Efficient and Explainable Convolutional Neural Networks for Neonatal Jaundice Recognition

Syllas Otutey
Department of Health Informatics
Michigan Technological University
Houghton, MI, USA
sotutey@mtu.edu

***Abstract*—Neonatal jaundice affects up to 90% of term newborns and 80% of preterm infants worldwide. While most cases resolve spontaneously, severe or delayed-recognized jaundice poses serious clinical risks, including bilirubin-induced neurological dysfunction and irreversible kernicterus. Current diagnostic pathways rely on subjective clinician visual assessment combined with serum bilirubin laboratory testing, both resource-intensive and prone to delays, particularly in low-resource healthcare settings.**

This project develops and evaluates efficient and explainable convolutional neural network (CNN) architectures for automated neonatal jaundice recognition from photographic images. Four models were investigated: a custom CNN trained from scratch, and three transfer learning models based on ResNet50, EfficientNetB0, and VGG16. The study emphasizes computational efficiency, predictive performance, and explainability via Grad-CAM. Results demonstrate that transfer learning substantially outperforms the baseline CNN, with ResNet50 achieving the best overall performance (89.47% accuracy, 92.98% F1-score, 0.9536 ROC AUC), while EfficientNetB0 offers a favorable accuracy–efficiency trade-off suitable for deployment on hospital virtual machines or mobile devices. Grad-CAM visualizations confirm that the models attend to clinically meaningful regions (facial skin and sclera), supporting use as an explainable clinical decision support tool.

***Index Terms*—neonatal jaundice, deep learning, convolutional neural networks, transfer learning, Grad-CAM, explainable AI, medical image analysis**

I. Introduction

Neonatal jaundice is a highly prevalent condition, affecting up to 90% of term newborns and 80% of preterm infants globally [1]. While most cases are benign, severe or prolonged hyperbilirubinemia can lead to bilirubin-induced neurological dysfunction and irreversible kernicterus, with lifelong neurodevelopmental consequences.

Current diagnostic pathways rely on a combination of subjective clinician visual assessment and serum bilirubin measurements. Visual assessment is inherently variable and may be unreliable across different skin tones and lighting conditions [2]. Laboratory testing, while precise, is resource-intensive and may be delayed in low-resource or high-volume clinical settings, resulting in late recognition and treatment.

Artificial intelligence (AI) and deep learning offer a promising opportunity to augment neonatal jaundice screening via non-invasive, image-based methods [3], [5]. In particular, convolutional neural networks (CNNs) can learn discriminative features from infant facial and skin images that correlate with jaundice severity. However, several challenges remain:

- Designing models that are accurate yet computationally efficient for deployment on constrained hardware.
- Ensuring that predictions are transparent and clinically interpretable through explainable AI (XAI) methods such as Grad-CAM [4].
- Addressing class imbalance and limited dataset size typical of medical imaging applications.

This project addresses these challenges by developing and comparing four CNN architectures for neonatal jaundice recognition, with a focus on efficiency and explainability. The underlying motivation is to create a deployable tool that can operate on hospital virtual machines or mobile platforms where clinical expertise and laboratory infrastructure may be limited.

II. Dataset Description

The experimental dataset comprises 760 neonatal images organized into two classes:

- Normal newborns: 560 images (73.7%).
- Jaundiced newborns: 200 images (26.3%).

This natural class imbalance reflects real-world clinical prevalence. Images were sourced from a clinical jaundice image collection and preprocessed uniformly to a standardized resolution of 224×224 pixels. Pixel values were normalized to the range [0, 1].

A. Sample Images and Class Distribution

Figure 1 illustrates sample images of normal and jaundiced infants, showing characteristic visual differences in skin and sclera coloration.

The overall class imbalance is visualized in Figure 2, which displays the proportion of normal versus jaundiced images.

B. Data Augmentation

To improve generalization and combat overfitting, several data augmentation techniques were applied during training:

- Rotation: $\pm 15^{\circ}$.
- Width and height shifts: up to 10%.
- Brightness jittering.
- Random horizontal flipping.

Fig. 1: Sample images from the dataset showing characteristic visual differences between normal newborns and those with jaundice.

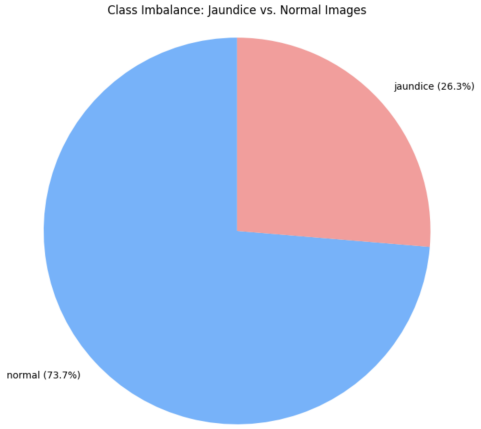


Fig. 2: Class distribution of the dataset: 560 normal (73.7%) vs. 200 jaundiced (26.3%) images.

Augmentation was applied only to the training set, preserving the integrity of validation data for unbiased performance evaluation. This design reflects clinical realism by simulating variability in lighting, infant positioning, and imaging angles without distorting the subtle yellow pigmentation patterns critical for diagnosis.

III. Exploratory Data Analysis

Exploratory data analysis (EDA) characterized dataset composition, imaging conditions, and class balance.

A. Variability in Imaging Conditions

Figure 3 highlights variability in lighting, skin tone, and background conditions.

All images contain frontal or semi-frontal views of the infant head/torso with varying illumination and backgrounds. Jaundiced images display yellow–orange discoloration of skin and sclera, whereas normal images show pink-to-pale skin tones.

B. Class Balance and Split Distribution

The initial dataset contained 760 images:

- Normal: 560 images (73.7%).
- Jaundice: 200 images (26.3%).

Fig. 3: Variability in dataset conditions including lighting, skin tone, and background.

A stratified split was used to preserve this ratio in training and validation subsets. Figure 4 summarizes the distribution across splits.

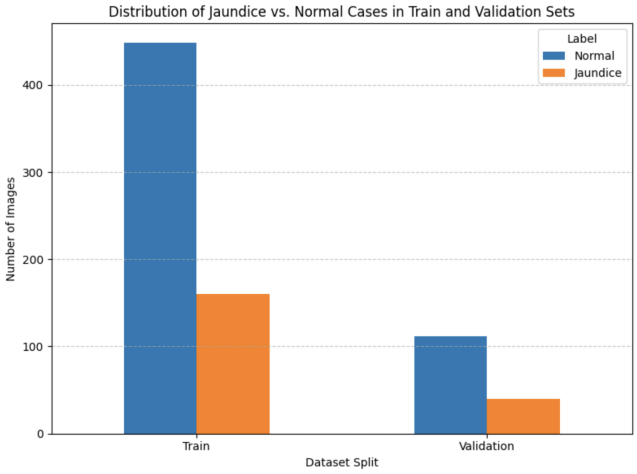


Fig. 4: Distribution of normal and jaundiced images across training and validation subsets (stratified split).

IV. Preprocessing Pipeline

All images were resized to 224×224 pixels using bilinear interpolation. Pixel values were scaled to [0, 1] and then processed with model-specific preprocessing functions:

- ResNet50: ImageNet-style mean subtraction and standard deviation scaling.
- EfficientNetB0: EfficientNet-specific preprocessing.
- VGG16: VGG-specific preprocessing.
- Simple CNN: Direct [0, 1] normalization.

A. Augmentation Examples

Figure 5 shows representative augmentation samples.

Aggressive geometric distortions were intentionally avoided to maintain diagnostically relevant skin and sclera patterns.

V. CNN Model Design

Four CNN architectures were designed and trained in parallel to enable fair comparative evaluation.

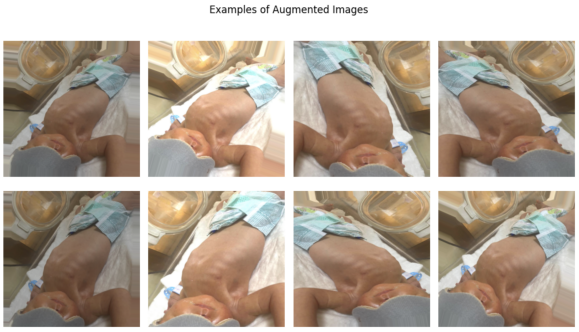


Fig. 5: Data augmentation examples (rotation, shift, brightness adjustment, horizontal flip).

A. Model A: Simple CNN (Baseline)

A custom CNN trained from scratch served as a baseline:

- Input: $224 \times 224 \times 3$.
- Three Conv2D + MaxPooling2D blocks (32, 64, 128 filters).
- Flatten (86,528 features).
- Dense (256 units, ReLU) + Dropout (0.4).
- Output Dense (1 unit, sigmoid).

Total parameters: 22,244,929 ($\approx$ 84.86 MB).

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 222, 222, 32)	896
max_pooling2d (MaxPooling2D)	(None, 111, 111, 32)	0
conv2d_1 (Conv2D)	(None, 109, 109, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 54, 54, 64)	0
conv2d_2 (Conv2D)	(None, 52, 52, 128)	73,856
max_pooling2d_2 (MaxPooling2D)	(None, 26, 26, 128)	0
flatten (Flatten)	(None, 86528)	0
dense (Dense)	(None, 256)	22,151,424
dropout (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 1)	257

Total params: 22,244,929 (84.86 MB)
Trainable params: 22,244,929 (84.86 MB)
Non-trainable params: 0 (0.00 B)

Fig. 6: Model summary for the custom Simple CNN baseline (from `model.summary()`).

B. Model B: ResNet50 (Transfer Learning)

ResNet50 was employed with ImageNet-pretrained weights and a frozen convolutional base:

- Base: ResNet50 (frozen).
- GlobalAveragePooling2D.
- Dropout (0.3).
- Dense (128 units, ReLU).
- Dropout (0.3).
- Dense (1 unit, sigmoid).

Total parameters: 23,850,113; trainable parameters: 262,401.

Model: "sequential_1"

Layer (type)	Output Shape	Param #
resnet50 (Functional)	(None, 7, 7, 2048)	23,587,712
global_average_pooling2d (GlobalAveragePooling2D)	(None, 2048)	0
dropout_1 (Dropout)	(None, 2048)	0
dense_2 (Dense)	(None, 128)	262,272
dropout_2 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 1)	129

Total params: 23,850,113 (90.98 MB)
Trainable params: 262,401 (1.00 MB)
Non-trainable params: 23,587,712 (89.98 MB)

Fig. 7: Model summary for the ResNet50-based transfer learning architecture.

C. Model C: EfficientNetB0 (Transfer Learning)

EfficientNetB0 provided a lightweight, parameter-efficient architecture:

- Base: EfficientNetB0 (frozen).
- GlobalAveragePooling2D.
- Dropout (0.3).
- Dense (128 units, ReLU).
- Dropout (0.3).
- Dense (1 unit, sigmoid).

Total parameters: 4,213,668 ($\approx$ 16.07 MB); trainable: 164,097.

Model: "sequential_2"

Layer (type)	Output Shape	Param #
efficientnetb0 (Functional)	(None, 7, 7, 1280)	4,049,571
global_average_pooling2d_1 (GlobalAveragePooling2D)	(None, 1280)	0
dropout_3 (Dropout)	(None, 1280)	0
dense_4 (Dense)	(None, 128)	163,968
dropout_4 (Dropout)	(None, 128)	0
dense_5 (Dense)	(None, 1)	129

Total params: 4,213,668 (16.07 MB)
Trainable params: 164,097 (641.00 KB)
Non-trainable params: 4,049,571 (15.45 MB)

Fig. 8: Model summary for the EfficientNetB0-based transfer learning architecture.

D. Model D: VGG16 (Transfer Learning)

VGG16 was evaluated as a deep, sequential architecture:

- Base: VGG16 (frozen).
- GlobalAveragePooling2D.
- Dropout (0.3).
- Dense (128 units, ReLU).
- Dropout (0.3).
- Dense (1 unit, sigmoid).

Total parameters: 14,780,481 ($\approx$ 56.38 MB); trainable: 65,793.

VI. Hyperparameters and Training Strategy

All models shared a consistent core configuration for fair comparison:

Model: "sequential_3"

Layer (type)	Output Shape	Param #
vgg16 (Functional)	(None, 7, 7, 512)	14,714,688
global_average_pooling2d_2 (GlobalAveragePooling2D)	(None, 512)	0
dropout_5 (Dropout)	(None, 512)	0
dense_6 (Dense)	(None, 128)	65,664
dropout_6 (Dropout)	(None, 128)	0
dense_7 (Dense)	(None, 1)	129

Total params: 14,780,481 (56.38 MB)
Trainable params: 65,793 (257.00 KB)
Non-trainable params: 14,714,688 (56.13 MB)

Fig. 9: Model summary for the VGG16-based transfer learning architecture.

- Batch size: 32.
- Optimizer: Adam (learning rate 0.001).
- Loss function: binary cross-entropy.
- Hidden activation: ReLU.
- Output activation: sigmoid.
- Epochs: up to 20.
- Early stopping: patience of 5 epochs on validation loss (restoring best weights).
- Model checkpointing: best model saved by validation loss.

A `ReduceLROnPlateau` callback further reduced the learning rate by a factor of 0.5 (minimum 0.0001) when validation loss plateaued. This was especially helpful for the Simple CNN and VGG16 models.

Rather than exhaustive grid search, a principled tuning strategy leveraged robust transfer learning initialization, conservative batch sizes, early stopping, and dropout regularization (0.3 for transfer models, 0.4 for the Simple CNN).

VII. Training and Validation Curves Analysis

Understanding the training and validation behavior of all four models is crucial for evaluating how well they learn, generalize, and avoid overfitting. The combined accuracy and loss plots provide a clear visual comparison of performance dynamics.

A. Simple CNN — Unstable Learning and Poor Generalization

For the Simple CNN, training accuracy increases only modestly and the validation accuracy fluctuates sharply, with noticeable drops at several epochs. Training loss decreases initially but plateaus at a relatively high value, while validation loss exhibits spikes.

These patterns indicate that the Simple CNN:

- struggles to extract robust hierarchical features,
- is highly sensitive to the small dataset size, and
- shows both underfitting (limited capacity) and instability.

This behavior is typical of shallow networks applied to complex medical imaging tasks.

B. ResNet50 — Smooth, Consistent Convergence

ResNet50 exhibits the most stable convergence pattern:

- Training and validation accuracy rise steadily and remain close to each other.
- Training and validation loss decrease smoothly with no major divergence.

The close alignment of curves suggests strong generalization and minimal overfitting. Residual connections and ImageNet pretraining contribute to this stability. These learning dynamics are consistent with ResNet50's superior performance metrics.

C. EfficientNetB0 — Stable and Efficient Learning

EfficientNetB0 also shows smooth, well-aligned training and validation curves:

- Accuracy curves increase in parallel, with validation accuracy often matching or slightly exceeding training accuracy.
- Loss curves decrease to low values (≈ 0.28–0.32) without spikes.

This indicates excellent generalization, good calibration, and efficient use of parameters, reinforcing EfficientNetB0 as a strong deployment candidate on resource-limited hardware.

D. VGG16 — Good Learning with Mild Underfitting

VGG16 improves steadily in early epochs but demonstrates:

- validation accuracy that is sometimes higher than training accuracy, and
- validation loss that plateaus earlier than for ResNet50 or EfficientNetB0.

This pattern suggests mild underfitting and less efficient convergence. The large parameter count and frozen base layers limit how much adaptation is possible on this dataset.

E. Overall Interpretation

From the combined curves:

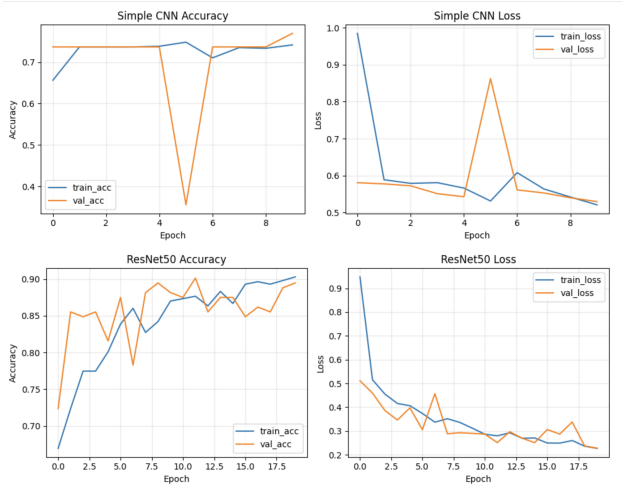
- ResNet50 and EfficientNetB0 show the most reliable and stable learning behavior, confirming their suitability for jaundice classification.
- The Simple CNN is clearly the weakest performer, with unstable validation behavior and limited feature extraction capacity.
- VGG16 performs reasonably well but is less efficient and slightly less stable than the other transfer-learning models.

VIII. Model Performance Results

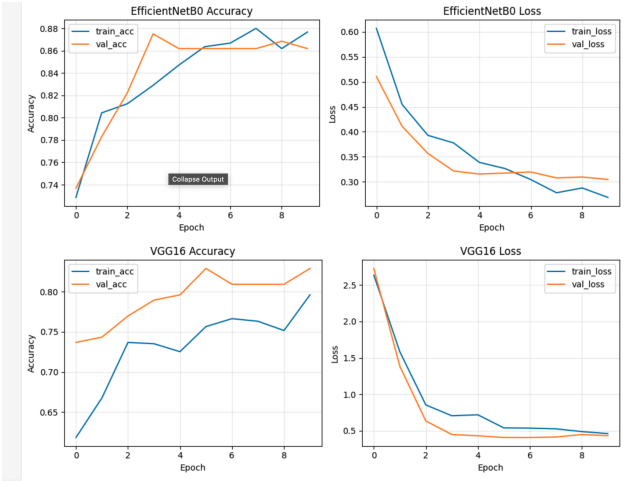
A. Validation Performance Summary

All models were evaluated on the validation set using accuracy, precision, recall, F1-score, and ROC AUC. Table I summarizes the results (values from the 80/20 stratified split used in this report).

Transfer learning models clearly outperform the baseline Simple CNN. ResNet50 achieves the highest accuracy and F1-score, while EfficientNetB0 offers a strong efficiency–accuracy trade-off.



(a) Training/validation curves for the Simple CNN and ResNet50.



(b) Training/validation curves for EfficientNetB0 and VGG16.

Fig. 10: Training and validation accuracy/loss curves for all four models.

TABLE I: Validation Performance of Evaluated Models

Model	Acc.	Prec.	Recall	F1	AUC
Simple CNN	0.7632	0.7923	0.9196	0.8512	0.7252
ResNet50	0.8947	0.9138	0.9464	0.9298	0.9536
EfficientNetB0	0.8618	0.9027	0.9107	0.9067	0.9250
VGG16	0.8224	0.8696	0.8929	0.8811	0.8895

B. Confusion Matrices

Confusion matrices were computed for all four models to analyze error patterns. Figure 11 presents the matrices side by side.

The Simple CNN shows a higher false-positive rate (normal infants misclassified as jaundiced). ResNet50 and EfficientNetB0 strike a better balance between sensitivity and specificity, with fewer misclassifications overall.

C. Classification Report

Detailed class-wise precision, recall, and F1-score for ResNet50 are visualized in Figure 12.

D. ROC Curves

ROC curves were generated for all models, as illustrated in Figure 13.

ResNet50 achieves the highest AUC (0.9536), followed closely by EfficientNetB0 (0.9250), confirming strong discriminative ability.

IX. Model Comparison and Statistical Testing

To assess whether performance differences between ResNet50 and EfficientNetB0 were statistically significant, McNemar's test was applied using paired validation predictions. The contingency table summarizing disagreements between the two models is shown in Figure 14.

The resulting chi-squared statistic was 0.640 with a p-value of 0.424, indicating no statistically significant difference ($p > 0.05$). Although ResNet50 achieved slightly higher accuracy, the misclassification patterns of both models are statistically comparable.

From a clinical perspective, this supports EfficientNetB0 as a pragmatic deployment choice, given its substantially smaller model size and faster inference.

X. Explainability via Grad-CAM

A. Grad-CAM Methodology

Grad-CAM (Gradient-weighted Class Activation Mapping) was integrated for ResNet50 and EfficientNetB0 to generate visual explanations of model predictions [4]. The method computes gradients of the predicted class score with respect to feature maps in a convolutional layer, producing a spatial heatmap highlighting regions most influential for the prediction.

B. Grad-CAM Visualizations

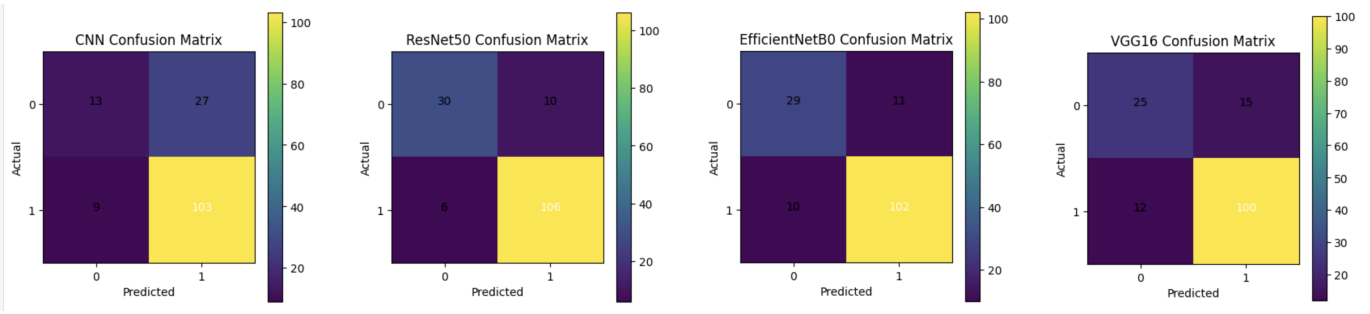
For correctly classified jaundiced cases, both ResNet50 and EfficientNetB0 exhibit strong activation on the face, forehead, sclera, and upper chest—regions clinically associated with jaundice. For normal cases, activation patterns are more diffuse with lower intensity, reflecting absence of pathological discoloration.

These patterns suggest that the models leverage clinically meaningful visual cues rather than spurious background artifacts, enhancing trust and interpretability for clinical end-users.

XI. Critical Analysis and Insights

Several key findings emerged from this project:

1) **Transfer learning dominates.** All three transfer learning models outperformed the Simple CNN by 13–17 percentage points in accuracy, underscoring the value of ImageNet pre-training for small medical datasets.



(a) Simple CNN (b) ResNet50 (c) EfficientNetB0 (d) VGG16

Fig. 11: Confusion matrices for all four models on the validation set.

Classification Report — ResNet50

Metric	Value
Accuracy	0.8947
Precision	0.9211
Recall	0.9375
F1-score	0.9292
ROC AUC	0.9571

Fig. 12: Classification report for ResNet50 showing precision, recall, F1-score, and support per class.

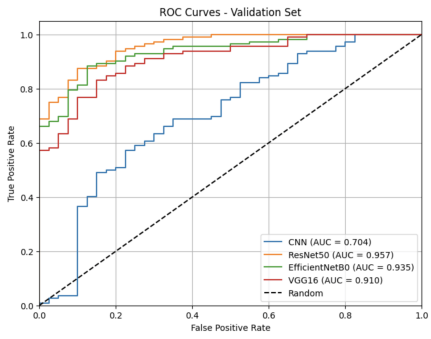


Fig. 13: ROC curves for Simple CNN, ResNet50, EfficientNetB0, and VGG16 with corresponding AUC values.

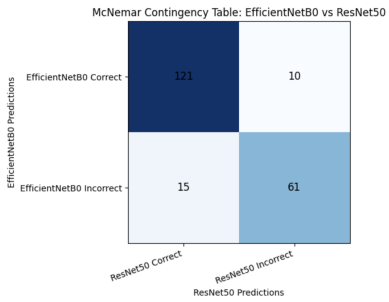


Fig. 14: McNemar contingency table comparing ResNet50 and EfficientNetB0 predictions.

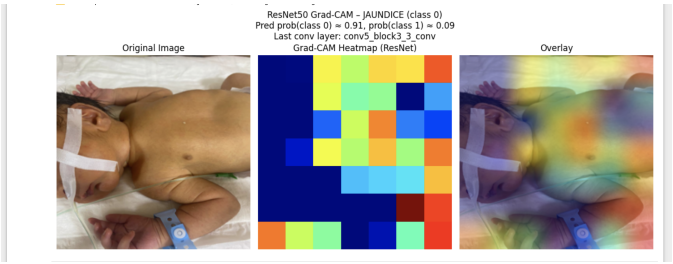


Fig. 15: Grad-CAM heatmap for a jaundiced infant, showing strong activation over facial skin and periocular regions.

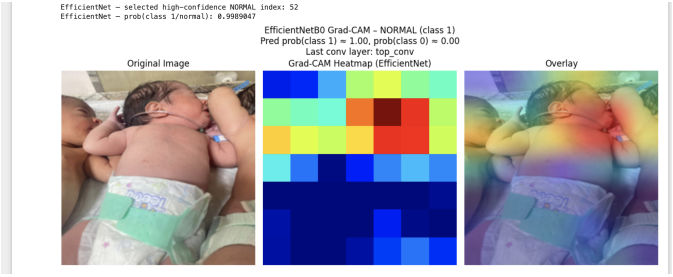


Fig. 16: Grad-CAM heatmap for a normal infant, showing more diffuse activation and absence of focal high-intensity regions.

2) **ResNet50 is optimal for accuracy.** With 89.47% accuracy, 92.98% F1-score, and 0.9536 AUC, ResNet50 delivers the strongest overall performance and stable training dynamics.
3) **EfficientNetB0 is optimal for deployment.** A 3-point accuracy sacrifice yields a 5.7× reduction in model size (16.07 MB vs. 90.98 MB) and faster inference—highly advantageous for deployment on mobile or resource-limited systems.
4) **Class imbalance impacts calibration.** The 73.7%/26.3% normal-to-jaundice ratio biased models toward normal predictions, with higher precision for normal cases than recall for jaundiced cases. Stratified splitting and binary cross-entropy mitigated but did not eliminate this effect; future work should explore class weighting or oversampling.

5) **Statistical parity of ResNet50 and EfficientNetB0.** McNemar's p-value of 0.424 indicates statistically equivalent misclassification patterns, further supporting EfficientNetB0 as a clinically viable, efficient alternative.

Clinically, all models achieved $>89\%$ recall on jaundiced cases, limiting false negatives—the most critical error type. False positives, while non-trivial (9–21% depending on model), translate into additional follow-up testing and are therefore more acceptable from a safety standpoint.

XII. Conclusion

This project successfully developed and evaluated four CNN architectures for automated neonatal jaundice recognition from photographic images. ResNet50 emerged as the accuracy leader, while EfficientNetB0 provided the best efficiency–accuracy compromise for practical deployment. Transfer learning proved essential, dramatically improving performance over a from-scratch CNN baseline.

Grad-CAM visualizations demonstrated that the models focus on clinically relevant regions, providing transparent explanations suitable for clinical decision support. McNemar's statistical testing further indicated that ResNet50 and EfficientNetB0 achieve statistically equivalent performance despite modest accuracy differences.

Overall, the study demonstrates the feasibility of efficient, explainable deep learning systems for neonatal jaundice screening, particularly in low-resource settings where laboratory infrastructure is limited. Future work should expand dataset diversity, incorporate multi-center data, explore multi-task learning for bilirubin regression, and conduct prospective clinical validation.

References

References

[1] T. W. R. Hansen *et al.*, "Narrative review of the epidemiology of neonatal jaundice," *Translational Pediatrics*, vol. 10, no. 1, pp. 1–6, 2021. [Online]. Available: https://doi.org/10.21037/pm-21-4

[2] StatPearls, "Neonatal jaundice," *StatPearls Publishing*, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK532930/

[3] "Assessment, management, and incidence of neonatal jaundice in near-term neonates," *Scientific Reports*, vol. 12, 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-17612-y

[4] "Grad-CAM-based explainable artificial intelligence related to clinical medicine," *Journal of Biomedical Informatics*, 2023. [Online]. Available: https://doi.org/10.1016/j.jbi.2023.104360

[5] "Artificial intelligence non-invasive methods for neonatal jaundice detection," *Scientific Reports*, 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-10321-4