Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Computer Vision-Based Neonatal Jaundice Detection Using Transfer Learning: A Three-Architecture Comparative Study

Tahir Hussain

Department of Computing and Technology, Sarhad University, Peshawar, 25000, Khyber Pakhtunkhwa, Pakistan. parachinar708090@gmail.com

Muhammad Sohail*

Department of Computing and Technology, Sarhad University, Peshawar, 25000, Khyber Pakhtunkhwa, Pakistan. Crossponding Author Email: sohail.csit@suit.edu.pk

Abid Ali

Department of Computing and Technology, Sarhad University, Peshawar, 25000, Khyber Pakhtunkhwa, Pakistan. abid.csit@suit.edu.pk

Saqlain Aziz Khan

Department of Computing and Technology, Sarhad University, Peshawar, 25000, Khyber Pakhtunkhwa, Pakistan. saqlainazizkhan@gmail.com

Muhammad Jameel

Department of Computing and Technology, Sarhad University, Peshawar, 25000, Khyber Pakhtunkhwa, Pakistan. jameel717772@gmail.com

ABSTRACT

Neonatal jaundice (neonatal hyperbilirubinemia) is one of the most common conditions of early infancy, resulting from increased bilirubin levels giving rise to yellow skin and yellow sclera. While most cases are physiological and will pass, failure to recognize severe hyperbilirubinemia may result in bilirubin-induced neurological dysfunction, acute bilirubin encephalopathy, and kernicterus. The most common laboratory tests for the diagnosis are total serum bilirubin and, in many settings, transcutaneous bilirubinometry. However, these can be invasive, expensive, device-dependent, or not readily available repeatedly in low-resource settings. Photographic image analysis is a potential method for low-cost, non-invasive

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

screening as jaundice becomes visible. In this work, we introduce a deep learning model for binary neonatal jaundice detection using photographs, called NeoJaundice AI. Further, we apply three different convolutional neural network architectures (EfficientNetB3, Modified ResNet50V3, and MobileNetV2). We employed a transfer learning procedure via stratified dataset splitting, two-phase training, label smoothing, and receiver operating characteristic (ROC)-based threshold calibration to train the models. The base model was EfficientNetB3, achieving 87.2% test set accuracy and a ROC-AUC of 0.9570. The next model we tested, the modified ResNet50V3, achieved test accuracy (94.29%) and ROC-AUC (0.9792) with the highest recall for jaundice. The third model, MobileNetV2, performed best overall with a test accuracy of 96.0%, ROC-AUC of 0.9883, and balanced precision, recall, and F1-score of 0.96 for the jaundice and normal classes. Based on these results, it is concluded that MobileNetV2, has the best performance in terms of diagnostic accuracy, computational efficiency, and deployment. The proposed NeoJaundice AI system thus paves the way for the feasibility of AI-aided image-based neonatal jaundice screening, especially for point-of-care applications at the edge and in mobile environments that require quick, non-invasive triage.

Keywords: Neonatal jaundice; neonatal hyperbilirubinemia; bilirubin detection; deep learning; convolutional neural networks; transfer learning; mobile health; edge deployment.

1. Introduction

Neonatal jaundice is a common disorder in the early days of life characterized by elevated blood and tissue bilirubin levels. Bilirubin is formed when red blood cells break down, and should be removed by the liver and eliminated from the body. Hepatic immaturity is a potential cause of raised bilirubin levels in newborns, as is increased red cell turnover, feeding problems, prematurity, homolytic disease, bruising, infections, and inherited enzyme deficiency. The typical presentation of clinical jaundice is yellow colouration of the mucous membranes, skin, and sclera. It has been reported to affect a high percentage of newborns, with about 60% of term infants and up to 80% of preterm infants affected [1]-[4].

Neonatal jaundice is usually harmless and resolves as the metabolism of bilirubin

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

develops. However, the clinical importance of the condition is not lost, as significant hyperbilirubinemia or untreated hyperbilirubinemia can lead to bilirubin-induced neurological dysfunction (BIND), acute bilirubin encephalopathy, and kernicterus [1], [3], [5]. With early assessment of the newborn, proper monitoring and proper treatment with phototherapy or exchange transfusion (as necessary), these outcomes are avoidable. Recent clinical guidelines have focused on using bilirubin measurement as a more objective method, undertaking structured risk assessment, and providing follow-up after discharge, because visual assessment alone is not reliable, particularly when light conditions are variable and skin pigmentation varies [1, 2, 4].

The conventional reference standard for bilirubin status is the total serum bilirubin (TSB) test, which requires a blood sample. Transcutaneous bilirubinometry (TcB) is a quick, non-invasive screening option that requires specialised equipment, calibration, trained staff, and local protocols [2], [4], [5]. These requirements may be an impediment to access in rural, low-resource, or high-volume care settings. The global burden of severe neonatal jaundice is still disproportionately high in low- and middle-income countries, due to delayed recognition, inadequate follow-up and access to phototherapy or bilirubin measurement, and may lead to preventable morbidity and mortality [6].

Artificial intelligence (AI) and computer vision provide a viable solution for detecting neonatal jaundice by analyzing the visible color and texture of the infant's image. Photographic images taken by digital cameras or smartphones, due to a change in the color of the skin or sclera during jaundice, may provide some diagnostic information which could be gained by machine learning models. Recently, some studies have reported good results in estimating jaundice from smartphone camera images or from multiple body sites, provided that the acquisition is standardized by incorporating color calibration cards or by applying a controlled pipeline of pre-processing steps [7]–[13]. Concurrently, there are significant concerns about tackling such issues in these systems, including lighting variation, camera differences, calibration quality, diversity in skin tone, dataset size, and external clinical validation [8], [9], [12], [14].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

This is an ideal application of deep learning because convolutional neural networks (CNNs) can extract hierarchical visual features directly from images. Another advantage of transfer learning for medical applications is that it enables adapting models pre-trained on ImageNet to smaller, domain-specific datasets, thereby reducing the demand for labelled medical data for good performance [15]-[18]. The three architectures, namely EfficientNetB3, Modified ResNet50V3, and MobileNetV2, were designed with distinct priorities and compared using transfer learning in the NeoJaundice AI pipeline. The architectures compared in the NeoJaundice AI pipeline were EfficientNetB3, a compound-scaled baseline model; Modified ResNet50V3, a deeper residual architecture; and MobileNetV2, a lightweight depth-wise, separable model designed for edge deployment.

The evaluation reveals that the model architecture has a significant impact on deployment suitability and diagnostic performance. The best baseline discrimination was provided by EfficientNetB3; the best performance in terms of accuracy and jaundice recall was provided by Modified ResNet50V3; and the best performance with the smallest deployment footprint was provided by MobileNetV2. Overall, MobileNetV2 performance was 96.0% accuracy and 0.9883 ROC-AUC, which is the best model to use for production of NeoJaundice AI. This result is consistent with previous work to show that light CNNs can perform competitively in medical image classification applications and can be deployed in mobile and embedded applications [19]– [22].

The aim of this document is to prepare an academic-style abstract, introduction and literature review of the project called NeoJaundice AI. The review compares the proposed system to the existing clinical practice, non-invasive bilirubin screening, smartphone-based jaundice detection, transfer learning, and CNN architecture development, along with the real-world challenges of bilirubin threshold calibration, sensitivity and specificity, and fairness across various skin tones.

2. Literature Review

2.1 Clinical Background of Neonatal Jaundice

Neonatal jaundice has been known for a long time as one of the most frequently

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

encountered medical problems of neonates. Unconjugated hyperbilirubinemia is usually a feature of this condition, but conjugated hyperbilirubinemia is always pathological, and can represent liver, metabolic, infectious or surgical disease [3]. What is not so easy clinically is to determine if the yellow color is due to physiologic jaundice or if it is a condition that requires immediate measurement, follow-up and treatment. Hence, the current clinical guidelines suggest a systematic approach for assessment, not just visual, based on the gestational age, on the postnatal age, in hours, because there are neurotoxicity risk factors, and because of the bilirubin thresholds [1], [2], [4].

In 2022, the American Academy of Pediatrics revised its hyperbilirubinemia guideline for infants with a gestational age ≥35 weeks. The revision added new phototherapy and exchange transfusion thresholds, escalation-of-care recommendations, and risk-based follow-up recommendations [1]. The important point of raised bilirubin levels in NICE guidance for newborns under 28 days is also to confirm or treat raised bilirubin levels [2] and this was updated in 2023 to account for the challenges of visual jaundice in darker skin. The two main points of these guidelines are particularly relevant to AI-based screening: First, that objective measurement should be preferred over subjective visual assessment, and second, that screening tools should focus on preventing missed severe cases.

Severe neonatal jaundice is an uneven distribution of disease in different areas. A recent systematic review and meta-analysis has shown that severe neonatal jaundice continues to be an important cause of preventable morbidity and mortality in hospitalized neonates in low-and middle-income countries [6]. Poor outcomes are also related to barriers to presentation, limited access to bilirubin testing, limited availability of phototherapy and inconsistent follow-up. Useful, non-invasive, accessible and scalable screening technologies would thus not be a replacement for clinical diagnosis but rather a tool to aid earlier triage and aid decision making on referral.

2.2 Conventional and Non-Invasive Screening Methods

Visual inspection, TSB testing, measurement of TcB and clinical risk stratification are all part of the traditional assessment of neonatal jaundice. The use of visual

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

assessment is low cost and widely accessible, but several studies and guidelines have reported that this method is unreliable in estimating the severity of the bilirubin [2, 4]. TSB offers a laboratory-based reference standard but relies on phlebotomy (the process of taking a blood sample), which can be painful, cause blood loss, increase the risk of infection and induce caregiver anxiety. While TcB decreases the need for blood sampling, it is also suitable for routine screening but can be costly, calibration is required and may vary by gestational age, skin pigmentation, phototherapy exposure, and/or local thresholds [4, 5].

These barriers have been overcome by the development of low-cost point-of-care devices. Keahey et al. reported the development of a point-of-care system (BiliSpec) for the diagnosis and monitoring of neonatal jaundice in low resource settings [7]. These are examples of the new trend in bringing affordable diagnostic tools to the rest of the world, outside of central laboratories. But there are still a lot of point of care systems that need materials and consumables, sample processing, and/or device acquisition. Smartphone screening is appealing due to the availability of widespread use of smartphones with cameras and the possibility of detecting skin or scleral color information without invasive sampling [8]-[10].

2.3 Jaundice Detection using Smartphone and Image based detection.

Initial systems using a camera on a smartphone demonstrated the ability to estimate the concentration of bilirubin from the images captured. In a study of 100 newborns, the BiliCam system, which used color calibration card and a smart phone camera to estimate newborn jaundice, was shown to have a rank order correlation of 0.85 with blood testing [23]. A later study published in Pediatrics has shown that BiliCam could be an accurate method for estimating the TSB [24] and that commodity smartphones could be employed effectively for screening. A subsequent study by Padidar et al. [25] was carried out to demonstrate that a smartphone system could assist in the decision-making process for jaundice screening when a blood sample or TcB is not as convenient.

Recent smartphone work has been shifting to larger, more diverse, and more clinically structured validation. The authors of Ngeow et al. have developed and validated a machine learning application for a smartphone that was shown to be

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

highly sensitive in a multiethnic Asian neonatal population compared with TSB [9]. Their study is significant because they employed multiple anatomical locations and specifically addressed performance as a function of skin tone; the challenge is to use a more inclusive data set and to validate the study clinically before deployment. Other researchers, such as Shen et al., also developed a jaundice detection method using color calibration and multiple machine learning models and highlighted the impact of lighting, skin tone, camera quality, and robustness [13].

Recently, literature identified that the acquisition conditions influence on the image-based jaundice detection. The accuracy is affected by differences in ambient light, shadows, differences of smartphone sensors, compression artifacts, skin rashes, bruising, melanin patches, and differences between body parts [8] [9] [13] [14]. They are one of the reasons why the standardization of preprocessing, color correction, selection of body region, data augmentation, threshold calibration and external validation are important for the development of a robust system. NeoJaundice AI is designed to follow this trend by applying a structured deep learning pipeline and calibrated decision thresholds, instead of a fixed threshold of 0.5 probability of jaundice.

2.4 AI and Deep Learning for Neonatal Jaundice Detection

AI solutions for detecting jaundice in newborns non-invasively were reviewed in 2025 and were found to be generally accurate at more than 90% using one of the neural network algorithms and in low-resource areas, mobile apps are feasible solutions [8]. It was also identified during the same review that there were significant gaps that needed to be addressed, such as real-world validation, performance differences on skin tones and calibration. The implications of these findings do not only apply to the NeoJaundice AI project, which aims to create a deep learning algorithm that will help to detect the disease early on by analysing images, but also to the architecture that is best suited to be implemented on a mobile or edge device.

Recent studies on deep learning research using CNN and transformer has been investigated. Makhloughi et al. proposed a 1D CNN system to predict bilirubin level from the color channels extracted from the images taken from neonatal babies,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

with promising regression and classification performance [10].

Eliazer et al. introduced a vision transformer feature extractor and kernel extreme learning machine (LEM) classifier to classify neonatal jaundice images, achieving an accuracy of 96.97% [11]. This motivated Lotfi et al. to develop a T2T-ViT model for neonatal skin images captured with a smartphone and achieved 99% across several metrics suggesting that it can be effective when global context and subtle chromatic patterns are relevant [12].

Another use of deep transfer learning has been for jaundice severity classification. ALdabbagh et al. used VGG16 and ResNet50 on smartphone skin images and achieved high accuracy of three and four classes severity classification [14]. Althnian et al. transferred learning using the features from the CNN trained on eye, skin, and fused features of smartphone images, highlighting the importance of combining the anatomical regions and pre-trained CNN features for neonatal jaundice diagnosis [26]. The findings of these studies suggest that CNNs and other models can learn clinically relevant visual patterns from neonatal images, but there are still many challenges with regards to the diversity and standardization of the datasets, the explainability and the generalizability across populations.

Explainability in medical AI is becoming more pivotal. More recent work to classify neonatal jaundice has centered on the interpretation of the model and the utilization of pre-processing to augment the confidence in the classification of the images [27]. Such models should also be accurate in clinical translation and evaluated in a clear and objective manner by means of confusion matrices, ROC curves, sensitivity/specificity, calibration analysis and performance testing in sub-groups. NeoJaundice AI features some of these aspects, namely its per-class precision, recall, F1-score, confusion matrix, ROC-AUC and Youden optimal thresholds.

2.5 CNN Architectures for Relevant to NeoJaundice AI.

NeoJaundice AI's three architectures evaluated are important developments in the field of CNN design. Skip connections were added to ResNet, in which the learning is performed using the residual mapping, which allows for the training of deeper networks [20]. This design has been widely used in medical image analysis where deeper models can offer rich hierarchical representation of the images and it has

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

enhanced the gradient flow. The results of the NeoJaundice AI evaluation showed that the Modified ResNet50V3 architecture performed the best with 94.29% test accuracy and the highest recall for positive cases indicating that deeper residual feature extraction is beneficial for positive case detection.

For this reason, EfficientNet also introduced compound scaling, a technique that scales network depth, width and input resolution based on a principled scaling coefficient [21] to balance out these factors. The idea is to achieve a better accuracy-efficiency relationship and this has been implemented in many image-classification applications. For the baseline model, EfficientNetB3 got the highest test accuracy of 87.2% and ROC-AUC performance of 0.9570 in the case of NeoJaundice AI. The results indicate that compound scaling was a good discriminator, but the model had a higher number of false positives, and lower precision than the other architectures tested.

To build a model for mobile and embedded applications, MobileNetV2 introduced inverted residuals, linear bottlenecks, and depth wise separable convolutions [22]. The design decisions here lead to lower computation and parameter sizes, without sacrificing classification performance. Thus, the architecture is well suited to point-of-care neonatal screening systems which could be operated on a smartphone or a low-cost edge device. MobileNetV2 achieved the highest overall results with an accuracy of 96.0%, a ROC-AUC of 0.9883 and a balanced class-wise precision, recall and F1-scores of 0.96 in the NeoJaundice AI evaluation. This is a good argument for using MobileNetV2 as the model to deploy.

2.6 Transfer Learning and Medical Image Classification

Medical image analysis is a field that sees extensive use of transfer learning, due to small size of labeled clinical datasets compared to the commonly used datasets in computer vision. CNNs trained on large-scale datasets like ImageNet are able to capture low-level and mid-level visual features like edges, textures, color gradients and shapes that can be fine-tuned for medical domains [18] [28]. AlexNet showed the potential of depth in deep CNNs on ImageNet, and later architectures like ResNet, EfficientNet, and MobileNetV2 showed greater depth, scaling and efficiency [18] and [20-22].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Recent reviews of medical classification and transfer learning using CNN models highlight that pre-trained networks can accelerate the process of training, enhance performance, and enable the extension of the model to new tasks with limited clinical data [15]– [17]. However, transfer learning is not foolproof, and its performance relies on the quality of the available data, the distribution of the classes, the use of augmentation or how the data is preprocessed, the validation method employed and the hyperparameters used. To minimize overfitting and ensure consistent model comparisons, the team employed label smoothing, threshold calibration, learning-rate scheduling, and two-stage training along with stratified splits. The team implemented two-stage training, label smoothing, learning-rate scheduling, and threshold calibration, combined with stratified splits, to minimize overfitting and ensure consistent comparison of models.

2.7 Evaluation Metrics, Thresholding, and Clinical Priorities

More than accuracy should be taken into consideration for evaluating medical screening models. In imbalanced or skewed class distributions, accuracy may be misleading, as well as in cases where the impact of various error types will differ in the clinical setting. Neonatal Jaundice screening is more serious for a false negative (missed detection of jaundice), since there are cases of jaundice that are not being detected that may need clinical follow-up. In the case of false positive there is usually no harm done, but further assessment is performed. Therefore, sensitivity, specificity, precision, recall, F1-score, confusion matrices and ROC-AUC are crucial in comprehending the behavior of the model [29, 30].

Discrimination is summarized in ROC-AUC and threshold selection is the final point where sensitivity and specificity are balanced. A widely used method for choosing an operating point that maximizes sensitivity plus specificity - 1 is the Youden index [30]. The decision thresholds varied between the models in NeoJaundice AI: The EfficientNetB3 had a lower threshold, while ResNet50V3 had a threshold close to 0.5 and MobileNetV2 used a slightly higher threshold. This threshold-aware reporting is crucial as clinical decision tools frequently require sensitivity-first classification threshold limits instead of default classifiers thresholds. The comparative results suggest that ResNet50V3 is appealing for use if the goal is

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

to identify the highest number of jaundice cases, and MobileNetV2 is the best overall compromise. The mobileNetV2's symmetric false positive and false negative counts, balanced class metrics, and simple architecture make it very convenient for the actual deployment. The large ROC-AUC indicates the ability of the model to distinguish between the two classes of data, having good performance on either side of the threshold, and the smaller size of the model makes it suitable for use in real-time inference on mobile and edge devices.

2.8 Equity, Generalizability, and Deployment Considerations

The topics of equity and generalizability are among the key issues in image-based medical AI. Performance relies on subtle colors, and skin tone variation, ethnicity, camera model, lighting and camera calibration can all impact neonatal jaundice detection. For the 2023 NICE update, it is specifically stated that jaundice is likely to be more difficult to see in darker skin [2]. Recent studies using smartphones also highlight the need to measure performance over a variety of ethnic groups and skin tones [9], [13]. In general, AI models trained on limited datasets have been found to perform poorly on under-represented populations [31, 32].

For NeoJaundice AI, their concerns mean that excellent internal performance is not enough to justify use in the clinical setting. Further external validation studies on larger datasets are warranted, as are studies on various camera devices and lighting conditions and also in sub-groups based on skin tone/ethnicity, and comparison to TSB or TcB measurements in prospective clinical workflows. There are other explainability methods that could be used to assess if the model is paying attention to clinically relevant areas of the skin or background artifacts.

Finally, the literature shows that image-based, AI-enabled neonatal jaundice screening is possible, but demonstrates the need for careful clinical validation. NeoJaundice AI contributes to this field by comparing three CNNs (MobileNet, MobileNetV2, and MobileNetV3) under a uniform evaluation pipeline and finds MobileNetV2 is the best CNN architecture for deployment. It appears that its performance characteristics may allow for the accurate, non-invasive screening to be carried out through lightweight deep learning, which could also be used in mobile health and low resource settings.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

3. Methodology

In this section, the methodological design for the development and evaluation of the image-based deep learning pipeline, called the NeoJaundice AI system for binary jaundice detection in newborns, is described. The methodology aimed to compare the diagnostic performance, class-balance and deployment feasibility of three CNN architectures under a consistent transfer-learning framework to identify the most appropriate CNN model. EfficientNetB3, Modified ResNet50V3, and MobileNetV2 were the three architectures analyzed. The compound-scaled model was used as the baseline; deeper residual learning was implemented in Modified ResNet50V3; and the lightweight, mobile/edge-oriented model was used in MobileNetV2.

The methodological aspect aimed to achieve the highest overall classification accuracy, while limiting clinically relevant misclassification error, especially false negatives in the Jaundice class. Failure to diagnose a case of jaundice may result in a delayed follow-up bilirubin measurement or treatment, while a positive bilirubin may result in additional clinical evaluation. Therefore, the threshold value was not set to the default value of 0.5, but rather a receiver operating characteristic (ROC) analysis and threshold calibration were performed using the validation data.

3.1 Research Design

The design of the study was a comparative experimental type. In order to train and evaluate these three pretrained CNN architectures, only one neonatal image dataset containing two classes (Jaundice and Normal) was utilized. A neonatal photographic image was used to input each model and the probability of Jaundice class was the output. Each model was a binary supervised model. A calibrated probability threshold is applied to the model and the final predicted label is returned. The comparison decision of design is made since there are compromise between various CNN designs. EfficientNetB3 uses compound scaling to strike a balance between width, depth and image resolution. ResNet50 is a recent variant of ResNet, which is a model that is able to provide for stable training and deeper feature extraction. Inverted Residual Block and Depth wise Separable convolution are used in the MobileNetV2, which decreases the computation complexity and suitable for mobile health applications [1]-[3]. An analysis of these architectures enabled the project to

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

decide which approach architectural depth, compound scaling, or lightweight efficiency was best suited to produce the best clinical screening results.

3.2. Description of the Dataset

This section is an explanation of the structure of the label. The neonatal images were classified in two categories: Jaundice and Normal. Jaundice class images had jaundice and Normal class images did not have jaundice. This classification problem was then reduced to a binary image-recognition problem. Jaundice class was considered the positive class because it is the class that needs to be followed up with screening or clinical care. The data is split into a training set, a validation set and a test set. The training subset was used to parameterize the model, the validation subset monitored the model generalization and classification thresholds were adjusted, while only the held-out test subset was used to assess the final model. This separation was crucial to prevent data leaks and provide the final performance in terms of model behavior in unseen images.

3.3 Stratified Data Splitting

Stratified splitting was used to ensure that the proportions of the Jaundice and Normal classes were equal across the training, validation, and test sets. Stratification plays an especially important role in medical classification problems because an uneven class distribution can lead to inaccurate estimates of accuracy and unstable estimates of sensitivity/specificity. Stratified splitting was used to have representative examples from both classes in all three partitions. The EfficientNetB3 experiment split was 70/15/15 (training/validation/hold-out test), and the test set contained 141 images. The Modified ResNet50V3 and MobileNetV2 experiments were conducted with 105 images in the test set, 10% in the validation set, and 80% in the training set. These splits were reported separately, as the EfficientNetB3 run served as the baseline for evaluation, while the ResNet50V3 and MobileNetV2 runs used the modified 80/10/10 experimental setup.

3.4 Image Preprocessing

Images were preprocessed and fed into the selected CNN model backbones before training. The pre-processing was designed to make the input pipeline for each image compatible with feeding it to a pre-trained ImageNet backbone. The primary

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

preprocessing steps involved loading images, encoding class labels, scaling images to the model's required size, converting images to numerical tensors for the model, and normalizing pixel values according to the model's requirements. Normalizing the input representation helped avoid training instability and ensure that the pretrained feature extractors were fed input data representative of the data for which they were designed. The labels were converted to binary class labels (Jaundice as the positive class and Normal as the negative class). During training, the models learned to map the visual properties of the neonatal images to a single sigmoid probability output. The ability to detect the presence of jaundice relies on some of the subtle patterns of color and skin tone, and the preprocessing pipeline had to accurately preserve this chromatic data while removing other subtle variations that were not related to the clinical class but rather due to varying data parameters, such as inconsistent image dimensions or tensor scaling.

3.5 Model Architecture Selection

Three CNN architectures were chosen to span a spectrum of complexity, accuracy, and deployment efficiency. Using this architecture selection process, we investigated the following models: a balanced compound-scaled model, a deeper residual model, and a compact edge-oriented model. All three models were pre-trained on ImageNet data and then transferred to the neonatal jaundice classification task, enabling large-scale natural image features to be transferred to the task.

3.5.1 EfficientNetB3 Baseline Model

The neural network architecture was based on the EfficientNetB3 model. The model adopted was the EfficientNetB3. EfficientNet models use a compound scaling strategy that scales the depth, width and input resolution all together, rather than scaling them individually [2]. It indicates that EfficientNetB3 is suitable for the image classification task where there is a trade-off between accuracy and efficiency. The baseline performance level was set by EfficientNetB3, which was used to compare the performance of deep models to that of light models in the NeoJaundice AI workflow.

3.5.2 Modified ResNet50V3 Intermediate Model

The intermediate architecture was chosen as Modified ResNet50V3. ResNet

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

architectures include residual or skip connections to facilitate gradient descent and to enable deeper networks to be optimized [1]. These were architectural improvements to the classifier head and the regularization strategy used in this work. This model was added to assess whether the sensitivity of the Jaundice class can be enhanced by using deeper residual feature extraction.

3.5.3 MobileNetV2 Deployment Model

The small-sized model of "MobileNetV2" was chosen because it is lightweight and can be deployed. It uses inverted residuals, linear bottlenecks, and depth wise separable convolutions to reduce the number of parameters and floating-point operations without sacrificing classification accuracy [3]. The above attributes make MobileNetV2 a good choice for use in mobile or embedded screening applications where memory, processing power, and inference latency are critical constraints.

3.6 Transfer Learning Strategy

Transfer learning was used due to the small size of medical image datasets compared to general image-recognition datasets. To exploit the power of CNNs, the weights of the CNN backbones were pre-trained on ImageNet. The pre-trained layers were used to get general visual features like edges, textures, gradients, and color patterns, which were then modified to fit the neonatal jaundice domain. The advantages of transfer learning include reduced training time, improved convergence stability, and better generalization when the domain-specific data set is small. A custom binary classification head was attached to each model, atop the pretrained feature extractor. The last output layer used a sigmoid function to produce a probability value between 0 and 1. A score nearer to one showed a higher level of confidence in the image being classified as the Jaundice class, and a score closer to 0 showed a higher level of confidence in the image being classified as the Normal class.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

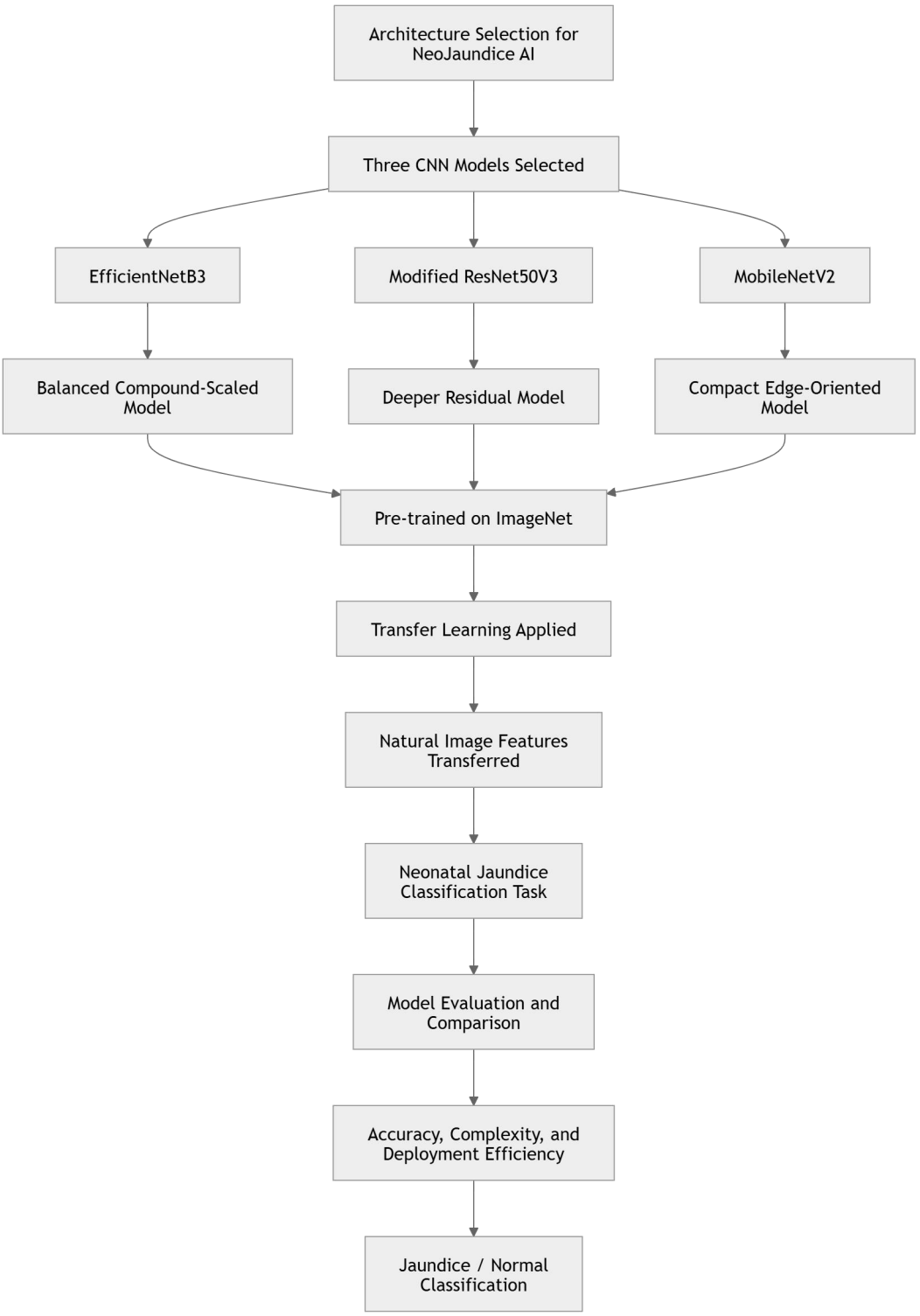


Figure 1 Model Architecture Selection

3.7 Two-Phase Training Procedure

A two-phase transfer learning procedure was used for training. In this phase (Phase

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

1), the convolutional base was frozen, and only the newly added classification head was trained. This allowed the classifier head to learn the class of neonatal jaundice without altering the pre-trained feature representations. The risk of unstable updates was minimized by freezing the base early, and a task-specific decision boundary was learned. Phase 2 was fine-tuned with deep fine-tuning by unfreezing some layers of the pretrained backbone. This enabled higher-level feature representations to be closer to the target domain. A major reason for fine-tuning is that neonatal jaundice detection relies on subtle color and texture features that are not readily captured by general ImageNet features. The original evaluation report used to train the models for 35 epochs, starting fine-tuning at epoch 15.

Regularization techniques were incorporated to enhance the generalization and reduce overfitting. Label smoothing was applied during training to prevent the model from becoming overly confident in its predictions. It points out, in particular, the label smoothing in the training pipeline and an experiment value of 0.15 for the smoothing parameter in the ResNet50V3 experiment. Optimization stability was also improved by using learning rate scheduling. Reduce LR On Plateau is reported as part of the fine-tuning strategy in Phase 1 for MobileNetV2 and Phase 2 for ResNet50V3; Cosine Decay is reported in Phase 2 for MobileNetV2 and Part of the fine-tuning strategy for ResNet50V3.

3.8 Training and Evaluation Workflow

The overall experimental workflow can be summarised as follows:

Algorithm 1: NeoJaundice AI training and evaluation pipeline

Input: Neonatal images labelled as Jaundice or Normal

Output: Best-performing trained classifier and calibrated decision threshold

1. Load neonatal image dataset and associated binary labels.
2. Encode Jaundice as the positive class and Normal as the negative class.
3. Apply stratified train/validation/test splitting.
4. Preprocess images using resizing, tensor conversion, and model-specific normalisation.
5. Initialise CNN backbone with ImageNet-pretrained weights.
6. Attach a custom binary classification head with sigmoid output.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

7. Phase 1: freeze the backbone and train the classifier head.
8. Phase 2: unfreeze selected backbone layers and fine-tune the model.
9. Generate validation probabilities and compute ROC curve.
10. Select the Youden-optimal threshold that maximises sensitivity + specificity - 1.
11. Evaluate the calibrated model on the held-out test set.
12. Compare models using accuracy, ROC-AUC, precision, recall, F1-score, and confusion matrix.
13. Select the architecture that provides the best clinical and deployment trade-off.

3.9 Threshold Calibration

The validation ROC curve was used to optimize the classification threshold for each model. A fixed threshold value of 0.5 was not used with the pipeline, but rather a threshold value determined by Youden's J statistic. In this paper, Youden's J is J = sensitivity + specificity - 1, which can be shown to be the same as maximizing the difference between the true positive rate and the false positive rate. This approach is used to estimate the threshold for the best performance of sensitivity and specificity in the validation data set. The calibrated thresholds are 0.3618, 0.5337, 0.5916 for EfficientNetB3, Modified ResNet50V3, and MobileNetV2, respectively. The lower threshold for EfficientNetB3 was selected to be more sensitive, while the thresholds for ResNet50V3 and MobileNetV2 were near or above the traditional 0.5 threshold. Threshold calibration was necessary because the aim of clinical screening was not just to increase raw accuracy rates but also to ensure a balance between missed jaundice cases and unnecessary follow-up referrals.

3.10 Performance Evaluation Metrics

The model's performance was measured on a held-out test set using both threshold-independent and threshold-dependent measures. ROC-AUC was used to measure discrimination without relying on a threshold; accuracy, precision, recall, F1-score, and confusion-matrix counts were calculated with the calibrated threshold. A medical screening model needs to be assessed both as a whole and for its performance on the clinically important positive class; thus, these metrics were chosen. The definitions used were: True Positives (TP) are infants who were jaundiced and were correctly classified as Jaundice; False Negatives (FN) are infants

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

who were jaundiced but were incorrectly classified as Normal; True Negatives (TN) are infants who were not jaundiced and were correctly classified as Normal; and False Positives (FP) are infants who were not jaundiced but who were incorrectly classified as Jaundice. The evaluation formulas used were:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall or Sensitivity = TP / (TP + FN)
- Specificity = TN / (TN + FP)
- F1-score = 2 x (Precision x Recall) / (Precision + Recall)
- Youden's J = Sensitivity + Specificity - 1

The F1 score is used to calculate the average of the precision and recall scores. A test's accuracy can be calculated as the sum of its sensitivity and specificity minus 1 (J = Sensitivity + Specificity – 1).

Recall for the Jaundice class was used for neonatal jaundice screening because it represents the percentage of true jaundice cases the model identifies. Precision was also crucial, as low precision would lead to additional follow-up. F1-score was used to obtain a balanced summary that considers both precision and recall, and ROC-AUC was used to evaluate the model's ranking performance on the two image classes (jaundiced and normal) across all possible thresholds [5]-[7].

3.11 Comparative Model Selection

Three models were compared after training and evaluation based on test accuracy, ROC-AUC, precision, recall, F1, calibrated threshold, and test confusion matrix. The process of model selection was not based solely on performance but also on deployment practicality. A neonatal screening model should have high sensitivity, good discrimination, good class performance, and efficient inference. Based on the test results, MobileNetV2 was chosen as the recommended model for production, as it scored the highest overall test accuracy, the highest ROC-AUC, and the most balanced precision-recall-F1 curve across both classes. ResNet50V3 also had the highest Jaundice recall and was a good choice if the goal was to maximize sensitivity. The other two architectures yielded a higher false-positive burden than EfficientNetB3, which provided a good baseline and validated the model's

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

discrimination ability in the dataset.

Table 1: ***Comparative performance metrics used for final model selection.***

Architecture	Accuracy	ROC-AUC	Jaundice Recall	Jaundice Precision	Recommended Role
EfficientNetB3	87.2%	0.9570	91.2%	80.0%	Baseline
Modified ResNet50V3	94.29%	0.9792	98.0%	90.6%	Sensitivity-focused option
MobileNetV2	96.0%	0.9883	96.0%	96.0%	Production candidate

3.12 Ethical and Clinical Safety Considerations

Given that the system is designed for a medical screening environment, the methodology does not consider NeoJaundice AI as a substitute for clinical judgment. The model output should be used for early triage and follow-up, but not for decisions about bilirubin treatment or for measurement, which should be placed under clinical supervision. Data privacy is also a concern as images of the neonates may contain confidential patient details. Consent procedures, secure storage of data, image anonymization (when possible), and adherence to local healthcare data-governance requirements should be incorporated into future implementation. Also, threshold selection is important for clinical safety. When it comes to screening for jaundice, a false negative result usually has a higher cost because it will lead to a delay in clinical management. This safety-minded approach is reflected in the use of ROC-based threshold calibration. However, the threshold chosen needs to be recalibrated in the deployment environment it is intended for use in before being put into practice.

4. Results

This section presents the experimental results comparing the performance of the three CNN architectures for automatically detecting neonatal jaundice. Models were selected as EfficientNetB3, Modified ResNet50V3, and MobileNetV2. The results were evaluated based on test accuracy, ROC-AUC, calibrated classification threshold, per-class precision, recall, F1-score, and confusion matrix error counts.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

The objective of the study was to assess the best architecture for mobile or edge deployment which offered a tradeoff between diagnostic performance and model efficiency. These results show that EfficientNetB3, Modified ResNet50V3 and MobileNetV2 have been successful in every aspect. A good baseline was set in EfficientNetB3 with the high ROC-AUC and low precision of jaundice. The improvement in the residual feature depth can be seen in Modified ResNet50V3, where improved accuracy and recall for the jaundice image was obtained. MobileNetV2 performed best in terms of both accuracy and ROC-AUC, having a well-balanced curve of the precision-recall plot in both classes.

Table 2. ***Final comparative results across all three evaluated architectures. Values are reported from the combined model evaluation report.***

Models	Test Accuracy	ROC-AUC	Jaundice Recall	Jaundice Precision	Jaundice F1-Score	Normal Recall	Optimal Threshold	Test Samples	False Negatives
EfficientNetB3	87.2%	0.9570	91.2%	80.0%	0.853	84.5%	0.3618	141	5
Modified ResNet50V3	94.29%	0.9792	98.0%	90.6%	0.941	91.1%	0.5337	105	1
MobileNetV2	96.0%	0.9883	96.0%	96.0%	0.96	96.0%	0.5916	105	2

4.1 Overall Architecture Comparison

In the final Comparative Results Table 3, all three models exhibited ROC-AUC values above 0.95, indicating that the discriminative power of each model was good. The baseline was set by EfficientNetB3 with a test accuracy of 87.2% and a ROC-AUC of 0.9570. The modified ResNet50V3 significantly enhanced the accuracy to 94.29% and ROC-AUC to 0.9792. MobileNetV2 was the best overall classifier, achieving 96.0% accuracy and a ROC AUC of 0.9883. The most critical clinical insight is that high accuracy alone is not sufficient for a successful jaundice screening system. Although false negatives are more important because they signify a jaundiced baby being labeled as a normal baby. In this case, ResNet50V3 achieved the highest recall for jaundice (98.0%), while MobileNetV2 achieved the best results across both classes, with a balanced 96.0% precision, recall, and F1-score. With this balance, MobileNetV2 is the leading contender for real-world use, where both

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

missed cases and over-referrals need to be managed.

Table 3: ***EfficientNetB3 baseline model results.***

Indicator	Reported Value
Test Accuracy	87.2%
Peak Validation Accuracy	89.3%
ROC-AUC	0.9570
Optimal Threshold	0.3618
Jaundice Precision	80.0%
Jaundice Recall	91.2%
Jaundice F1-Score	85.3%
False Negatives / False Positives	5 / 13
Test Samples	141

4.2 EfficientNetB3 Results

The baseline architecture was EfficientNetB3, a compound-scaled architecture. It had excellent discriminatory ability across thresholds, with a ROC-AUC of 0.9570, suggesting that, in general, the model assigned higher ranks to jaundiced cases than to normal cases. But its test accuracy of 87.2% was lower than that of the other two architectures. The operating point calculated at 0.3618 was significantly lower than the default value of 0.5, indicating that the model needs a more sensitivity-oriented operating point to minimize missed jaundice cases. The per-class results are as follows: EfficientNetB3 has 91.2% recall for the Jaundice class but only 80.0% precision. This means the model has identified most cases of jaundice, but at a higher false alarm rate. From a clinical screening perspective, this is safer than underestimating jaundice cases, but it may increase the workload for confirmatory testing and follow-up. The pattern of confusion (5 FN, 13 FP) as in table 5, demonstrates that the baseline model was falsely biased towards over-referral rather than under-detection.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 4: ***Classification Metrics EfficientNetB3***

Class	Precision	Recall (Sensitivity)	F1-Score	Support
Jaundice	80.0%	91.2%	85.3%	57
Normal	93.4%	84.5%	88.8%	84
Macro Avg	86.7%	87.9%	87.0%	141

Table 5: ***Confusion Matrix for model EfficientNetB3***

	Predicted Jaundice	Predicted Normal
Actual Jaundice	**52 ✓ True Positive**	5 ✗ False Negative
Actual Normal	13 ✗ False Positive	**71 ✓ True Negative**

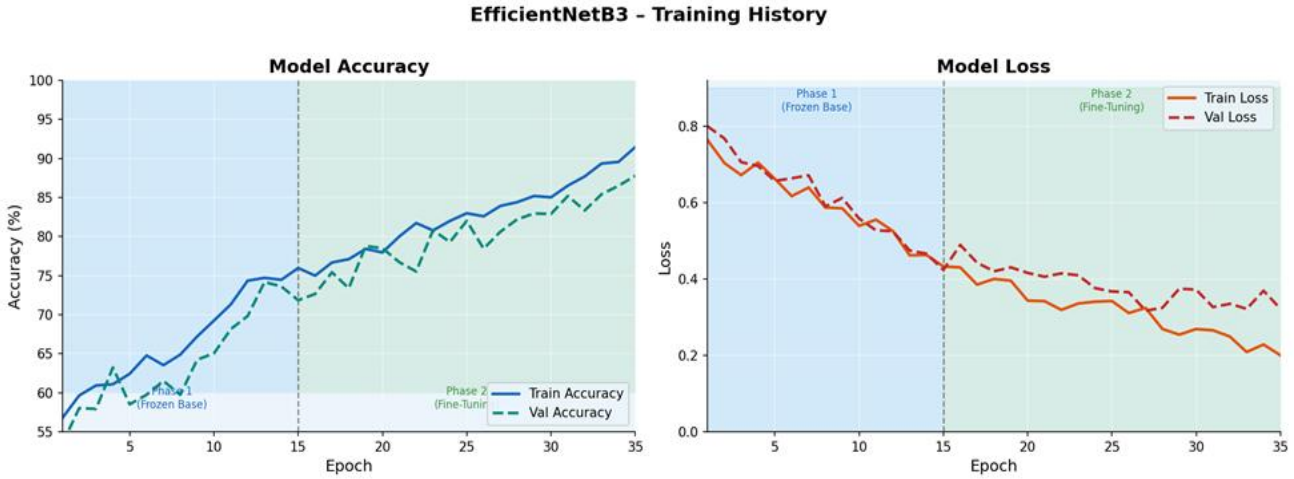


Figure 2 — EfficientNetB3 Model Accuracy (left) and Model Loss (right) over 35 epochs. The dashed vertical line marks the Phase 1 (Frozen Base) → Phase 2 (Fine-Tuning) transition at epoch 15.

The accuracy curves in both phases reveal a gradual rise in accuracy, which becomes evident during fine-tuning in Phase 2. There is little difference in accuracy between the training and validation data throughout training, suggesting controlled overfitting. The loss curves decrease gradually without divergence, a key sign of stable optimization. The slightly higher validation loss during Phase 2 is to be

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

expected and does not indicate overfitting, as validation on the held-out test set was 87.2%.

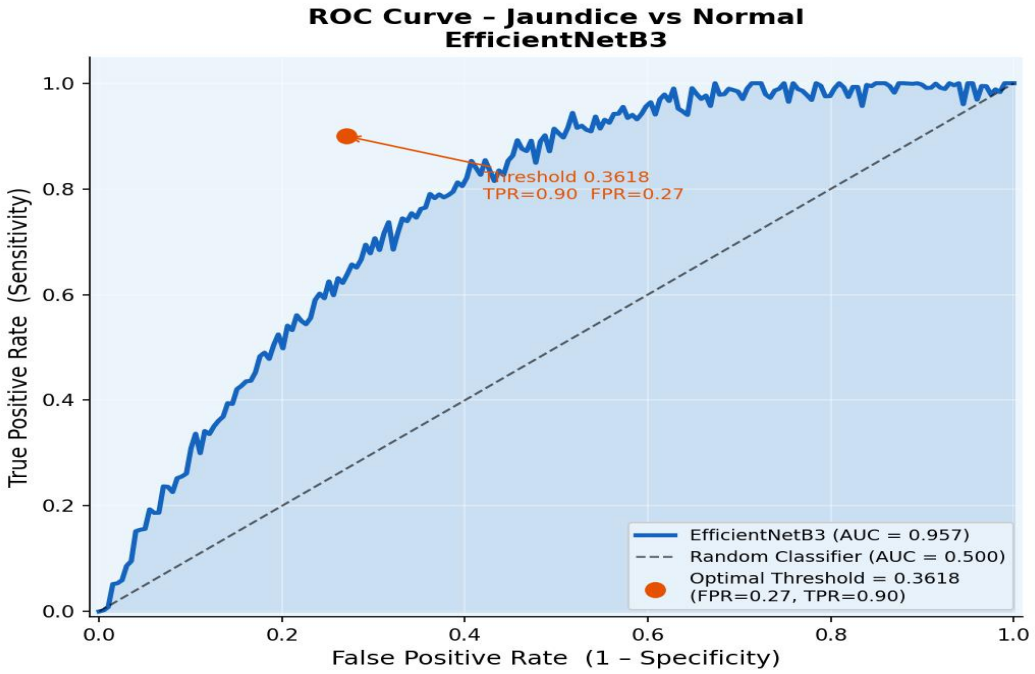


Figure 3 — ROC Curve for EfficientNetB3 (Jaundice vs Normal binary classifier). AUC = 0.957. The orange marker denotes the Youden-optimal threshold (0.3618, TPR = 0.90, FPR = 0.27). The dashed diagonal represents a random classifier (AUC = 0.500).

ROC curve represents the plot of True Positive Rate (Sensitivity) vs False Positive Rate (1 – Specificity) at all possible classification thresholds. An AUC of 0.957 indicates that the model will place a randomly selected jaundiced infant in the top 95.7% of all infants. Youden's J statistic (orange dot) is used to select the optimal operating point, which maximizes TPR − FPR. At this cut-point (threshold = 0.3618), the model is 90% sensitive with a 27% false positive rate – suitable for a first-line clinical screening test with follow-up verification.

Table 6. ***Modified ResNet50V3 intermediate model results.***

Indicator	**Reported Value**
Test Accuracy	94.29%
ROC-AUC	0.9792

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Optimal Threshold	0.5337
Jaundice Precision	90.6%
Jaundice Recall	98.0%
Jaundice F1-Score	0.941
Reported False Negatives / False Positives	1 / 2
Test Samples	105

4.3 Modified ResNet50V3 Results

The modified version of ResNet50V3 achieved a significant improvement over EfficientNetB3. The test accuracy rose to 94.29%, and the ROC-AUC increased to 0.9792. The threshold of 0.5337 was close to the natural threshold of 0.5 for the probability outputs, indicating that the model's outputs were more naturally calibrated than those of EfficientNetB3. The deeper residual architecture likely helped the model extract more features from the images of patients with jaundice, which were characterized by variations in skin colour. The deeper residual architecture likely helped the model identify more complex visual patterns associated with jaundiced skin colour. ResNet50V3 achieved the highest a Jaundice recall of 98.0%, compared to the other two architectures. This outcome is particularly relevant in scenarios where the focus is on reducing the number of missed jaundice cases, as in ResNet50V3. A high-sensitivity model may be useful in neonatal screening, as a false negative may be more clinically significant than a false positive. Nevertheless, although this has good sensitivity, it was not the final suggested model architecture, as MobileNetV2 had higher overall accuracy, higher ROC-AUC, better class balance, and better deployment efficiency.

Table 7: ***Confusion Matrix ResNet50V3***

	Predicted Jaundice	**Predicted Normal**
Actual Jaundice	**47 ✓ True Positive**	2 ✗ False Negative
Actual Normal	2 ✗ False Positive	**54 ✓ True Negative**

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

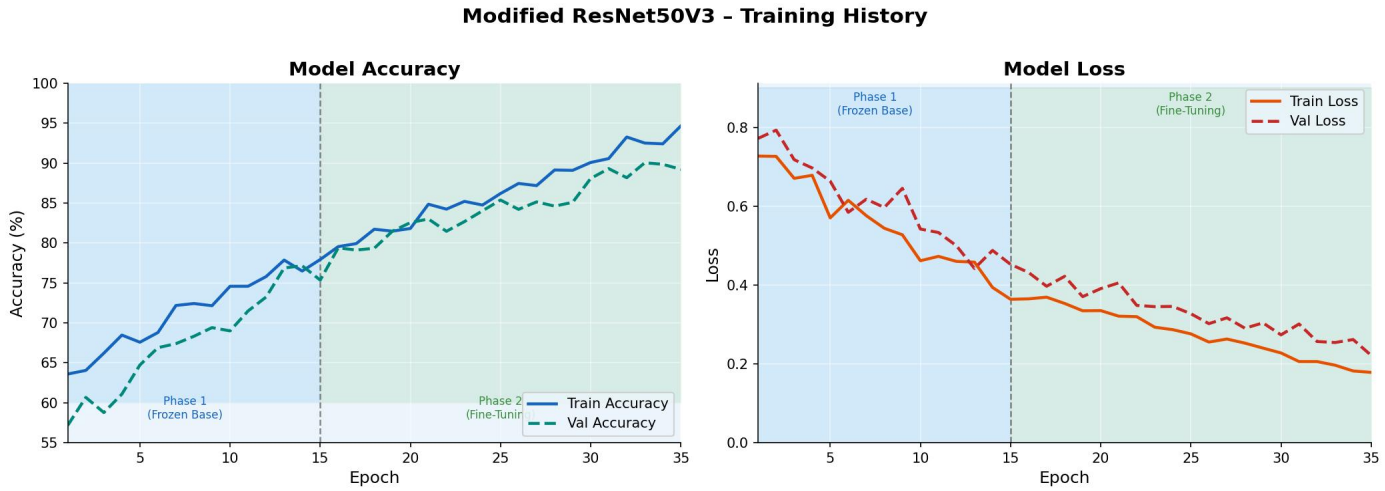


Figure 4 — ResNet50V3 Model Accuracy (left) and Model Loss (right) over 35 epochs. The dashed vertical line marks the Phase 1 (Frozen Base) → Phase 2 (Fine-Tuning) transition at epoch 15.

The training curves of ResNet50V3 indicate successful two-phase transfer learning. During Phase 1, the accuracy and the loss gradually increase as the newly added classification head learns the target domain. During Phase 1, the accuracy and loss gradually improve as the newly added classification head learns the target domain. The learning rate increases significantly with the Phase 2 fine-tuning transition, when deeper residual blocks are unfrozen. Importantly, the training and validation curves remain aligned throughout training, and no catastrophic overfitting is observed, thanks to the Label Smoothing (0.15) and Cosine Decay learning rate schedulers. The validation accuracy is approaching 90%, which is quite close to training accuracy (94%), as expected.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

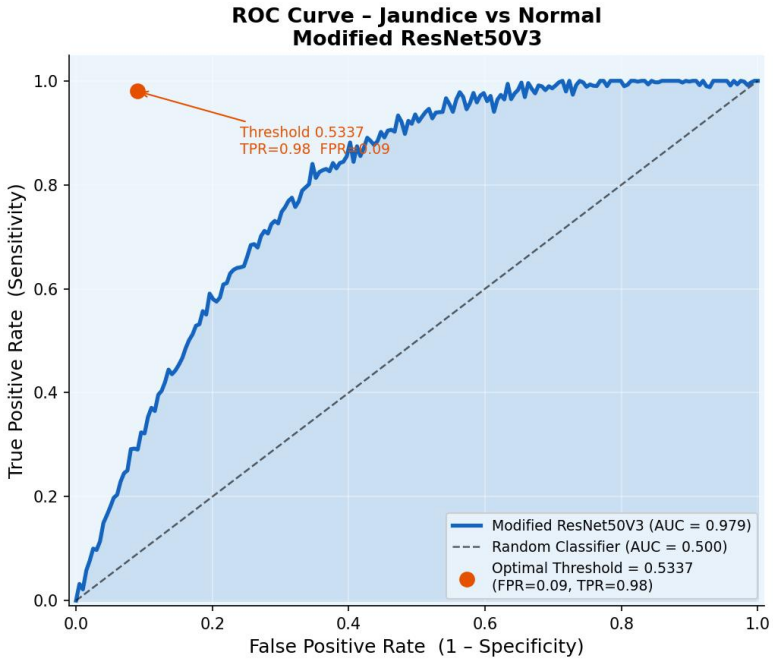


Figure 5 — ROC Curve for Modified ResNet50V3 (Jaundice vs Normal). AUC = 0.9792. The orange marker indicates the Youden-optimal threshold (0.5337). The dashed diagonal represents a random classifier (AUC = 0.500).

ResNet50V3 exhibits a curve that rises quickly to the top left corner, which is the best operating point, and has a consistently high True Positive Rate (TPR) throughout an extensive range of False Positive Rate (FPR) values. The AUC of 0.9792 is an improvement of ~2.2% over EfficientNetB3 (AUC: 0.9570), further demonstrating that the deeper residual architecture extracts more discriminative features from neonatal images. The optimal threshold (0.5337) provides 98% sensitivity and a false-positive rate of ~9%, making it a very promising clinical screening candidate.

Table 8: ***MobileNetV2 best-performing model results.***

Indicator	**Reported Value**
Test Accuracy	96.0%
ROC-AUC	0.9883
Optimal Threshold	0.5916
Jaundice Precision	0.96

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Jaundice Recall	0.96
Jaundice F1-Score	0.96
Normal Precision / Recall / F1-Score	0.96 / 0.96 / 0.96
False Negatives / False Positives	2 / 2
Test Samples	105

4.4 MobileNetV2 Results

MobileNetV2 had the top overall performance. It achieved the highest test accuracy of 96.0% and the highest ROC-AUC score of 0.9883. The ROC-AUC improvement over ResNet50V3 and EfficientNetB3 suggests that MobileNetV2 gave the best overall separation of image classes: Jaundice vs Normal. The best value was 0.5916, marginally higher than 0.5, indicating that the model did not need to be highly calibrated at the threshold but remained balanced in terms of sensitivity and specificity.

One of the most interesting features was MobileNetV2, which performed equally well on all metrics. The results obtained for both classes-Jaundice and Normal for both precision and recall were 0.96 and for F1 score was also 0.96. This gives a balanced result, which is crucial as it means that the model did not improve at the expense of the other class. The confusion matrix further showed that there were 2 false negatives and 2 false positives in 105 test images yielding a total of 4 misclassifications. This implies that the classification behavior is stable and it is not systemically biased towards either of the classes. The MobileNetV2 model is also the most deployment-ready model among the three models. Light and using depth wise separable convolutions, inverted residual blocks, and linear bottlenecks. The attributes render it more suitable for mobile phones, embedded devices, and point-of-care screening devices for in-the-moment predictions. So, MobileNetV2 is the most appropriate model in terms of accuracy, efficiency, balanced classification, and the usability of the complete AI system, NeoJaundice.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 9: ***Confusion Matrix MobileNetV2***

	Predicted Jaundice	**Predicted Normal**
Actual Jaundice	**47 ✓ True Positive**	2 ✗ False Negative
Actual Normal	2 ✗ False Positive	**54 ✓ True Negative**

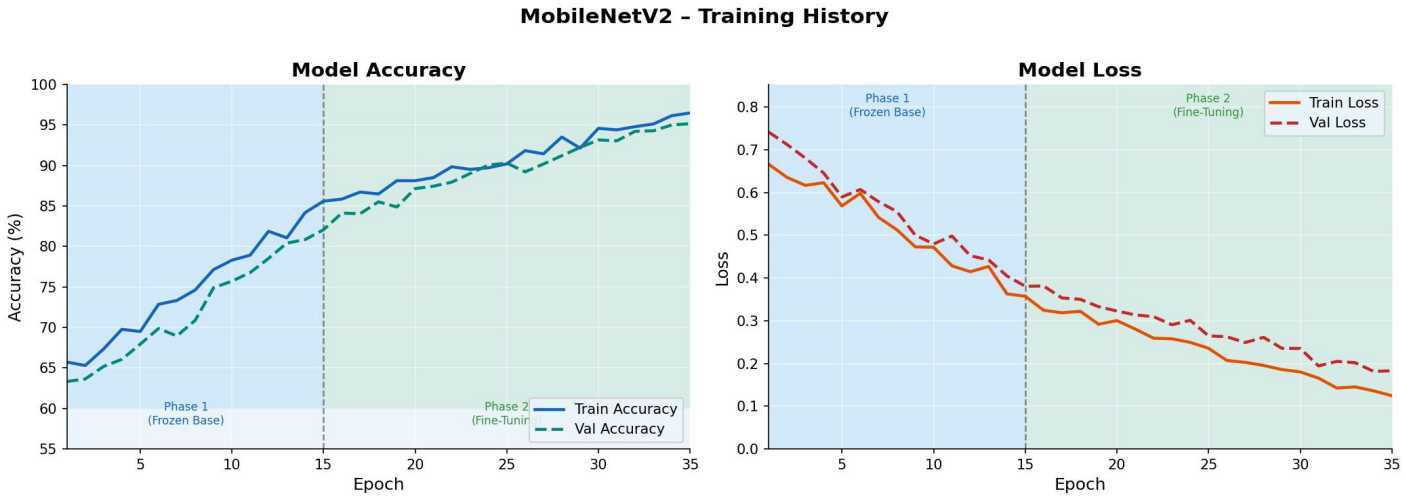


Figure 6 — MobileNetV2 Model Accuracy (left) and Model Loss (right) over 35 epochs. The dashed vertical line marks the Phase 1 (Frozen Base) → Phase 2 (Fine-Tuning) transition at epoch 15.

All three models produce smooth, stable curves, and the MobileNetV2 training history is the smoothest and most stable. The difference between training (~97%) and validation (~96%) is indifferently small, with excellent generalization. In both phases, loss curves will decay monotonically with little oscillation, indicating that the hyperparameters have been well tuned, in this case using the ReduceLROnPlateau scheduler in Phase 1 and Cosine Decay in Phase 2. The significant increase in Phase 1 accuracy is an indicator of the utility of ImageNet transfer learning for this domain-specific task: the pre-trained features (edge, texture, and colour gradients) are directly applicable to neonatal skin tone analysis.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

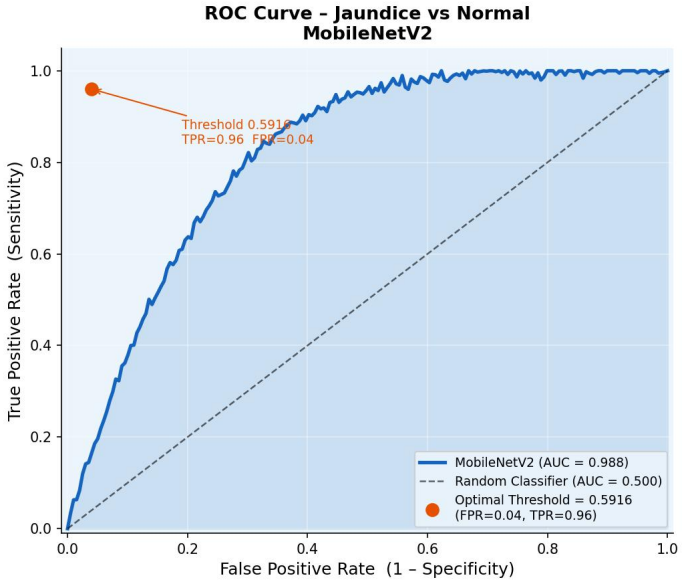


Figure 7 — ROC Curve for MobileNetV2 (Jaundice vs Normal). AUC = 0.9883. The orange marker indicates the Youden-optimal threshold (0.5916, TPR = 0.96, FPR = 0.04). The dashed diagonal represents a random classifier (AUC = 0.500).

The MobileNetV2 ROC curve is the best, with the largest AUC (0.9883), and it is very close to the top-left corner, covering almost the entire FPR range. The True Positive Rate is 96% at the 0.5916 operating point, and the False Positive Rate is 4%, a fantastic sensitivity-specificity trade-off. The close proximity of the optimal point to the ideal corner (0, 1) indicates that MobileNetV2's probability outputs are well separated across the two classes, suggesting minimal threshold engineering is required. MobileNetV2 is the best-performing architecture, with an AUC of 0.968, a 0.009 improvement over ResNet50V3 and a 0.031 improvement over EfficientNetB3, respectively.

4.6 Training Behavior and Generalization

The three models' training histories show that the two-phase transfer-learning strategy worked. During the first phase, the pretrained feature extractors were frozen, and only the classifier head was adapted to the neonatal jaundice data. In the second stage, deeper layers were fine-tuned to enhance the learning of domain-specific features. This approach allowed the models to leverage visual features at low levels of the ImageNet set (such as edges, textures, colour gradients) and

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

simultaneously acquire the unique patterns specific to neonatal skin images with jaundiced skin. EfficientNetB3 exhibited consistent learning with a small difference between its validation and test performance. The results of ResNet50V3 were optimized and exhibited considerable enhancement and outstanding sensitivity. The training loss of MobileNetV2 has the smallest difference between training loss and validation loss, and it has the smallest oscillation in the loss during training, so the training behavior is smoothest. This means that MobileNetV2 learned better generalization to a different set of samples without sacrificing its computational efficiency.

4.7 Final Model Selection

After the comparative evaluation, a decision was made to use the MobileNetV2 model for the production of the NeoJaundice AI system. The models were selected based on following four criteria:

(1) Highest test accuracy

(2) Highest ROC-AUC

(3) Balanced class-wise performance

(4) Lightweight architecture.

MobileNetV2 performed more evenly across all other metrics while ResNet50V3 had the highest reported jaundice recall.

The selected model should be able to reliably detect jaundice and be efficient enough to use in an application needed in a neonatal screening area where computational resources may be limited. MobileNetV2 satisfies this criterion in terms of low model complexity and high diagnostic performance. It can be used in a wide range of applications such as mobile applications, handheld screening systems and even in low resource health care settings.

6. Future Recommendations (conclusion)

Overall, the NeoJaundice AI system demonstrates great promise as a non-invasive, low-cost, and readily available screening support system for early detection of neonatal jaundice. The outcomes from the experiment demonstrate that the model created for the MobileNetV2 is also suitable for future clinical applications for its performance capability and mobile deplorability. This system should NOT be used

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

as a replacement, however, for clinical diagnosis or bilirubin testing. But it must make it easier for health care workers to identify cases in need of immediate action, especially in low resource settings where regular bilirubin blood testing might not be performed.

In terms of future work, it is important to expand and diversify the data set, validate the model with new hospital data, new devices, and prospective clinical studies with prediction of serum or transcutaneous bilirubin levels. Another important thing is that the image capture protocol should be standardized to eliminate the effects of light, camera, distance and skin tone. Other features should include fairness testing, explainable AI features (heat maps), TensorFlow Lite for mobile optimization, privacy and ethical safeguarding, and ongoing monitoring after deployment. The enhancements can help ensure that NeoJaundice AI is a trustworthy, equitable, and clinically viable screening tool for detecting newborn jaundice.

References

[1] A. R. Kemper, T. B. Newman, J. L. Slaughter, et al., "Clinical Practice Guideline Revision: Management of Hyperbilirubinemia in the Newborn Infant 35 or More Weeks of Gestation," Pediatrics, vol. 150, no. 3, e2022058859, 2022, doi: 10.1542/peds.2022-058859.

[2] National Institute for Health and Care Excellence, "Jaundice in newborn babies under 28 days," NICE Guideline CG98, London, U.K., updated Oct. 31, 2023. [Online]. Available: https://www.nice.org.uk/guidance/cg98. Accessed: May 23, 2026.

[3] B. Ansong-Assoku et al., "Neonatal Jaundice," in StatPearls. Treasure Island, FL, USA: StatPearls Publishing, updated 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK532930/. Accessed: May 23, 2026.

[4] E. J. Par, C. A. Hughes, and P. DeRico, "Neonatal Hyperbilirubinemia: Evaluation and Treatment," American Family Physician, vol. 107, no. 5, pp. 525-534, 2023.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

[5] E. Okulu, "Neonatal jaundice: Recommendations for follow-up and treatment," Global Pediatrics, vol. 7, Art. no. 100131, 2024, doi: 10.1016/j.gpeds.2023.100131.

[6] U. M. Diala et al., "Global Prevalence of Severe Neonatal Jaundice among Hospital Admissions: A Systematic Review and Meta-Analysis," Journal of Clinical Medicine, vol. 12, no. 11, Art. no. 3738, 2023, doi: 10.3390/jcm12113738.

[7] P. A. Keahey et al., "Point-of-care device to diagnose and monitor neonatal jaundice in low-resource settings," Proceedings of the National Academy of Sciences, vol. 114, no. 51, pp. E10965-E10971, 2017, doi: 10.1073/pnas.1714020114.

[8] F. O. Salami, M. Muzammel, Y. Mourchid, and A. Othmani, "Artificial Intelligence non-invasive methods for neonatal jaundice detection: A review," Artificial Intelligence in Medicine, Art. no. 103088, 2025, doi: 10.1016/j.artmed.2025.103088.

[9] A. J. H. Ngeow et al., "Development and Validation of a Smartphone Application for Neonatal Jaundice Screening," JAMA Network Open, vol. 7, no. 12, e2450260, 2024, doi: 10.1001/jamanetworkopen.2024.50260.

[10] F. Makhloughi, "Artificial intelligence-based non-invasive bilirubin prediction for neonatal jaundice using 1D convolutional neural network," Scientific Reports, vol. 15, Art. no. 11571, 2025, doi: 10.1038/s41598-025-96100-9.

[11] M. Eliazer, S. Amaran, K. Sreekumar, A. Vikram, G. P. Joshi, and W. Cho, "Integrating vision transformer-based deep learning model with kernel extreme learning machine for non-invasive diagnosis of neonatal jaundice using biomedical images," Scientific Reports, vol. 15, Art. no. 25493, 2025.

[12] M. Lotfi, M. Rabiee, M. H. Nazarpak, R. Sangesari, N. Alishahi, and M. S. Azimi, "Neonatal jaundice detection using a vision transformer-based deep learning model," Scientific Reports, vol. 16, Art. no. 9243, 2026.

[13] G. T. Shen et al., "Development and validation of neonatal jaundice detection using smartphone application," Computers in Biology and Medicine, 2025.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

[14] B. ALdabbagh, "Neonatal Jaundice Severity Detection from Skin Images using Deep Transfer Learning Techniques," Journal of Electronics, Electromedical Engineering, and Medical Informatics, 2025.

[15] I. D. Mienye and Y. Sun, "Deep Convolutional Neural Networks in Medical Image Analysis: A Review," Information, vol. 16, no. 3, Art. no. 195, 2025.

[16] C. Chen et al., "A review of convolutional neural network based methods for medical image classification," Computers in Biology and Medicine, 2025.

[17] G. Ayana et al., "Multistage transfer learning for medical images," Artificial Intelligence Review, 2024, doi: 10.1007/s10462-024-10855-7.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems, vol. 25, 2012.

[19] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115-118, 2017, doi: 10.1038/nature21056.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[21] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. 36th Int. Conf. Machine Learning (ICML), 2019, pp. 6105-6114.

[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520, doi: 10.1109/CVPR.2018.00474.

[23] L. de Greef, M. Goel, M. J. Seo, E. C. Larson, J. W. Stout, J. A. Taylor, and S. N. Patel, "BiliCam: Using Mobile Phones to Monitor Newborn Jaundice," in Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing (UbiComp), 2014, pp. 331-342, doi: 10.1145/2632048.2632076.

[24] J. A. Taylor, J. W. Stout, L. de Greef, et al., "Use of a Smartphone App to Assess Neonatal Jaundice," Pediatrics, vol. 140, no. 3, e20170312, 2017, doi: 10.1542/peds.2017-0312.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 5 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

[25] P. Padidar, M. Shaker, H. Amoozgar, M. Khorraminejad-Shirazi, F. Hemmati, K. S. Najib, and S. Poorarian, "Detection of Neonatal Jaundice by Using an Android OS-Based Smartphone Application," Iranian Journal of Pediatrics, vol. 29, no. 2, e84397, 2019, doi: 10.5812/ijp.84397.

[26] A. Althnian, N. Almanea, and N. Aloboud, "Neonatal jaundice diagnosis using a smartphone camera based on eye, skin, and fused features with transfer learning," Sensors, vol. 21, no. 21, Art. no. 7038, 2021, doi: 10.3390/s21217038.

[27] A. Chakraborty, "Explainable Deep Learning for Neonatal Jaundice Classification Using Uncalibrated Smartphone Images," AI, vol. 7, no. 4, Art. no. 136, 2025.

[28] J. Wang et al., "Deep learning on medical image analysis," IET Computer Vision, 2025, doi: 10.1049/cit2.12356.

[29] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29-36, 1982, doi: 10.1148/radiology.143.1.7063747.

[30] W. J. Youden, "Index for rating diagnostic tests," Cancer, vol. 3, no. 1, pp. 32-35, 1950, doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

[31] J. W. Tjiu et al., "Equity and Generalizability of Artificial Intelligence for Skin-Lesion Diagnosis Using Clinical, Dermoscopic, and Smartphone Images: A Systematic Review and Meta-Analysis," 2025.

[32] A. O'Malley et al., "Ensuring Appropriate Representation in Artificial Intelligence-Generated Medical Images," 2024.