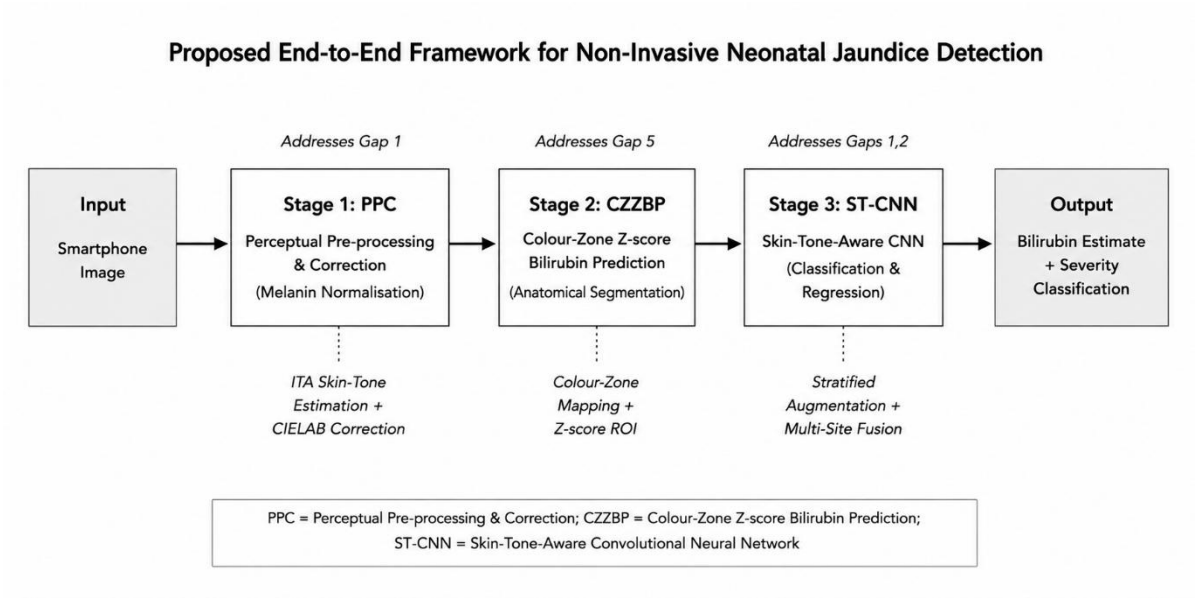

Figure 3: Proposed End-to-End Framework Architecture - PPC (Stage 1) performs melanin-normalised colour preprocessing; CZZBP (Stage 2) segments anatomical regions of interest; ST-CNN (Stage 3) performs skin-tone-aware classification and bilirubin regression

6.1 Perceptual Pre-processing Correction (PPC) Module

The first component of the framework is the PPC module, which directly addresses Gap 1 (melanin interference and skin tone bias). The module is based on the premise that melanin interference must be resolved prior to segmentation and classification. It is not possible to recover lost information from the spectral overlap of melanin-contaminated features. The PPC module computes skin tone for each neonate using the Individual Typology Angle (ITA), derived from the CIELAB L* and b* coordinates as follows: ITA = arctan((L* - 50)/b*). Melanin-corrected color transformations are used to adjust the b* component, establishing a reference for bilirubin-related color signals across skin tone strata. The PPC module is camera agnostic and is designed not to require color calibration targets, thus supporting deployment in LMIC environments.

6.2 Colour-Zone-based Z-score Bilirubin Prediction (CZZBP) Segmentation Algorithm

The second stage of the framework and Gap 5 (limitations of segmentation and preprocessing) is addressed by the CZZBP segmentation algorithm. The main innovation of this algorithm is the ability to develop a segmentation of the neonatal skin and sclera based on the established concepts of anatomy and photophysics, thus avoiding the need for large annotated datasets. This replaces the learned feature representations utilized in U-Net, which are significantly more data demanding. This algorithm creates logical color zones applicable to anatomy where chromophores are expected to appear across certain facial regions. For example, it describes the scleral conjunctiva (little melanin, mostly made of bilirubin), facial skin (includes melanin and oxyhemoglobin), and the nasal bridge (a melanin reference site). For each of these zones, the statistcal analysis of pixel color distribution in CIELAB space identifies, as a statistical outlier in the b* dimension, the bilirubin-relevant color zone, provided the PPC module's ITA estimates of melanin are in place. Within each color zone, Z-score statistical analysis of pixel color distributions identifies the bilirubin-relevant color zone as a statistical outlier in the b* dimension.

6.3 Skin-Tone-Aware Convolutional Neural Network (ST-CNN)

The ST-CNN is the framework's classification and regression component. It closes Gaps 1 and 2. Its architecture is built on a standard transfer learning foundation with three skin tone-focused enhancements. These enhancements are: (1) skin tone-focused, stratified data augmentation targeted to improve representation across ITA-specified skin tone groups, with photometric augmentations closely bounded to the realistic limits of brightness and contrast that preserve bilirubin and melanin color relationships, (2) explicit incorporation of skin tone as an input covariate to the classification head to allow the network to learn feature-bilirubin relations under varying skin tones, and (3) multi-site color feature fusion that utilizes a learned attention mechanism to combine color features from disparate anatomical sites, weighed by the estimated diagnostic signal quality at each site.

7. Expected Research Contributions

The integrated framework embodies five original research contributions, each addressing one or more of the five research gaps:

Contribution 1 - PPC Preprocessing Module (Gap 1): The first melanin decoupling pre-processing step for neonatal jaundice detection allows all subsequent components of the framework to learn the color space where bilirubin exists across all the Fitzpatrick skin types (I-VI), with normalised melanin.

Contribution 2 - CZZBP Algorithm (Gap 5): An innovative segmentation algorithm that relies on a knowledge framework that considers both anatomical photophysical domains, rather than learning from the data, and is hypothesized to provide a reliable decision boundary across diverse skin tones and lighting conditions, in contrast to data-driven approaches, where the training data is often inadequate.

Contribution 3 - ST-CNN Architecture (Gaps 1 and 2): The first neonatal jaundice detection architecture with systematic, evidence-based skin-tone stratification in both training phase and deployment, integrating stratified augmentation, skin tone as an explicit covariate, and using multi-site attention-weighted feature fusion.

Contribution 4 - Fairness-Aware Evaluation Framework (Gaps 3 and 4): A review framework integrating skin-tone stratification and explicit fairness metrics with lighting robustness disassociating the Fitzpatrick phototype, ITA quintiles, and the Monk Skin Tone Scale, which aligns with the shifting global thresholds for responsible AI in healthcare. (Obermeyer et al., 2019).

Contribution 5 - LMIC-Deployable Integrated System (Gaps 3 and 4): An end-to-end system designed for use with consumer-grade smartphones, without any specialized optical accessories or post calibration of colors on a per-image basis, with an target inference latency being below three seconds, possibly making this the first AI neonatal jaundice screening tool available to the populations that bear the highest suffering from global kernicterus.

8. Discussion

8.1 The Problem Formulation Imperative

A central argument of this paper is that the principal bottleneck in non-invasive neonatal jaundice detection is not algorithmic sophistication but problem formulation. Two decades of research have produced progressively more sophisticated architectures - from SVM classifiers to Vision Transformers - applied to essentially unchanged problem framings: train on small, demographically homogeneous datasets; evaluate on within-cohort test sets; report aggregate

performance metrics. This paradigm has produced impressive within-distribution performance numbers while making essentially no progress on the challenges of skin tone equity, real-world robustness, and multi-centre validation that are prerequisites for clinical deployment.

8.2 Equity as a Design Principle, Not an Afterthought

The systematic skin tone bias documented across all categories of non-invasive jaundice detection reflects a structural inequity that will not be resolved by better algorithms applied to the same biased datasets. The path to equitable performance requires addressing the problem at its source: the under-representation of darkly pigmented neonates in training data, the absence of melanin-correction in optical measurement models, and the systematic use of aggregate performance metrics that obscure disparities. Obermeyer et al. (2019) demonstrated that systems performing acceptably on aggregate metrics while performing unacceptably in specific subgroups will, if deployed, amplify existing health inequities. The fairness-aware evaluation framework proposed in this research attempts to make such disparities visible as a core assessment criterion.

8.3 Towards Clinical Translation

The path from research prototype to clinical tool requires progress on at least three fronts beyond algorithmic performance: prospective multi-centre external validation, regulatory pathway engagement, and implementation science. The framework proposed in this research is designed with clinical translation as an explicit design goal. However, we acknowledge that these design choices are necessary but not sufficient: prospective validation in diverse clinical settings remains the essential next step. The absence of large, diverse, openly available neonatal imaging datasets is the single most important structural obstacle, requiring coordinated multi-centre data collection initiatives with standardised protocols and deliberate over-sampling of under-represented skin tone strata.

9. Conclusion

This work proposes a systematic problem formulation for non-invasive neonatal jaundice detection that critically synthesizes the abundant literature in the fields of clinical visual assessment and transcutaneous bilirubinometers, as well as a broad spectrum of classical and contemporary machine learning and deep learning techniques. Five interrelated, complex challenges of the real world were incorporated operationally, presenting examples of model failure: the melanin interference and pigmentation bias leading to a 2-3 fold increase in error when estimating bilirubin concentration; insufficient and poorly diverse datasets resulting in a training distribution that was poorly aligned with the target populations bearing the greatest disease burden; poor resilience to imaging divergence in the real world; the absence of any planned, multi-center external validation; and segmentation methods in anatomy that remain rudimentary.

These four diagrams – the Research Gap Mind Map (Figure 1), the Skin Tone Bias Chart (Figure 2), the Challenge-to-Failure-Mode Causal Chain (Figure 3), and the Framework Architecture (Figure 4) – consist of brief and organized representations of the issue and the integrated answer. These illustrations capture the overall interconnectedness of the five research gaps and the design logic for the PPC + CZZBP + ST-CNN framework.

Kernicterus is completely avoidable. The neurodevelopmental sequela of salsa and the resulting intellectual deficit can be completely avoided if neonatal hyperbilirubinemia is hospital diagnosed and treated in a timely manner. The path to prevention, for the more than 10,000 neonates who suffer kernicterus annually - predominantly in South Asia and sub-Saharan Africa - runs through the development of accurate, accessible, and equitable non-invasive detection tools. This paper has attempted to define that path with greater precision than the existing literature has provided, arguing that progress requires not merely better algorithms, but a principled reformulation of the problem itself.

References

1. Aune, A., Vartdal, G., Bergseng, H., Randeberg, L. L., & Darj, E. (2020). Bilirubin estimates from smartphone images of newborn infants' skin correlated highly to serum bilirubin levels. *Acta Paediatrica*, *109*(12), 2532–2538. https://doi.org/10.1111/apa.15287
2. Deora, K., & Khanal, L. (2023). Hypogammaglobulinemia Causing Multiple Abscesses and Osteomyelitis of Calcaneus Following a Heel Puncture in a Preterm Neonate. *Cureus*. https://doi.org/10.7759/cureus.36992
3. Engle, W. D., Jackson, G. L., Sendelbach, D., Manning, D., & Frawley, W. H. (2002). Assessment of a Transcutaneous Device in the Evaluation of Neonatal Hyperbilirubinemia in a Primarily Hispanic Population. *Pediatrics*, *110*(1), 61–67. https://doi.org/10.1542/peds.110.1.61

4. Gupta, K., Sharma, V., & Kathait, S. S. (2024). Transfer Learning Based Neonatal Jaundice Detection Using MobileNet and EfficientNet. *International Journal of Computer Applications*, *186*(35), 35–43. https://doi.org/10.5120/ijca2024923924
5. Jangaard, K., Curtis, H., & Goldbloom, R. (2006). Estimation of bilirubin using BiliChek™, a transcutaneous bilirubin measurement device: Effects of gestational age and use of phototherapy. *Paediatrics & Child Health*, *11*(2), 79–83. https://doi.org/10.1093/pch/11.2.79
6. Joshi, G., & Singhvi, M. (2024). Comparison of pain during heel prick in preterm neonates receiving only expressed breast milk and expressed breast milk with kangaroo mother care. *International Journal of Contemporary Pediatrics*, *11*(2), 127–132. https://doi.org/10.18203/2349-3291.ijcp20240086
7. Kramer, L. I. (1969). Advancement of Dermal Icterus in the Jaundiced Newborn. *Archives of Pediatrics & Adolescent Medicine*, *118*(3), 454. https://doi.org/10.1001/archpedi.1969.02100040456007
8. Lin, Q., Zhu, D., Chen, C., Feng, Y., Shen, F., & Wu, Z. (2022). Risk factors for neonatal hyperbilirubinemia: A systematic review and meta-analysis. *Translational Pediatrics*, *11*(6), 1001–1009. https://doi.org/10.21037/tp-22-229
9. Lyra, S., Mustafa, A., Rixen, J., Borik, S., Lueken, M., & Leonhardt, S. (2023). Conditional Generative Adversarial Networks for Data Augmentation of a Neonatal Image Dataset. *Sensors*, *23*(2), 999. https://doi.org/10.3390/s23020999
10. Mariakakis, A., Banks, M. A., Phillipi, L., Yu, L., Taylor, J., & Patel, S. N. (2017). BiliScreen: Smartphone-Based Scleral Jaundice Monitoring for Liver and Pancreatic Disorders. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(2), 1–26. https://doi.org/10.1145/3090085
11. Mishra, S., Chawla, D., Agarwal, R., Deorari, A. K., & Paul, V. K. (2010). Transcutaneous bilirubin levels in healthy term and late preterm Indian neonates. *The Indian Journal of Pediatrics*, *77*(1), 45–50. https://doi.org/10.1007/s12098-010-0007-3
12. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342
13. Olusanya, B. O., Imosemi, D. O., & Emokpae, A. A. (2016). Differences Between Transcutaneous and Serum Bilirubin Measurements in Black African Neonates. *Pediatrics*, *138*(3), e20160907. https://doi.org/10.1542/peds.2016-0907
14. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (Vol. 9351, pp. 234–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28
15. Schumacher, R. E. (1990). Noninvasive Measurements of Bilirubin in the Newborn. *Clinics in Perinatology*, *17*(2), 417–435. https://doi.org/10.1016/S0095-5108(18)30576-1
16. Slusher, T. M., Zamora, T. G., Appiah, D., Stanke, J. U., Strand, M. A., Lee, B. W., Richardson, S. B., Keating, E. M., Siddappa, A. M., & Olusanya, B. O. (2017). Burden of severe neonatal jaundice: A systematic review and meta-analysis. *BMJ Paediatrics Open*, *1*(1), e000105. https://doi.org/10.1136/bmjpo-2017-000105
17. Watchko, J. F. (2009). Identification of Neonates at Risk for Hazardous Hyperbilirubinemia: Emerging Clinical Insights. *Pediatric Clinics of North America*, *56*(3), 671–687. https://doi.org/10.1016/j.pcl.2009.04.005
18. *WHO Recommendations on Maternal and Newborn Care for a Positive Postnatal Experience* (1st ed). (2022). World Health Organization.