scientific reports

Check for updates

OPEN

Feasibility study of texture-based machine learning approach for early detection of neonatal jaundice

Nanthida Phattraprayoon[1], Teerapat Ungtrakul[1], Patiparn Kummanee[1], Sunisa Tavaen[1], Tanin Pirunnet[2]✉ & Todsaporn Fuangrod[1]✉

Untreated neonatal jaundice can have severe consequences. Effective screening for neonatal jaundice can prevent long-term complications in infants. Non-invasive approaches may be beneficial in settings with limited resources. This feasibility study explores a texture-based machine learning approach for early detection of neonatal jaundice. Clinical data and skin images of 200 infants were captured from four body locations using the Neonatal Jaundice Screening and Assessment Plate. Data were split into training/validating ($n = 160$) and blind testing ($n = 40$) datasets. Ninety-two features (three clinical, 89 texture-based) were extracted after image processing. Eight machine learning models were compared for bilirubin level prediction. The best performing model, Support Vector Machine (SVM), was implemented in a web-based application (AmberSNAP) and tested using blind testing dataset. SVM paired with RRelief-F feature selection achieved optimal performance for head and sternum measurements, while SVM with Univariate Regression performed best for abdomen and lower leg measurements. Blind testing demonstrated good performance in bilirubin level prediction (mean absolute error: 1.675 mg/dL; root mean square error: 2.192 mg/dL), with moderate correlation between predicted and measured values ($r = 0.644$, $p < 0.001$). These findings suggest that texture-based machine learning is a feasible approach for neonatal jaundice screening in low-resource settings.

Keywords Jaundice, Screening, Neonates, Non-invasive, Machine learning

Newborn jaundice is a common condition characterized by elevated blood bilirubin levels, a condition known as hyperbilirubinemia. It is particularly concerning because unconjugated bilirubin can cross the blood-brain barrier, potentially causing kernicterus and permanent brain damage[1,2]. Newborns are more susceptible due to their higher hematocrit, shorter red blood cell lifespan with approximately 90 days, and immature liver function. Risk factors include Glucose-6-Phosphate Dehydrogenase (G6PD) deficiency[3] and blood type incompatibility in Rh and ABO[4]. Jaundice screening in newborns is therefore critically important for early and accurate jaundice detection in neonatal care.

Traditional screening methods for neonatal jaundice include invasive blood tests and non-invasive methods such as transcutaneous bilirubinometer. Although blood tests offer high precision, they can cause discomfort and pain to the infant. Transcutaneous bilirubinometer provides a non-invasive alternative but can be costly, limiting its accessibility in resource-constrained settings[5–7]. Recent advancements in image processing and artificial intelligence have enabled the development of new approaches for non-invasive jaundice detection. For example, Abiha et al. used a visual inspection method to develop a noninvasive approach to a preliminary neonatal jaundice screening test[8]. Bilirubin accumulation results in yellow discoloration of the skin and conjunctiva of the eyes. Jaundice progression can be assessed at the zone level by dividing the body into five zones: the face and neck, chest and back, abdomen below the umbilicus to the knees, arms and legs below the knees, and hands and feet[8].

Several studies have developed smartphone-based approaches to neonatal jaundice detection, each offering valuable insights into the strengths and limitations of these methods across different populations. Aune et al.[9] developed a smartphone-based method that estimates bilirubin levels by correlating skin color with a physics-based simulation model. Their study demonstrated good performance across diverse ethnic groups, but

[1]Princess Srisavangavadhana College of Medicine, Chulabhorn Royal Academy, Bangkok, Thailand. [2]Department of Pediatrics, Phramongkutklao Hospital and Phramongkutklao College of Medicine, Bangkok, Thailand. ✉email: pirunpediatrician@gmail.com; todsaporn.fua@cra.ac.th

most participants were Caucasian. The authors emphasized the need for further validation in non-Caucasian populations to ensure broader applicability. Taylor et al.[10] introduced the BiliCam app, which paired smartphone images with total serum bilirubin (TSB) measurements. Conducted across seven sites in the United States, their study demonstrated BiliCam's accuracy and cost-effectiveness as a jaundice screening tool. However, over half of the study participants were from white ethnic groups, highlighting the need for more diverse datasets to improve the generalizability of such models. Munkholm et al.[11] investigated the use of smartphone cameras, with and without dermatoscope assistance, for neonatal jaundice screening. Their findings indicated that smartphones, when paired with consistent light sources like dermatoscopes, could effectively estimate bilirubin levels. However, this study was also conducted on Caucasian newborns, necessitating further refinement for broader clinical use.

In China, Rong et al.[12] evaluated the accuracy of the Automated Image-Based Bilirubin (AIB) testing app compared to transcutaneous and total bilirubin measurements. The AIB app demonstrated a strong correlation ($r = 0.79$) with bilirubin levels, suggesting its potential as a non-invasive monitoring tool. However, Swarna et al.[13] reported that similar models underperformed when applied to Indian infants, underscoring the importance of training such models on the target population to address differences in skin color and other factors. These studies collectively highlight that while smartphone-based models for bilirubin estimation hold great promise, their accuracy depends on population-specific training. Differences in ethnicity, skin tone, and environmental conditions can influence the reliability of these methods, emphasizing the need for extensive validation across diverse demographic groups.

This study investigates the feasibility of a texture-based machine learning approach for early neonatal jaundice detection. Using the Neonatal Jaundice Screening and Assessment Plate (NJSNAP), the system analyzes smartphone-captured skin images from four body locations to predict bilirubin levels. A blind testing study validated the model's accuracy using RMSE, MAE, and Pearson correlation.

Research methods

Data collection

The study protocol was granted approval by the Chulabhorn Royal Academy Ethics Committee (114/2564) and the Institutional Review Board of the Royal Thai Army Medical Department (IRBRTA) (S081h/64). All methods were performed in accordance with the relevant guidelines and regulations. Only infants with a gestational age (GA) of 30 weeks or above and less than 28 days old were eligible for inclusion in the study. Written informed consent was obtained from the parents of all infants included in the study. This study adopted a cross-sectional design to test our proposed approach for infant jaundice screening through the integration of texture-based machine learning using the NJSNAP to capture skin color with a smartphone. We obtained sample images by gently applying pressure to the NJSNAP device on the infant's skin to ensure the clarity of the color rendition. Concurrently, we collected blood samples to measure bilirubin as per hospital protocol. Blood samples for bilirubin measurement were obtained using heparinized microcapillary tubes, which required less than 70 µL of blood. Samples were collected via a heel stick or venous blood draw and immediately transferred into hematocrit tubes. The samples were centrifuged at 12,000 rpm for 5 min, and the exterior of the tubes was cleaned using gauze or oil-blotting paper to ensure no contamination.

The bilirubin levels were analyzed using the NEO-BIL Plus neonatal bilirubin analyzer (Rome, Italy), which employs a direct spectrophotometric method. This device uses an LED light source with optical filters at 455 nm and 575 nm to quantify bilirubin levels, providing results within 3 s per sample. The detection range of the analyzer is 0 to 30 mg/dL. The device was calibrated according to the manufacturer's guidelines to ensure measurement accuracy and reliability and it was validated against certified reference method for serum bilirubin measurement[14].

Note that this study applied the measurement of total bilirubin levels using a direct spectrophotometric method. This approach is particularly beneficial for neonatal jaundice screening as it requires only a small blood sample and provides rapid results[15]. The method has been validated in multiple studies and implemented in various clinical settings[16,17], including Thailand[18]. Bilirubin measurement technique in this study enables efficient and accessible bilirubin quantification, which aligns with the goal of providing a low-cost, reliable screening tool for resource-limited regions.

NJSNAP design and photo capture protocol

The NJSNAP is a custom-designed acrylic plate developed specifically for this study. The plate was designed to apply gentle pressure to the infant's skin, squeezing blood from the capillaries. This technique reveals the true color of the infant's skin, mirroring the traditional physical examination method used by physicians to assess neonatal jaundice. Standard Pantone colors (Yellow) were attached to the acrylic plate. These served as reference points to calibrate and standardize color values across different smartphone cameras and lighting conditions. Figure 1a shows a technical illustration of the NJSNAP system, which has a length of 77 mm and a width of 28 mm. The system includes a convex lens with a thickness of 7 mm underneath the plate, which has a thickness of 2 mm. Figure 1b shows a three-dimensional-rendered image of the NJSNAP system.

In each examination, the researcher applied the NJSNAP to four specific locations on the infant's body: forehead, sternum, abdominal wall, and lower leg. Images were captured using an iPhone 11 smartphone (Apple Inc., California, USA). Standard Pantone colors were used to calibrate skin color in the system. Figure 2 shows examples of photos captured by the smartphone at the (a) forehead, (b) sternum, (c) abdominal wall, and (d) lower leg.

System overview

Our proposed system for neonatal jaundice detection, as shown in Fig. 3, consists of three main processes. The workflow begins with the collection of 200 patient datasets, each containing clinical data, including gestational

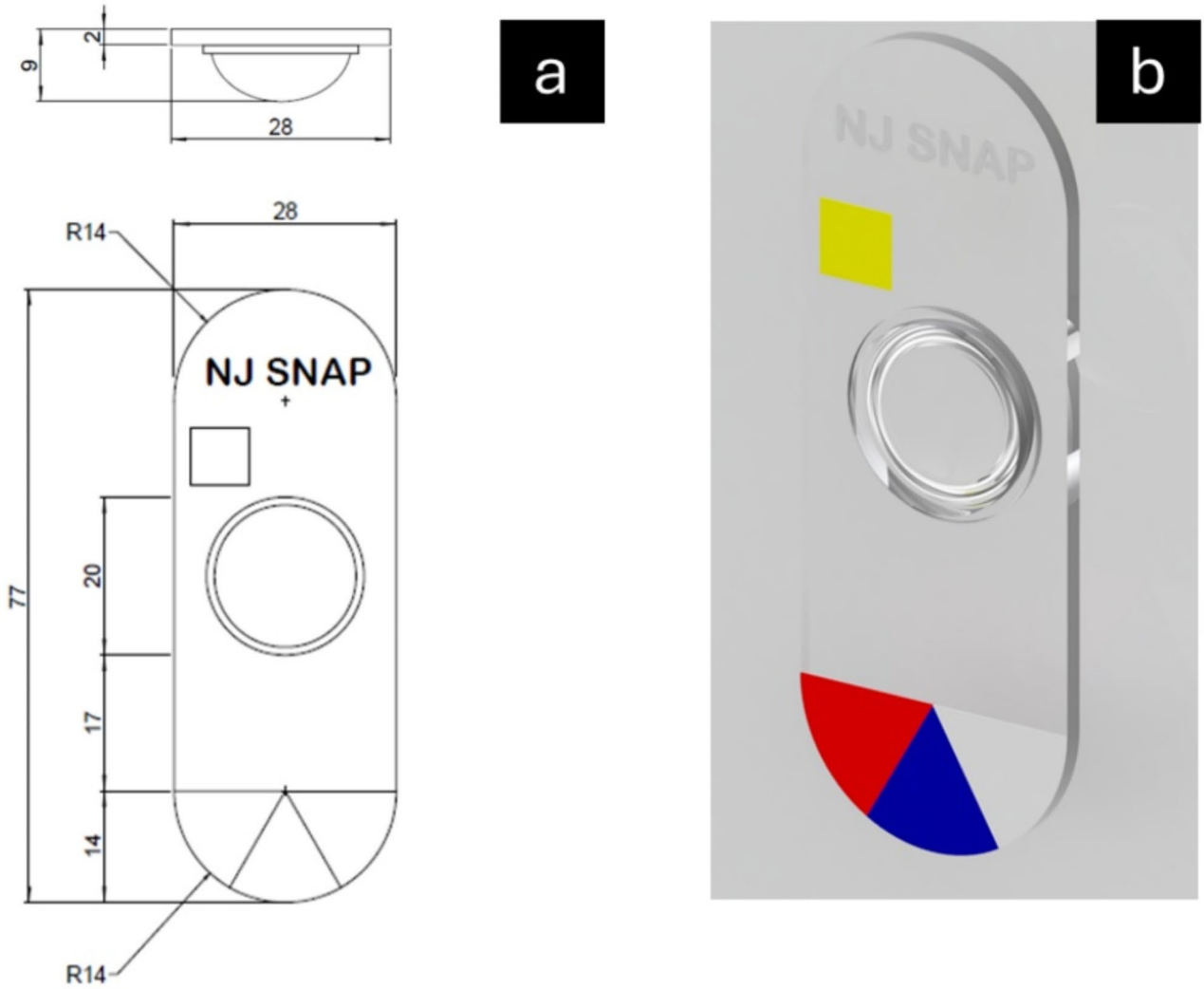

Fig. 1. NJSNAP design (**a**) technical illustration and (**b**) three-dimensional-rendered image.

age and infant age in hours, along with images captured using NJSNAP at four body locations: the forehead, sternum, abdomen, and lower leg. The dataset is randomly divided into two groups: (1) 160 datasets (80%) allocated for training and validating machine learning (ML) models for bilirubin level prediction and risk level classification and (2) 40 datasets (20%) reserved for blind testing.

In the image pre-processing step, the images were classified into regions of interest (ROIs) for two areas: ROI_{yellow} (yellow Pantone, which served as a reference for image calibration analysis) and ROI_{skin} (ROI for model training and prediction). ROI_{skin} was calibrated using the average intensity of ROI_{yellow} to reduce variations caused by different imaging devices and ambient lighting conditions.

Next, the machine learning model development, in which texture features were extracted from the calibrated images and combined with patient-specific data. These datasets were used to develop a regression model for predicting bilirubin level. Each model was trained with feature selection strategies to optimize performance, evaluated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The system was implemented using Python version 3.10, OpenCV version 4.8.1 for image processing tasks, NumPy, and Scikit-Learn version 1.5 for numerical computations and machine learning algorithms.

The best models for each location were selected and implemented into the AmberSNAP web application for blind testing. The maximum predicted bilirubin level was chosen as the final prediction, representing the infant's bilirubin level. The feasibility of our proposed method was evaluated using MAE, RMSE, Mean Absolute Percentage Error (MAPE), R^2 score, and Pearson correlation to assess bilirubin level prediction accuracy.

Region of interest (ROI) and color calibration

Images were converted to grayscale, and Gaussian blur was applied (kernel size 9×9, $\sigma = 2$) to reduce noise. Edge detection was then performed using the Canny algorithm, with lower and upper thresholds of 50 and 150, respectively. Two main tasks were performed for ROI extraction: detection of the Pantone yellow rectangle (ROI_{yellow}) and detection of a circular area of skin (ROI_{skin}).

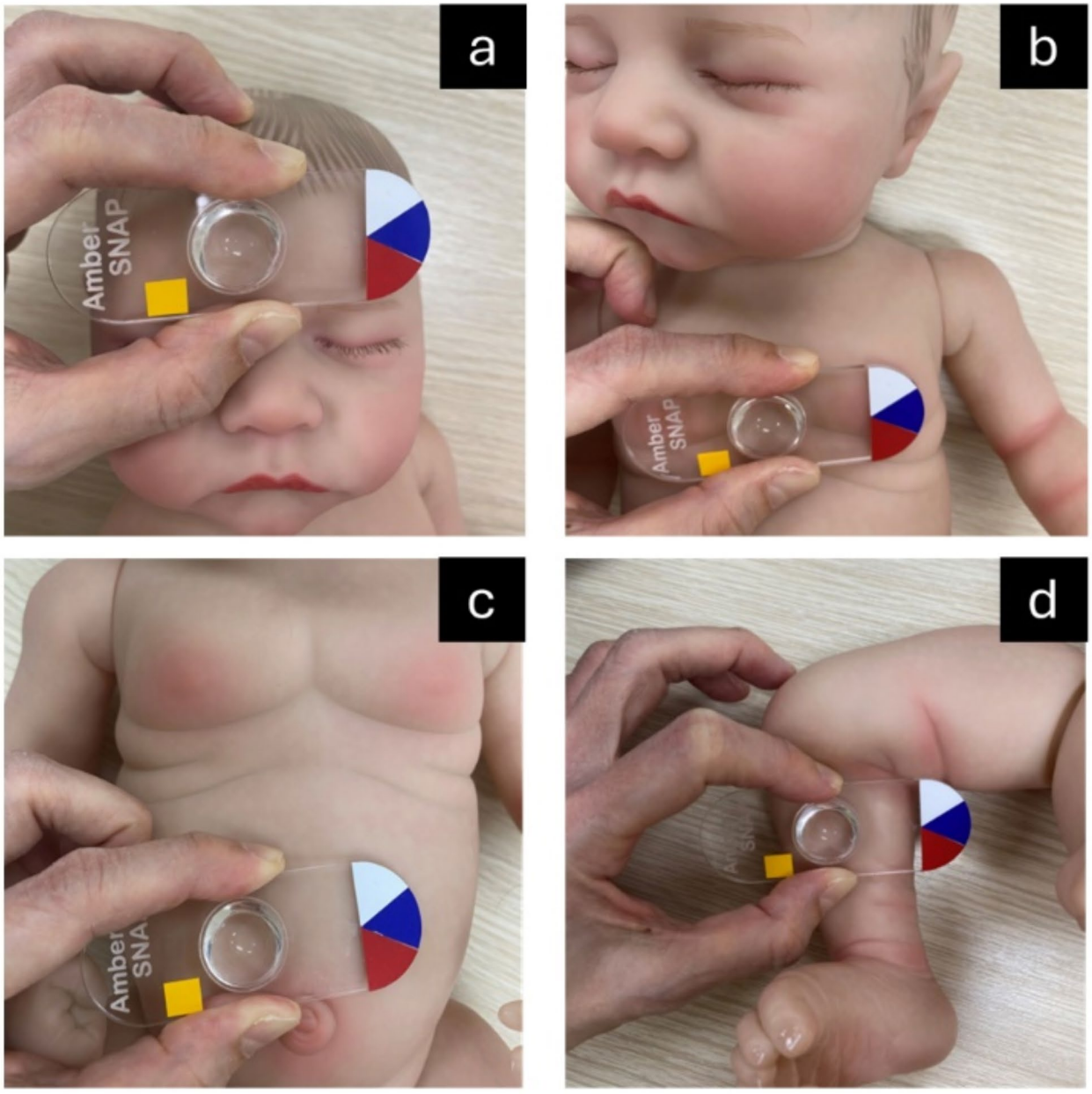

Fig. 2. Device measurement locations: (**a**) forehead, (**b**) sternum, (**c**) abdomen, (**d**) lower legs. Note that these pictures were taken using an infant manikin for demonstration purpose.

For detection of the yellow rectangle, color thresholding in the Hue, Saturation, Value (HSV) color space was applied. Using this method, the yellow region was isolated and then the largest contour was selected to form a polygon. If the polygon had exactly four vertices, it was identified as the yellow rectangle. For the circular region, we employed the Hough circle transform[19]. In cases where multiple circles were detected, the circle closest to the yellow rectangle was selected using Euclidean distance. However, for ROI_{skin}, only half of the detected circle's area was used (maintaining the same circle centroid but using half the radius). Figure 4 shows the original image and the processed image for ROI extraction, highlighting both the yellow rectangle (ROI_{yellow}) and the circle region target area (ROI_{skin}).

For the color calibration of the ROI corresponding to the skin and yellow reference regions, the red, green, blue (RGB) color space of ROI_{skin} was converted to the cyan, magenta, yellow (CMY) color space. This conversion allowed for the extraction of the yellow color channel, which was then compared to the yellow reference extracted from ROI_{yellow}. ROI_{skin} was calibrated by dividing the average intensity value from the yellow channel of ROI_{yellow}. This calibration provided discriminative information that minimized the variation in image sources caused by differences in smartphone color postprocessing and ambient lighting.

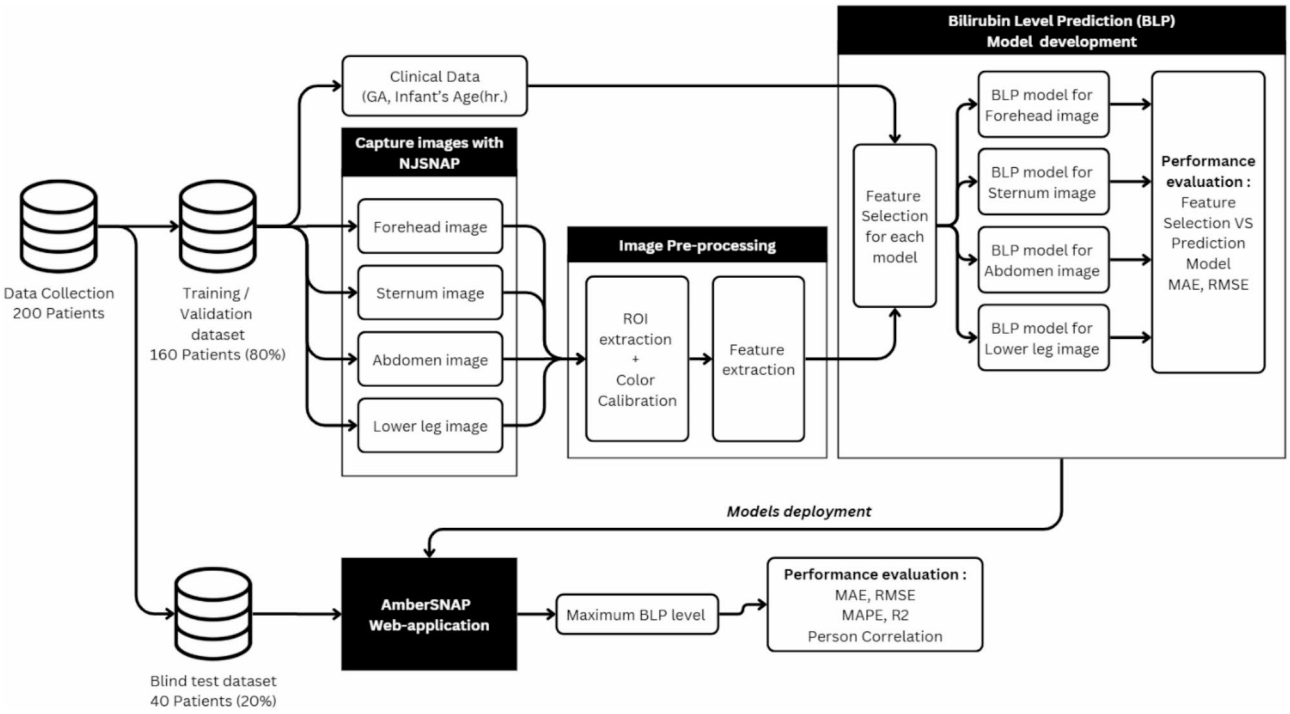

Fig. 3. A framework for developing models to assess newborns for jaundice.

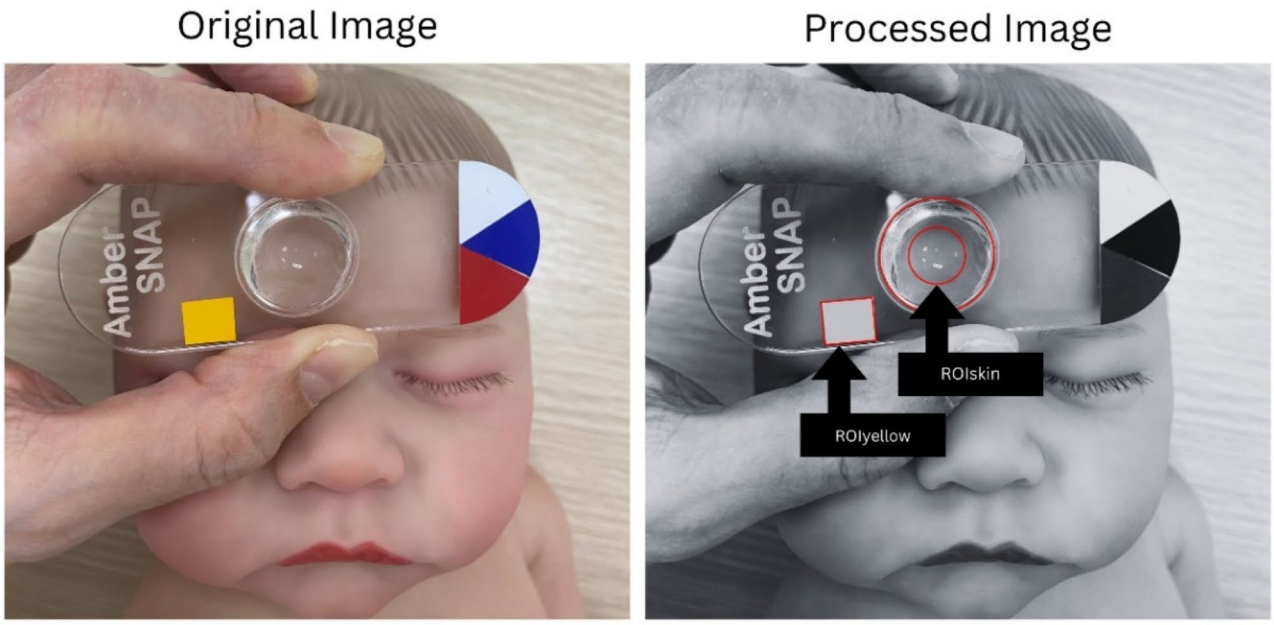

Fig. 4. Example of original image (left) and processed image (right) for ROI extraction (ROI_{yellow}, ROI_{skin}). Note that these pictures were taken using an infant manikin for demonstration purpose.

Texture-based feature extraction

After the ROI_{skin} was calibrated, feature extraction was performed. The extracted features were categorized into two main types: clinical features and texture-based features. The clinical features category included four features that provided clinical information about the image and the patient. The clinical features included the anatomical location of the image (forehead, sternum, abdomen, or lower leg), GA (in weeks), and the age at which blood was drawn (in hours). The texture-based features category covered a wide range of features, including the textural

and spatial characteristics of ROI_{skin}. This category was divided into two subcategories: histogram-based and texture-based features. Table 1 shows a list of all of the features.

Histogram-based features, or first-order features, described the distribution of intensity values within the ROI_{skin}, providing 15 features that represented the basic statistics of the histogram. Texture-based features focused on the spatial relationships and patterns of intensity values within the ROI_{skin}, derived from various matrix representations of the image, which highlighted different aspects of the ROI texture. These features included five types: gray level co-occurrence matrix (GLCM) with 24 features, gray level run length matrix (GLRLM) with 16 features, gray level size zone matrix (GLSZM) with 16 features, gray level dependence matrix (GLDM) with 14 features, and neighborhood gray tone difference matrix (NGTDM) with four features. In total, 92 features were extracted, consisting of three clinical features and 89 texture-based features (15 histogram-based and 74 texture-based features). All texture-based features were extracted using the PyRadiomics library[20].

Features and machine learning model selection

Two distinct feature selection methods were applied for the development of the bilirubin level prediction model: Regressional Relief-F (RRelief-F)[21] and univariate regression[22]. To evaluate the predictive performance of the selected features, we implemented and compared eight different machine learning models: SVM, linear regression, DT, gradient boosting, RF, AdaBoost, neural network, and kNN[23]. The evaluation metrics were root mean square error (RMSE) and mean absolute error (MAE). Each model was trained and tested using 10-fold cross-validation for the model performance assessment. The criteria for selecting the best model for bilirubin level prediction were based on the lowest error rates, considering both RMSE and MAE. RMSE provides a measure of the model's overall accuracy, while MAE indicates the average magnitude of prediction errors. A lower value for both metrics indicates better model performance.

For the risk classification model, a similar approach was applied. Four feature selection methods were used: information gain ratio, analysis of variance (ANOVA), chi-square, and RRelief-F[24]. Based on the classification task, seven ML models were applied, including SVM, DT, gradient boosting, RF, AdaBoost, neural network, and kNN. The evaluation metrics for this classification model included precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). Only the top 20 features from each feature selection method, plus three clinical features, were used. The selection of the best model for risk level classification was based on several key criteria, including overall performance across all evaluation metrics (precision, recall, F1-score, and AUC), robustness across different feature selection methods, focus on AUC scores for comprehensive class distinction, balance between precision and recall, and consistency in performance.

Web-based application development

Figure 5 presents a workflow diagram of our proposed neonatal jaundice screening method, implemented as a web-based application called AmberSNAP. The web-based application process begins with user authentication. Once logged in, the user creates a newborn profile by inputting GA (weeks), date of birth, gender (male, female), newborn name, and parent name. The user then uploads images captured using the NJSNAP device from four different locations on the newborn's body. It's crucial that the NJSNAP device is positioned as close to the center of each image as possible. These images are processed using the proposed algorithm to predict bilirubin level. The highest predicted bilirubin level from the four processed images is selected and displayed. For longitudinal monitoring, users can initiate neonatal jaundice screening at any time. However, images must be uploaded to the system immediately, as the newborn's age is calculated based on the uploaded image timestamp. For

Feature types	Feature category	Number of features	List of features
Clinical features	Clinical	3	Anatomical location of the image (forehead, chest, abdomen, or lower leg), gestational age (GA) (in weeks), age at which blood was drawn (in hours)
Histogram-based features	First-order features	15	10th percentile, 90th percentile, energy, entropy, inter-quartile range, kurtosis, mean absolute deviation, mean, median, robust mean absolute deviation, root mean square, skewness, total energy, uniformity, variance
Texture-based features	Gray level cooccurrence matrix (GLCM)	24	Auto correlation, cluster prominence, cluster shade, cluster tendency, contrast, correlation, difference average, difference entropy, difference variance, inverse difference (ID), inverse difference moment (IDM), inverse difference moment normalized (IDMN), inverse difference normalized (IDN), informational measure of correlation 1 (IMC1), informational measure of correlation 2 (IMC2), inverse variance, joint average, joint energy, joint entropy, maximal correlation coefficient (MCC), maximum probability, sum average, sum entropy, sum squares
	Gray level run length matrix (GLRLM)	16	Gray level non-uniformity, gray level non-uniformity normalized, gray level variance, high gray level run emphasis, long run emphasis, long run high gray level emphasis, low gray level run emphasis, run entropy, run length non-uniformity, run length non-uniformity normalized, run percentage, run variance, short run emphasis, short run high gray level emphasis, short run low gray level emphasis
	Gray level size zone matrix (GLSZM)	16	Gray level non-uniformity, gray level non-uniformity normalized, gray level variance, high gray level zone emphasis, large area emphasis, large area high gray level emphasis, large area low gray level emphasis, low gray level zone emphasis, size zone non-uniformity, size zone non-uniformity normalized, small area emphasis, small area high gray level emphasis, small area low gray level emphasis, zone entropy, zone%, zone variance
	Gray level dependence matrix (GLDM)	14	Small dependence emphasis, large dependence emphasis, gray level non-uniformity, dependence non-uniformity, dependence entropy, gray level variance, dependence variance, low gray level emphasis, high gray level emphasis
	Neighborhood gray tone difference matrix (NGTDM)	4	Busyness, coarseness, complexity, contrast

Table 1. List of features extracted from regions of interest on the infant's skin.

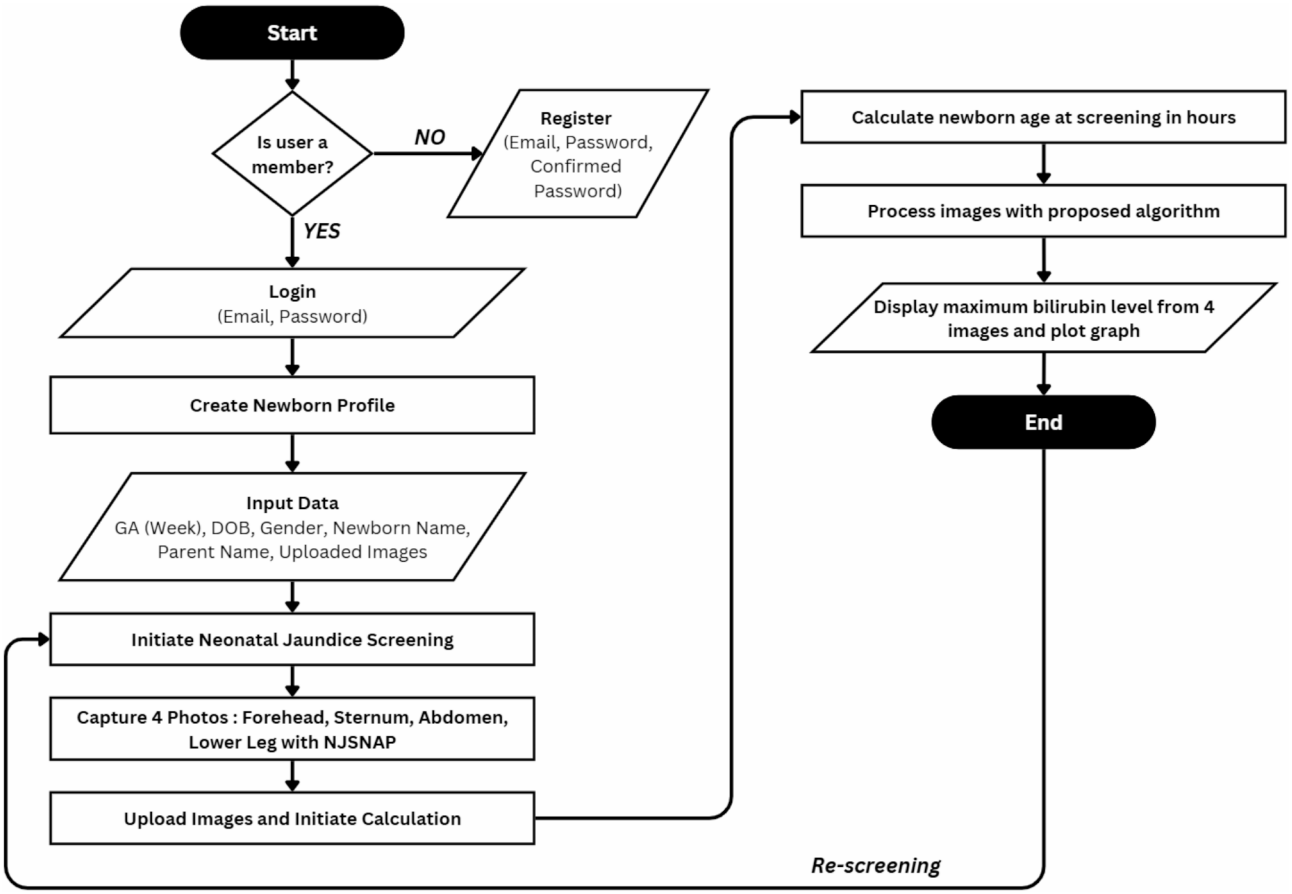

Fig. 5. Workflow diagram of neonatal jaundice screening process in the AmberSNAP web-based application.

blind testing, users were provided with the NJSNAP device and a Quick Response (QR) code to access the AmberSNAP program.

Blind testing

A total of 40 cases were included in a blind test to evaluate the accuracy of the AmberSNAP program in predicting bilirubin levels using the NJSNAP device. To simulate AmberSNAP usage, newborn profiles were manually created, and images were uploaded with artificially set timestamps into the system. The Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were calculated to assess the prediction accuracy of bilirubin levels. Additionally, the Mean Absolute Percentage Error (MAPE) and R^2 score were included to evaluate the relative error and explainability of the model's performance. To further validate the model, Pearson correlation analysis was performed to examine the relationship between predicted and measured bilirubin levels. Additionally, a Bland-Altman analysis was conducted to assess the agreement between the two methods and identify potential biases. The mean difference, standard deviation of differences, and limits of agreement (LOA) were reported to determine the method feasibility for clinical implementation.

Results

Infant data for model development

This feasibility study included 200 infants who met the screening criteria for neonatal jaundice or demonstrated clinical jaundice and whose parents agreed to their participation. The participants had a mean GA of 38.03 weeks (standard deviation [SD]: 1.63) and a mean screening age of 56.30 h (SD: 20.41). Infants with severe illnesses or skin lesions affecting the screening area were excluded from the study. Recruitment took place between March 2022 and October 2022 at Phramongkutklao Hospital in Bangkok, Thailand, from Neonatal intensive care unit, Nursery, and Outpatient departments. Table 2 provides a comprehensive overview of the infants' characteristics.

Bilirubin level prediction accuracy

Tables 3, 4, 5 and 6 present the RMSE and MAE values for each model across different feature selection approaches (RRelief-F and Univariate Regression) at four anatomical locations: head, sternum, abdomen, and lower leg. The results indicate variations in model performance depending on the location of measurement. For the lower leg region, AdaBoost achieved the lowest RMSE (2.888) under Univariate Regression, while SVM demonstrated the best performance in MAE minimization (2.151) under Univariate Regression.

Baseline characteristics	
Maternal age (years)	
Range	17–45
Mean (SD)	29.70 (5.85)
Mother's blood group	
Group O, n/N (%)	78/200 (39)
Group A, n/N (%)	39/200 (19.5)
Group B, n/N (%)	66/200 (33)
Group AB, n/N (%)	17/200 (8.5)
Gestational age (weeks),	
Range	30–40
30 weeks, n/N (%)	1/200 (0.5)
32 weeks, n/N (%)	1/200 (0.5)
33 weeks, n/N (%)	4/200 (2)
34 weeks, n/N (%)	2/200 (1)
35 weeks, n/N (%)	6/200 (3)
36 weeks, n/N (%)	11/200 (5.5)
37 weeks, n/N (%)	25/200 (12.5)
38 weeks, n/N (%)	70/200 (35)
39 weeks, n/N (%)	48/200 (24)
40 weeks, n/N (%)	32/200 (16)
Mean (SD)	38.03 (1.63)
Sex	
Male, n/N (%)	121/200 (60.5)
Birth weight classification	
Appropriate for gestational age, n/N (%)	168/200 (84.0)
Small for gestational age, n/N (%)	26/200 (13.0)
Large for gestational age, n/N (%)	6/200 (3.0)
Birth weight (g)	
Range	1434–4610
1000–1499, n/N (%)	1/200 (0.5)
1500–2500, n/N (%)	25/200 (12.5)
2501–4000, n/N (%)	173/200 (86.5)
≥ 4001, n/N (%)	1/200 (0.5)
Mean (SD)	3016.81 (491.54)
Infant's blood group*	
Group O, n/N (%)	44/94 (46.80)
Group A, n/N (%)	12/94 (12.77)
Group B, n/N (%)	34/94 (36.17)
Group AB, n/N (%)	4/94 (4.26)
Infant's glucose-6-phosphate dehydrogenase (G6PD) status**	
Deficiency, n/N (%)	11/84 (13.10)
Age of screening for jaundice (hours)	
Range	9–192
Mean (SD)	56.3 (20.41)
Microbilirubin level (mg/dL)	
Range	4.3–26.0
Mean (SD)	11.38 (3.10)
Continued	

Baseline characteristics	
Infant's hematocrit (%)	
Range	39.0–70.0
Mean (SD)	51.43 (6.10)
Required treatment	
Phototherapy, n/N (%)	88/200 (44)

Table 2. The infants' characteristics ($N = 200$). SD: Standard deviation. * Only participants whose blood group information was available. ** Only participants whose G6PD levels were measured.

	RMSE		MAE	
Model	RRelief-F	Univariate Regression	RRelief-F	Univariate Regression
SVM	2.678	2.692	2.016	2.032
Linear regression	2.852	2.895	2.262	2.211
Decision tree	3.805	3.754	3.106	2.960
Gradient boosting	3.156	3.188	2.436	2.424
Random forest	3.025	3.044	2.391	2.333
AdaBoost	3.021	2.885	2.284	2.201
Neural network	4.703	4.812	3.652	3.611
kNN	2.839	2.852	2.205	2.224

Table 3. Performance comparison of machine learning models and feature selection methods for bilirubin level prediction (mg/dL) at the head location. kNN: k-Nearest Neighbors; MAE: Mean absolute error; RMSE: Root mean square error; SVM: Support Vector Machine.

	RMSE		MAE	
Model	RRelief-F	Univariate Regression	RRelief-F	Univariate Regression
SVM	2.703	2.789	2.070	2.207
Linear regression	2.930	2.930	2.210	2.341
Decision tree	3.689	3.622	2.978	2.822
Gradient boosting	3.057	2.802	2.394	2.204
Random forest	2.846	2.938	2.282	2.298
AdaBoost	2.868	2.959	2.190	2.295
Neural network	3.555	4.838	2.830	3.720
kNN	2.975	3.030	2.362	2.426

Table 4. Performance comparison of machine learning models and feature selection methods for bilirubin level prediction (mg/dL) at the sternum location. kNN: k-Nearest Neighbors; MAE: Mean absolute error; RMSE: Root mean square error; SVM: Support Vector Machine.

	RMSE		MAE	
Model	RRelief-F	Univariate Regression	RRelief-F	Univariate Regression
SVM	2.921	2.886	2.286	2.222
Linear regression	2.922	3.140	2.368	2.439
Decision tree	3.749	4.170	2.893	3.288
Gradient boosting	3.332	3.276	2.594	2.575
Random forest	3.144	2.989	2.512	2.358
AdaBoost	3.095	3.272	2.353	2.558
Neural network	4.162	4.520	3.389	3.545
kNN	3.111	2.957	2.362	2.295

Table 5. Performance comparison of machine learning models and feature selection methods for bilirubin level prediction (mg/dL) at the abdomen location. kNN: k-Nearest Neighbors; MAE: Mean absolute error; RMSE: Root mean square error; SVM: Support Vector Machine.

Model	RMSE		MAE	
	RRelief-F	Univariate Regression	RRelief-F	Univariate Regression
SVM	2.832	2.812	2.165	2.151
Linear regression	3.101	3.498	2.313	2.403
Decision tree	3.780	3.526	3.034	2.833
Gradient boosting	3.207	3.269	2.491	2.565
Random forest	3.102	3.179	2.302	2.471
AdaBoost	2.973	2.888	2.286	2.204
Neural network	4.608	4.672	3.741	3.853
kNN	3.152	2.911	2.443	2.274

Table 6. Performance comparison of machine learning models and feature selection methods for bilirubin level prediction (mg/dL) at the lower leg location. kNN: k-Nearest Neighbors; MAE: Mean absolute error; RMSE: Root mean square error; SVM: Support Vector Machine.

For bilirubin prediction at the head location, Support Vector Machine (SVM) achieved the lowest RMSE (2.678) and MAE (2.016) using RRelief-F, making it the best-performing model for this feature selection method. For the sternum, Gradient Boosting achieved the lowest RMSE (2.802) under Univariate Regression, while Random Forest had the lowest RMSE (2.846) with RRelief-F. In terms of MAE, SVM demonstrated the lowest error (2.070) under RRelief-F, making it the most accurate model for this feature selection method. For abdomen, SVM achieved the lowest MAE (2.222) using Univariate Regression, making it the most accurate model. For RMSE, Random Forest demonstrated the best performance (2.989) under Univariate Regression, closely followed by kNN (2.957 RMSE). For the lower leg region, AdaBoost achieved the lowest RMSE (2.888) under Univariate Regression, while SVM demonstrated the best performance in MAE minimization (2.151) under Univariate Regression. Therefore, the best performance model for head, sternum, abdomen, and lower leg were SVM with RRelief-F, SVM with RRelief-F, SVM with Univariate Regression, and SVM with Univariate Regression, respectively. These models were selected for deployment in the AmberSNAP web application.

Web-based application development and blind testing

Figure 6 shows an example screenshot of the AmberSNAP program, which allows the user to upload four images of the forehead, sternum, abdomen, and lower leg using the NJSNAP device (Fig. 1a). The system predicts the bilirubin level (see Fig. 1b).

To evaluate the reliability of our method, we analyzed the performance of the blind testing dataset (40 cases) by comparing the predicted bilirubin levels to the measured bilirubin levels obtained via the NEO-BIL Plus neonatal bilirubin analyzer. The analysis resulted in a mean absolute error (MAE) of 1.675 mg/dL and a root mean squared error (RMSE) of 2.192 mg/dL, indicating an acceptable level of accuracy for clinical estimation. The mean absolute percentage error (MAPE) was 18.52%, suggesting a moderate level of relative prediction error. Additionally, the R^2 score was 0.327, showing that the model explains a portion of the variance in bilirubin levels but leaves room for further optimization.

Furthermore, the correlation analysis revealed a moderate positive correlation, with a Pearson correlation coefficient of 0.644 ($p < 0.001$), indicating a reasonable agreement between the predicted and measured bilirubin levels (Fig. 7a). The Bland-Altman plot demonstrated a mean difference of -0.61 mg/dL and a standard deviation of 2.13 mg/dL. The limits of agreement (LOA) ranged from − 4.79 mg/dL to 3.56 mg/dL (Fig. 7b). The bilirubin levels of 10–14 mg/dL demonstrated a better correlation based on Pearson correlation and the Bland-Altman plot. This can be attributed to the prevalence of bilirubin levels within this range in the study.

Discussion

This feasibility study introduced a new approach for neonatal jaundice screening using a custom-designed NJSNAP device and texture-based machine learning techniques. In machine learning development, Support Vector Machine (SVM) emerged as the consistently superior model across all anatomical locations, though with varying feature selection methods. For both head and sternum measurements (Tables 3 and 4), SVM with RRelief-F feature selection demonstrated optimal performance, achieving the lowest MAE (2.016 and 2.070, respectively) and RMSE values (2.678 for head). For abdomen and lower leg measurements (Tables 5 and 6), SVM paired with Univariate Regression showed the best performance, with MAE values of 2.222 and 2.151, respectively. While the results of the blind testing phase demonstrated good performance of the AmberSNAP program, with with an MAE of 1.675 mg/dL and RMSE of 2.192 mg/dL, demonstrating moderate correlation ($r = 0.644$, $p < 0.001$) between predicted and measured bilirubin levels. It should be note that the current sample size, while sufficient for initial validation, may not fully capture the variability seen in larger or more diverse populations. This raises the possibility that the reported metrics might be influenced by sampling bias. Increasing the size of the blind cohort would provide a more robust assessment of the model's performance and reduce the likelihood of overestimating its accuracy.

The choice of feature selection method plays a key role in this study, with different methods providing optimal for different anatomical locations when paired with SVM. RRelief-F demonstrated best performance for the head and sternum measurements due to its ability to handle non-linear relationships and identify features most relevant to textural-based features from skin area. On the other hand, Univariate Regression feature selection

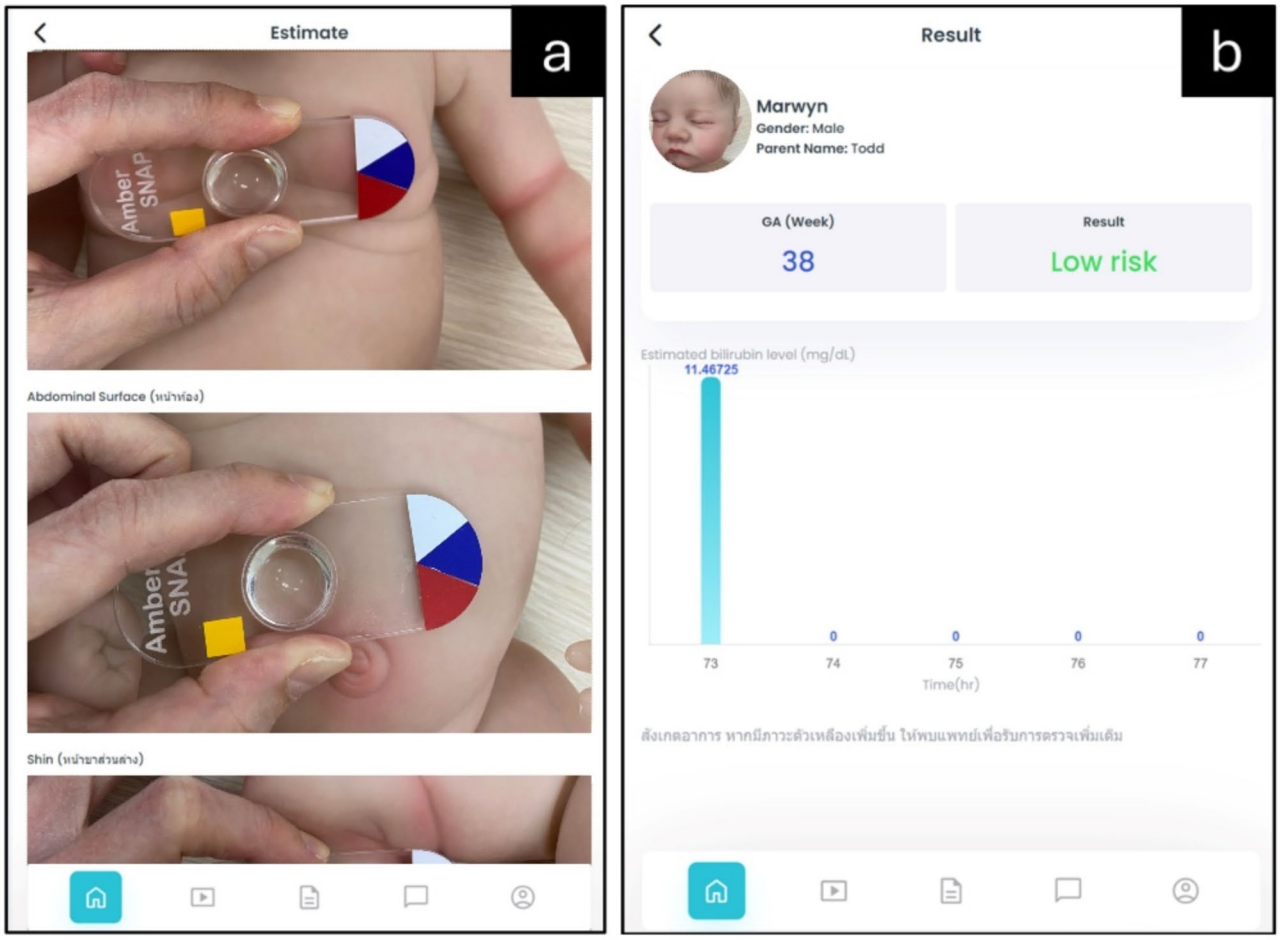

Fig. 6. Example of screenshot from the AmberSNAP software: (**a**) Uploaded images and (**b**) Results. This demonstration was performed using an infant manikin; no real patient name or identifying information was used.

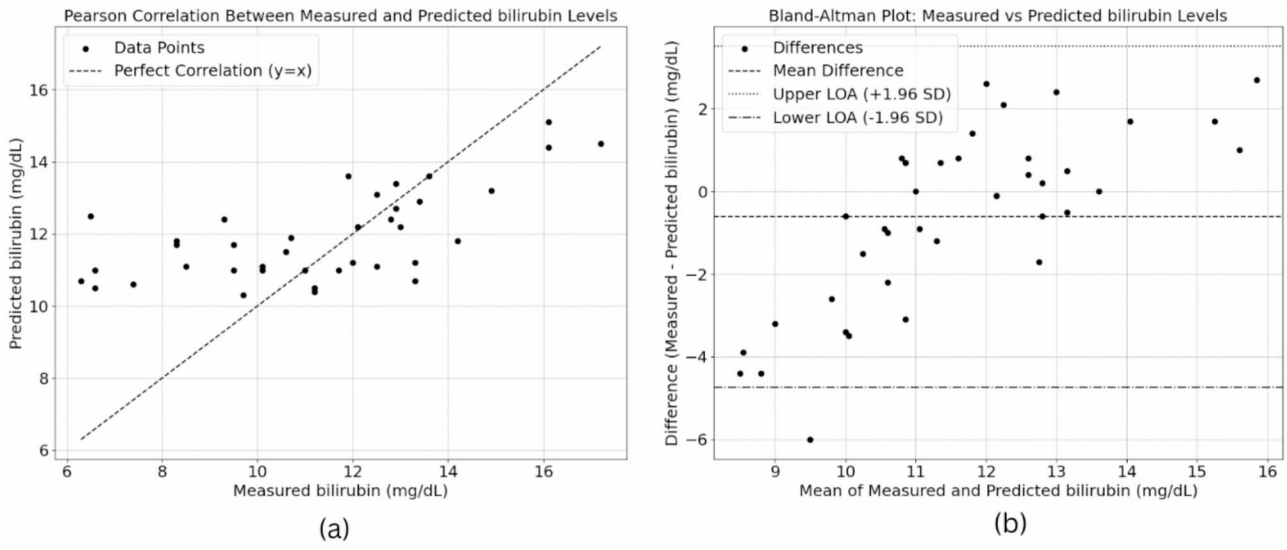

Fig. 7. (**a**) Pearson Correlation Between Predicted and Measured Bilirubin Levels and (**b**) Bland-Altman Plot Showing Agreement Between Predicted and Measured Bilirubin Levels.

is more effective when paired with SVM for the abdomen and lower leg because these regions may have more direct, linear relationships between texture features and bilirubin level.

In contrast to previous approaches, our method integrates physical examination principles with standardized image calibration. Aydın et al. pioneered this approach by developing a system that utilizes photos of a baby's skin alongside a color card for comparison[25]. Their method employed multiple color spaces (RGB [Red, Green, Blue], YCbCr [Luminance, Blue-difference, Red-difference], and LAB [lightness, A: Green-Red, B: Blue-Yellow]) and machine learning techniques (k-nearest neighbor [kNN] and support vector regression) to estimate bilirubin levels. This system achieved an 85% success rate, validated through F-statistical tests and receiver operating characteristic curve analysis. Munkholm et al. explored the use of iPhone 6 smartphones as a screening tool for neonatal hyperbilirubinemia[11]. Their study combined dermatoscopy with smartphone photography, focusing on the neonate glabella[11]. The results provided encouraging evidence supporting the efficacy of this method for hyperbilirubinemia testing. Althnian et al. conducted a comparative study of various machine learning models for jaundice detection[26]. The researchers trained classical models, such as random forest (RF), decision tree (DT), support vector machine (SVM), and multi-layer perceptron(MLP), and benchmarked them against a transfer learning model using VGG-16, a conventional convolutional neural network (CNN)[26]. Data collection utilized the camera of the Samsung Galaxy S7, with color cards placed on newborns' chests. The transfer learning model emerged as the top performer, achieving an accuracy of 86.83%, precision of 84.49%, recall of 81.05%, and an F1-score of 82.12%. Dissaneevate et al. expanded on these approaches by developing a mobile computer-aided diagnosis application[27]. Their study incorporated three different camera types (iOS, Android, and digital single-lens reflex [DSLR]) and employed various machine learning techniques, including DT, kNN, and CNN, to diagnose neonatal hyperbilirubinemia. The CNN model demonstrated superior performance across all camera types, achieving accuracy values of 0.9688, 0.9844, and 0.9688 for iOS, Android, and DSLR, respectively. Similarly, the CNN model yielded the highest precision, with values of 0.7289, 0.7321, and 0.7425 for iOS, Android, and DSLR, respectively.

Similar to Picterus commercial product, their findings highlighted the method's applicability across different ethnicities yet underscored the necessity for further validation in non-Caucasian groups, given the predominance of Caucasian participants. Conversely, our study targets a Southeast Asian demographic, particularly focusing on the complexities introduced by diverse skin tones and bilirubin profiles. Despite both methodologies employing smartphones for bilirubin measurement, our study presents notable differences. The NJSNAP device integrates physical examination principles, utilizing gentle skin compression to enhance true skin color visibility, paralleling traditional jaundice evaluation. This contrasts with other smartphone technologies, like Picterus, which depend exclusively on imaging and software analysis. Furthermore, we employed a Pantone color reference to standardize image calibration, thereby minimizing variability from differing smartphone cameras and ambient lighting.

In this study, the majority of the population consisted of term infants, with 96% having a gestational age (GA) of 35 weeks or more. Specifically, 87.5% (175 cases) involved infants with a GA of 37 weeks or more. Regarding birth weight, 86.5% (173 cases) were infants weighing between 2501 and 4000 g, while only one case had a birth weight between 1000 and 1499 g, and one exceeded 4000 g. Given that our study population primarily comprised term infants with birth weights of 2500 g or more, the bilirubin screening model is tailored for this demographic. Preterm infants or those with lower birth weights often exhibit variations in skin thickness and color, which could influence the accuracy of non-invasive bilirubin measurement. These characteristics, as shown in previous studies[28,29], are significant predictors of hyperbilirubinemia.

Moreover, the majority of participants were appropriate for gestational age (AGA), accounting for 168 out of 200 cases (84%). Additionally, 26 participants (13%) were classified as small for gestational age (SGA), and 6 participants (3%) were classified as large for gestational age (LGA). Although 13% of the population were SGA infants, the study primarily consisted of term infants with birth weights of 2500 g or more. For broader application, especially in preterm infants or those with lower birth weights, the model may require adjustments to account for these physiological differences. Future research should explore integrating birth weight as a feature to enhance the model's performance and ensure its applicability across diverse neonatal populations. This step will help address potential inaccuracies stemming from variations in skin properties and bilirubin metabolism in preterm and low-birth-weight infants.

Even though this feasibility study demonstrated model performance accuracy in bilirubin level prediction, several limitations need to be considered. First, the sample size of 40 infants in the blind testing phase is relatively small. This limited sample might not fully represent the diverse population of neonates, which could impact the generalizability of the findings. Future studies with larger, more diverse cohorts would help validate the model's performance across different ethnic groups and clinical scenarios. The variation between the initial analysis and blind testing performance, as evidenced by differences in RMSE and MAE values, suggests that some factors affecting performance were not fully addressed. These factors could include variations in image quality, lighting conditions, or patient characteristics that weren't fully captured in the training dataset.

Second, Infants of Southeast Asian descent tend to have higher bilirubin levels, and a greater incidence of neonatal jaundice compared to those of Caucasian or African descent[30]. Ethnic variations can also influence skin tone, which may affect the accuracy of bilirubin measurement. Even within the Asian population, notable differences exist. For example, Southeast Asians often have darker and more yellowish skin tones compared to East Asians, such as individuals from Japan or China. The model developed in this study was specifically trained and validated using a data from a Southeast Asian population, with study results specifically derived from Thailand. However, this model could be adapted to accommodate a broader range of populations with similar skin tones and diverse ethnic backgrounds. Further studies are necessary to test and refine the model for use in different demographic groups.

Third, the prediction and classification model development relied on traditional feature extraction and selection methods. Deep learning techniques, such as CNN, could be applied to capture more complex image features, potentially improving model performance. However, due to the limited sample size and the bilirubin level in this study, traditional machine learning approaches conceded better outcomes. Future studies should directly compare this method with established bilirubin measurement techniques to assess its clinical utility, particularly total serum bilirubin which is widely recognized and used in many countries. Moreover, predictive performance could be enhanced by incorporating additional clinical variables, such as variations in birth weight, gestational age, skin color shading, and other risk factors. Usability testing of the AmberSNAP web application is also recommended in future studies to ensure it meets the practical needs of medical professionals. Additionally, long-term follow-up studies should examine the clinical impact of this method, evaluating the efficacy of the neonatal jaundice screening tool.

Conclusion

This feasibility study demonstrates the potential of a texture-based machine learning approach as a non-invasive screening tool for early detection of neonatal jaundice. Our proposed system, AmberSNAP, utilizing the custom-designed NJSNAP device, showed high accuracy in bilirubin level prediction across different anatomical locations. Support Vector Machine (SVM) consistently emerged as the best model, demonstrating good performance when paired with RRelief-F feature selection for head and sternum measurements, and Univariate Regression for abdomen and lower leg measurements. In blind testing, AmberSNAP demonstrated good performance with an MAE of 1.675 mg/dL and RMSE of 2.192 mg/dL, showing moderate correlation ($r = 0.644$, $p < 0.001$) between predicted and measured bilirubin levels. Despite the need for larger-scale validation, AmberSNAP shows feasibility as a screening tool, particularly in resource-limited settings.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due data use agreement with participant under consent agreement that restrict public sharing but are available from the corresponding author on reasonable request.

Received: 14 October 2024; Accepted: 5 February 2025
Published online: 22 February 2025

References

1. Alkén, J., Håkansson, S., Ekéus, C., Gustafson, P. & Norman, M. Rates of extreme neonatal hyperbilirubinemia and kernicterus in children and adherence to national guidelines for screening, diagnosis, and treatment in Sweden. *JAMA Netw. Open.* **2**, e190858. https://doi.org/10.1001/jamanetworkopen.2019.0858 (2019).
2. Qian, S., Kumar, P. & Testai, F. D. Bilirubin Encephalopathy. *Curr. Neurol. Neurosci. Rep.* **22**, 343–353. https://doi.org/10.1007/s11910-022-01204-8 (2022).
3. Hansen, T. W. Kernicterus: An international perspective. *Semin Neonatol.* **7**, 103–109. https://doi.org/10.1053/siny.2002.0118 (2002).
4. Ho, N. K. Neonatal jaundice in Asia. *Baillieres Clin. Haematol.* **5**, 131–142. https://doi.org/10.1016/s0950-3536(11)80038-7 (1992).
5. Lee, B., Piersante, T. & Calkins, K. L. Neonatal hyperbilirubinemia. *Pediatr. Ann.* **51**, e219–e227. https://doi.org/10.3928/19382359-20220407-02 (2022).
6. Bhardwaj, K. et al. Newborn bilirubin screening for preventing severe hyperbilirubinemia and Bilirubin Encephalopathy: A rapid review. *Curr. Pediatr. Rev.* **13**, 67–90. https://doi.org/10.2174/1573396313666170110144345 (2017).
7. Okwundu, C. et al. (ed, I.) Transcutaneous bilirubinometry versus total serum bilirubin measurement for newborns. *Cochrane Database Syst. Rev.* **5** Cd012660 https://doi.org/10.1002/14651858.CD012660.pub2 (2023).
8. Abiha, U., Banerjee, D. S. & Mandal, S. Demystifying non-invasive approaches for screening jaundice in low resource settings: A review. *Front. Pead.* **11**, 1292678 (2023).
9. Aune, A., Vartdal, G., Bergseng, H., Randeberg, L. L. & Darj, E. Bilirubin estimates from smartphone images of newborn infants' skin correlated highly to serum bilirubin levels. *Acta Paediatr.* **109**, 2532–2538 (2020).
10. Taylor, J. A. et al. Use of a smartphone app to assess neonatal jaundice. *Pediatrics* **140** (2017).
11. Munkholm, S. B., Krøgholt, T., Ebbesen, F., Szecsi, P. B. & Kristensen, S. R. The smartphone camera as a potential method for transcutaneous bilirubin measurement. *PloS One.* **13**, e0197938 (2018).
12. Rong, Z. et al. Evaluation of an automatic image-based screening technique for neonatal hyperbilirubinemia. *Zhonghua Er Ke Za Zhi = Chin. J. Pediatr.* **54**, 597–600 (2016).
13. Swarna, S., Pasupathy, S., Chinnasami, B., Manasa, D. & Ramraj, B. The smart phone study: assessing the reliability and accuracy of neonatal jaundice measurement using smart phone application. *Int. J. Contemp. Pediatr.* **5** (2018).
14. Hulzebos, C. V. et al. Diagnostic methods for neonatal hyperbilirubinemia: Benefits, limitations, requirements, and novel developments. *Pediatr. Res.* **90**, 277–283 (2021).
15. Barko, H., Jackson, G. L. & Engle, W. Evaluation of a point-of-care direct spectrophotometric method for measurement of total serum bilirubin in term and near-term neonates. *J. Perinatol.* **26**, 100–105 (2006).
16. Akil, T., Avci, M., Ozturk, C., Akil, I. & Kavukcu, S. Is there any relationship between hyperbilirubinemia and pelvicaliceal dilatation in newborn babies? *Iran. J. Pediatr.* **21**, 431 (2011).
17. Leung, C., Soong, W. & Chen, S. Effect of light on total micro-bilirubin values in vitro. *Zhonghua Yi Xue Za Zhi = Chin. Med. Journal; Free China ed.* **50**, 41–45 (1992).
18. Wananukul, S. & Praisuwanna, P. Clear topical ointment decreases transepidermal water loss in jaundiced preterm infants receiving phototherapy. *J. Med. Association Thail. = Chotmaihet Thangphaet.* **85**, 102–106 (2002).
19. Sarosa, M. et al. in *Journal of Physics: Conference Series*. 077036 (IOP Publishing).
20. Van Griethuysen, J. J. et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**, e104–e107 (2017).
21. Robnik-Šikonja, M. & Kononenko, I. in *Machine learning: Proceedings of the fourteenth international conference (ICML'97)*. 296–304 (Citeseer).
22. Li, J. et al. Feature selection: A data perspective. *ACM Comput. Surv. (CSUR).* **50**, 1–45 (2017).
23. Bonaccorso, G. *Machine Learning Algorithms: Popular Algorithms for data Science and Machine Learning* (Packt Publishing Ltd, 2018).

24. Remeseiro, B. Bolon-Canedo, V. A review of feature selection methods in medical applications. *Comput. Biol. Med.* **112**, 103375 (2019).
25. Aydın, M., Hardalaç, F., Ural, B. & Karap, S. Neonatal jaundice detection system. *J. Med. Syst.* **40**, 1–11 (2016).
26. Althnian, A., Almanea, N. & Aloboud, N. Neonatal jaundice diagnosis using a smartphone camera based on eye, skin, and fused features with transfer learning. *Sensors* **21**, 7038 (2021).
27. Dissaneevate, S. et al. A mobile computer-aided diagnosis of neonatal hyperbilirubinemia using digital image processing and machine learning techniques. *Int. J. Innovative Res. Sci. Stud.* **5**, 10–17 (2022).
28. Daunhawer, I. et al. Enhanced early prediction of clinically relevant neonatal hyperbilirubinemia with machine learning. *Pediatr. Res.* **86**, 122–127 (2019).
29. Koch, G. et al. Leveraging predictive pharmacometrics-based algorithms to enhance perinatal care—application to neonatal Jaundice. *Front. Pharmacol.* **13**, 842548 (2022).
30. Hansen, T. W. R. Narrative review of the epidemiology of neonatal jaundice. *Pediatr. Med.* **4** (2021).

Author contributions
NP and TF conceived and designed the study. All authors contributed to the acquisition, analysis, and interpretation of the data. NP, TU, TP and TF drafted the manuscript. NP, TU, TP and TF revised the manuscript. NP, PK, ST and TF performed the statistical analysis and machine learning development. Administrative and technical support was provided by all authors. All authors had full access to all data and take responsibility for the integrity and accuracy of the data analysis.

Funding
This research project is supported by Chulabhorn Royal Academy, Bangkok, Thailand.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to T.P. or T.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.