Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Image-Based Bilirubin Estimation for Non-Invasive Neonatal Jaundice Assessment Using Swin Transformer V2

Hanan Muhajab

hmuhajab@jazanu.edu.sa

Jazan University

Areej Muhajab

Taif University

Article

Keywords: Neonatal Jaundice, Bilirubin Estimation, Non-Invasive Assessment, Swin Transformer V2, Deep Learning, Grad-CAM++

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10963208/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Image-Based Bilirubin Estimation for Non-Invasive Neonatal Jaundice Assessment Using Swin Transformer V2

Author-1 :

[1]Hanan Muhajab,

Department of Computer Science, College of Engineering and Computer Sciences,

Jazan University, Jazan, Saudi Araiba

[2]Engineering and Technology Research Center, Jazan University, P.O.Box 114,Jazan 82817, Saudi Arabia

hmuhajab@jazanu.edu.sa

https://orcid.org/0009-0000-3472-3723

Author-2 :

Areej Muhajab

Department of Information Technology,

College of Computer and Information Technology,

Taif University, P.O. Box 11099

Taif 21944, Saudi Arabia

Areej.n@tu.edu.sa

Abstract

Neonatal jaundice, is a prevalent and generally a benign condition in neonates that involves elevated levels of bilirubin, resulting in the yellowish discoloration of the skin and eyes. Accurate and timely measurement of bilirubin concentration is required to diagnose jaundiced neonates and to prevent serious complications resulting from untreated hyperbilirubinemia. Conventional bilirubin estimation methods commonly involve invasive blood sampling, indicating the need for reliable, accessible, and non-invasive image-based methods for bilirubin assessment. Therefore, there is a pressing need for a non-invasive, rapid, accessible, and explainable method for early bilirubin estimation and jaundice assessment. This study presents a Non-Invasive Bilirubin Estimation and Assessment Model (N-BEAM) for neonatal jaundice using Swin Transformer V2 with an explainable deep learning strategy. The presented N-BEAM model aims to estimate continuous Total Serum Bilirubin (TSB) values and provide a four-level assessment of neonatal jaundice without requiring blood samples during inference. Swin Transformer V2 is employed to extract hierarchical deep features from the preprocessed

neonatal skin images through window-based self-attention and shifted-window mechanisms. Besides, global average pooling and a fully connected regression network is employed to predict the continuous TSB value in mg/dL. Subsequently, the predicted TSB value is converted into four study-defined categories, namely Normal, Mild, Moderate, and Severe, providing an interpretable bilirubin assessment without using a separate classification model. Moreover, Grad-CAM++ is integrated into the proposed framework to provide visual interpretability by highlighting the neonatal skin regions that contribute to the predicted TSB value. To validate the effectiveness of the proposed N-BEAM model, extensive experiments are conducted using two independent neonatal jaundice image datasets, including the NJN dataset and the NeoJaundice dataset, under a cross-dataset evaluation protocol in which one dataset is used for training and the other as an independent testing dataset. The experimental results infer the capability of the N-BEAM framework to provide non-invasive bilirubin estimation, severity assessment, and visually interpretable predictions from neonatal images.

Keywords: Neonatal Jaundice; Bilirubin Estimation; Non-Invasive Assessment; Swin Transformer V2; Deep Learning; Grad-CAM++

1. Introduction

Neonatal jaundice refers to a prevalent condition in newborns with abnormally elevated levels of total serum bilirubin (TSB). Though normal physiologic jaundice has an ephemeral course, pathological accumulation of bilirubin can potentially result in irreversible brain damage [1]. Hence, efficient and effective assessment of bilirubin levels is crucial. Traditional evaluation of bilirubin requires invasive TSB measurement with blood withdrawal, which often makes newborns feel uncomfortable in a clinical environment, and a trained professional is needed for proper diagnosis [2]. Transcutaneous bilirubinometry is a non-invasive method, in which readings can be obtained with no blood test; however, this has been influenced by other factors like skin pigmentation, skin condition, lighting conditions, instrument-related factors, cost, and access to such tools [3]. Given the limitations of the existing methods in the field of image-based estimation, this framework seeks a non-invasive way for the assessment of bilirubin. Artificial intelligence (AI) and machine learning (ML) methods for newborn jaundice detection using images have recently received considerable attention [4]. In the previously proposed approaches, hand-crafted features (color, texture, and statistical features) are captured from the images of infant skin and faces and then used to predict the bilirubin level using classification techniques such as K-nearest neighbors (KNN), support vector machines (SVM), and random

forests (RF) [5]. Though these approaches proved skin appearance to be a key marker for jaundice detection, their performance depends greatly on lighting conditions, skin quality, and imaging equipment due to the extraction of hand-crafted features [6].

Deep learning (DL) can automate feature extraction, resulting in improved performance. Most medical image analysis methods employ convolutional neural networks (CNNs), and transfer learning has also been employed for jaundice assessment [7]. Vision Transformer (ViT)-based models are a recent class of models that effectively utilize attention-based mechanisms for processing images and demonstrate promising performance in medical image applications [8]. Swin Transformer V2 is one promising vision transformer architecture that employs a hierarchical design by introducing shifted-window-based self-attention [9]. With this approach, both fine-grained local image information and higher-level structural relationships between patches can be extracted. Such capability is particularly important for the assessment of neonatal jaundice [10]. In existing work on using images to analyze skin color appearance for bilirubin prediction, existing studies primarily focus on jaundice detection as binary or multi-class classification tasks instead of estimating bilirubin concentration [11]. Moreover, these studies evaluate the models primarily using data from the same dataset used for model development and rarely perform cross-dataset evaluation under varying acquisition conditions and different data distributions for more robust and generalizable validation [12, 13].

To address these limitations, a Non-Invasive Bilirubin Estimation and Assessment Model (N-BEAM) is presented for the non-invasive estimation of bilirubin concentration and assessment of jaundice severity. Specifically, the N-BEAM framework first extracts the neonatal skin region, preprocesses the images, and utilizes Swin Transformer V2 for learning deep hierarchical representations from these preprocessed images. The extracted features are then fed to a regression head to predict the continuous bilirubin concentration, which is further mapped to four study-defined assessment categories, namely normal, mild, moderate, and severe jaundice. Explainable artificial intelligence (XAI) is additionally integrated into the framework to demonstrate the regions contributing to the predicted bilirubin level. Cross-dataset experiments are also performed to demonstrate the generalization ability of the framework on an independent neonatal image dataset. The major contributions of the proposed N-BEAM framework are as follows:

- A non-invasive image-based bilirubin estimation framework is developed for neonatal jaundice assessment, eliminating the need for blood-based input during prediction.

- Swin Transformer V2 is employed as the deep feature extraction backbone to learn hierarchical and spatially distributed visual representations from neonatal skin images.
- Continuous TSB regression is formulated instead of directly performing jaundice classification, enabling the framework to estimate bilirubin concentration in mg/dL and subsequently derive four TSB-based assessment categories.
- Cross-dataset evaluation is conducted using two independent neonatal image datasets, with one dataset used for training and the other as an independent testing dataset, to examine the applicability and generalization of the proposed framework across different image collections.
- Grad-CAM++-based interpretability is incorporated to highlight the neonatal skin regions contributing to the predicted TSB value, improving the interpretability of the model.
- Comprehensive experimental and comparative analyses are performed to evaluate the effectiveness of N-BEAM in bilirubin estimation, TSB-based assessment, cross-dataset performance, and explainability.

2. Related Work

Recent studies have explored DL-based image analysis techniques for the detection of neonatal jaundice. Lotfi et al. [14] proposed a DL-based system called token-to-token vision transformer (T2T-ViT) with an effort to screen neonatal jaundice using image modality. It achieved better results compared to the conventional models when evaluated using a variety of metrics, proving an adequate prototype for jaundice screening in resource-scarce hospital settings. Akcan et al. [15] designed a new combined model from ViT and ML-based techniques to detect newborn jaundice. The ViT model can be differentiated using ML-based techniques like KNN, RF, and multilayer perceptron, and their results have been tested. For jaundice detection, a novel methodology is applied in a clinical setting that helps to better detect jaundice. Patel et al. [16] introduced an AI-based framework for neonatal jaundice detection called NeoJaundice-AI that uses images of an infant's skin to analyze jaundice and forecast the severity level within a second, without internet access. This method is developed using a dual-branch EfficientNet-B0 system, which individually examines skin images and addresses the challenges of limited reliable data in neonatal jaundice assessments.

Gogulakrishnan et al. [17] developed a transformer learning and test-time augmentation approach that incorporates CNN, MobileNetV2, and other DL techniques for neonatal jaundice detection. This study concluded that DL techniques provide promising results. Akarapanuvitaya et al. [18] engineered a modern model that comprises image preprocessing and ML-based detection techniques to support immediate detection and diagnosis of jaundice using image modality. Krishnaraj et al. [19] a transformer-based model utilizing ConvNeXt-Tiny and Focal Loss (CNXt-FL) was introduced for flexible, robust, and efficient detection. Premalatha et al. [20] developed a completely non-invasive DL approach that employs image processing and metadata to fuse features such as color attributes extracted from images. Finally, the method is trained and evaluated using an Adam optimizer, which provides an affordable solution to clinical jaundice detection. Salami et al. [21] designed an effective dual approach using visual transformers (JauNViT) for detecting and diagnosing newborn jaundice. The first stage deals with deep feature extraction from images using ViT, and the second stage utilizes an adapted fine-tuned pre-trained model on limited labeled data.

2.1. Research Gap

Although AI- and DL-based methods have already been studied for image-based neonatal jaundice assessment, several limitations still remain. The majority of previous works treat jaundice detection as a binary or multi-class classification problem and therefore provide limited estimation of the continuous bilirubin value. Moreover, traditional methods rely on handcrafted features such as skin color, texture, and statistical characteristics, which may fail to adequately describe image variations associated with different bilirubin levels. In addition, effective modeling of continuous bilirubin prediction using advanced vision transformer-based models remains limited. Furthermore, the generalization capability of existing jaundice assessment studies can be affected by variability among individuals due to differences in skin characteristics, illumination, and camera properties, which can affect image quality, and variations in data distributions arising from different devices and acquisition environments. Moreover, the lack of transparency in DL systems is another challenge, as there is limited knowledge about which regions of an input image influence a given prediction. Given these limitations, there is a need for a unified DL framework that not only predicts continuous bilirubin concentration but also provides corresponding TSB-based assessment, generalizes across different neonatal image datasets, and explains its predictions. To address these limitations, a N-BEAM is proposed, integrating Swin Transformer V2 for hierarchical deep

feature extraction, continuous TSB prediction, four-level TSB-based assessment, and Grad-CAM++ for visual interpretation of the predictions across two neonatal image datasets.

3. Materials and Methods

This section presents the N-BEAM approach for non-invasive TSB estimation and neonatal jaundice assessment from images. The dataset descriptions, overall architecture, preprocessing steps, Swin Transformer V2-based feature extraction, TSB regression, and Grad-CAM++-based explainability are described in the following subsections.

3.1. Dataset Description

The proposed N-BEAM method was evaluated using the NeoJaundice: Neonatal Jaundice Evaluation in Demographic Images dataset (Dataset 1) [22] and NJN: A Dataset for the Normal and Jaundiced Newborns (Dataset 2) [23]. A short description of the two datasets is provided below.

- **Dataset 1:** The dataset was gathered from Xuzhou Central Hospital and includes 745 infants with 2,235 image instances. The binary labels were initially generated using TSB measurements obtained from medical records, with a threshold of 12.9 mg/dL, and were further validated by an experienced dermatologist. Representative neonatal skin images corresponding to the TSB-based severity levels considered in this study are demonstrated in Fig. 1.



Fig. 1. Representative neonatal skin images from the NeoJaundice training dataset corresponding to the four TSB-based severity levels: (a) Normal, (b) Mild, (c) Moderate, and (d) Severe.

- **Dataset 2:** The dataset was collected from Al-Elwiya Maternity Teaching Hospital in Al Rusafa, Baghdad, Iraq. It contains images of jaundiced and normal infants captured from diverse angles and under varying lighting conditions. The dataset includes images from approximately 600 newborns, comprising 760 infant images, including 200 jaundiced and 560 normal images. Representative neonatal images from the NJN dataset are presented in Fig. 2.



Fig. 2. Representative neonatal images from the NJN dataset used as the independent testing dataset: (a) Normal and (b) Jaundiced

3.2. Overall Architecture of N-BEAM

The proposed N-BEAM is designed to estimate bilirubin concentration and assess neonatal jaundice from neonatal images using a non-invasive image-based approach. The main architecture consists of image preprocessing, transformer-based deep feature extraction, continuous bilirubin prediction through regression, TSB-based assessment of the predicted bilirubin value, and XAI. The complete architectural workflow is shown in Fig. 3. Firstly, the neonatal image is processed to extract the region of interest (ROI), where only the neonatal skin region is retained while removing the color card and irrelevant background regions. The extracted ROI is then resized to 224 × 224 pixels and subjected to pixel-intensity normalization. Finally, the preprocessed neonatal skin image is fed into Swin Transformer V2 to extract hierarchical low- and high-level features through attention mechanisms that encode interactions among local skin characteristics and spatial correlations across larger regions of the skin.

The deep features generated by Swin Transformer V2 are aggregated using a global average pooling (GAP) layer and then passed through fully connected (FC) layers with dropout layers in between to form the regression head, which produces a continuous predicted TSB value in mg/dL. The regression head is trained using the MSE loss. The predicted TSB value is then mapped to four predefined assessment levels, namely normal, mild, moderate, and severe, without requiring a separate classification head.

The Grad-CAM++ method is employed as a post-hoc explanation technique to visualize the important skin regions contributing to the predicted TSB value for an input neonatal image. Thus, N-BEAM generates three complementary outcomes: a continuous TSB measurement, a TSB-based assessment, and a visual explanation of the prediction.

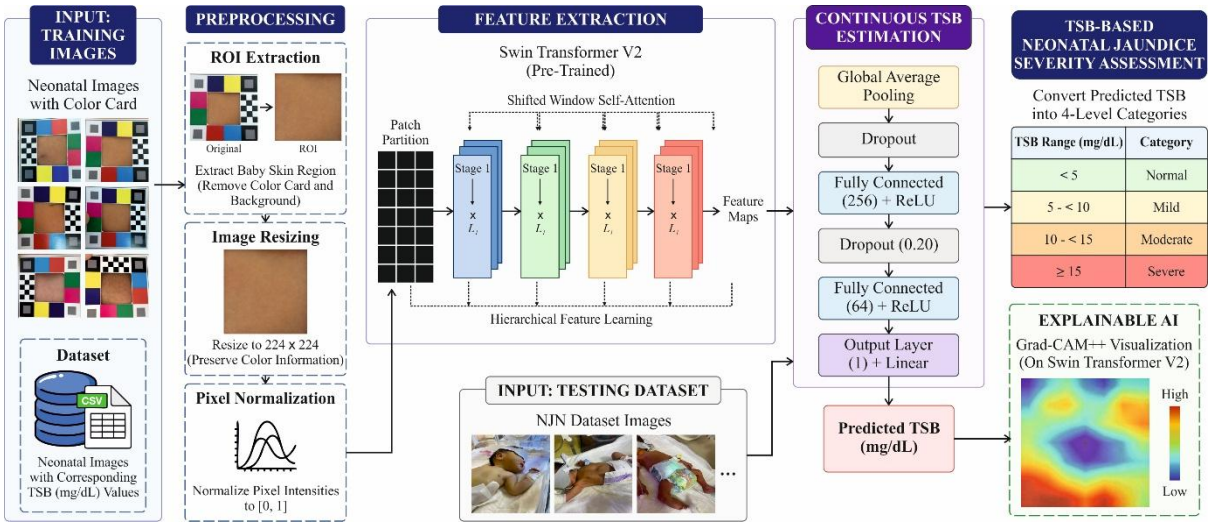


Fig. 3. Overall workflow of the proposed N-BEAM framework for non-invasive TSB estimation and TSB-based neonatal jaundice severity assessment

3.3. Data Preprocessing

The input neonatal images may exhibit variations in background content, image dimensions, illumination, and pixel intensity distributions because of differences in image acquisition conditions. To reduce these unwanted variations and obtain a consistent representation of the relevant neonatal skin region, three preprocessing techniques are applied to the neonatal images to obtain uniformly extracted skin regions and eliminate unwanted sources of variability [24, 25].

- **ROI Extraction:** Firstly, ROI extraction is performed. It aims to remove the outer color card, background, and any unnecessary regions, while focusing only on the neonatal

skin. This makes the model input focused on the visual features of neonatal skin related to bilirubin levels.

- **Image Resizing:** Secondly, the extracted ROI is resized to 224 × 224 pixels. The skin ROIs are thus standardized to a uniform size of 224 × 224 pixels. The original spatial information is preserved to provide a consistent input to Swin Transformer V2, while retaining the R, G, and B color channels.
- **Pixel Intensity Normalization:** Thirdly, pixel intensity normalization is performed. The R, G, and B images contain 8-bit pixel intensity values from 0-255. Each pixel intensity I is normalized to I_{norm} as:

$$I_{norm} = \frac{I}{255} \tag{1}$$

Thus, each pixel intensity is transformed to a value in the range $I_{norm} \in [0,1]$. This normalization helps reduce variations in input intensity values and provides standardized inputs for subsequent feature extraction. Fig. 4 presents the preprocessing of neonatal images.

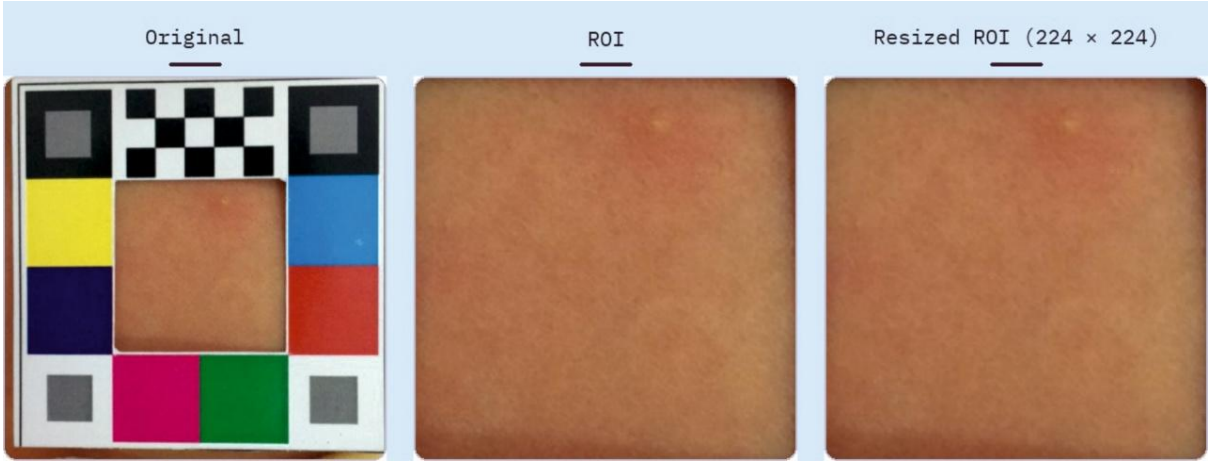


Fig. 4. Preprocessing of neonatal images through ROI extraction and resizing to 224 × 224 pixels

3.4. Swin Transformer V2-Based Feature Extraction

The Swin Transformer V2 network is leveraged as the deep feature extraction backbone for the proposed N-BEAM framework. Swin Transformer V2 is a state-of-the-art hierarchical vision transformer that improves upon the limitations of the original Swin Transformer, including scalability, training stability, and feature transferability across different resolution levels. While CNN-based networks generally learn local information through convolutional filters, the Swin Transformer V2 network models relationships among different regions of an image

through attention mechanisms. As it has a hierarchical structure that generates feature representations at different resolutions, both high-level global and low-level local features can be effectively learned.

The preprocessed neonatal skin ROIs are first split into a collection of small non-overlapping patches and subsequently embedded to represent the patches with feature representations before being passed into a series of hierarchical Swin Transformer V2 stages. Multi-head attention operates locally at each stage within individual windows, where the windows are shifted between consecutive transformer layers to enable interactions across larger areas without substantially increasing computational complexity compared with self-attention operations performed across all tokens. Features at different resolutions generated by the hierarchical window-based structure can simultaneously learn both low-level detailed and high-level structural image properties.

Moreover, it incorporates Scaled Cosine Attention, which uses cosine similarity instead of dot-product similarity for calculating the attention scores, as expressed in Eq. (2):

$$Attn(Q, K, V) = Softmax\left(\frac{\cos(Q, K)}{\tau} + B\right)V \tag{2}$$

here Q, K, and V represents the query, key, and value feature representations, respectively, τ indicates a learnable temperature parameter, $\cos(Q, K)$ is the cosine similarity between the query and key representations, and B denotes the position bias.

The scaled cosine attention reduces the dependence of feature similarity on the degree of feature vectors. In addition to the above technique, a residual post-normalization method and a log-spaced continuous position bias, as introduced in Swin Transformer V2, can further enhance optimization stability and facilitate effective feature propagation across different resolution stages [26]. These architectural designs render the Swin Transformer V2 highly applicable to neonatal image-based bilirubin estimation. Issues arising from local differences in bilirubin-associated coloration and appearance within neonatal skin cannot always be resolved based on a relatively small skin region alone. The appearance of neonatal skin tends to consist of multiple interrelated variations, requiring an effective learning process that can model local spatial appearance and capture relationships between different skin regions. Fig. 5 demonstrates the hierarchical feature extraction stages of the Swin Transformer V2 used in the N-BEAM framework.

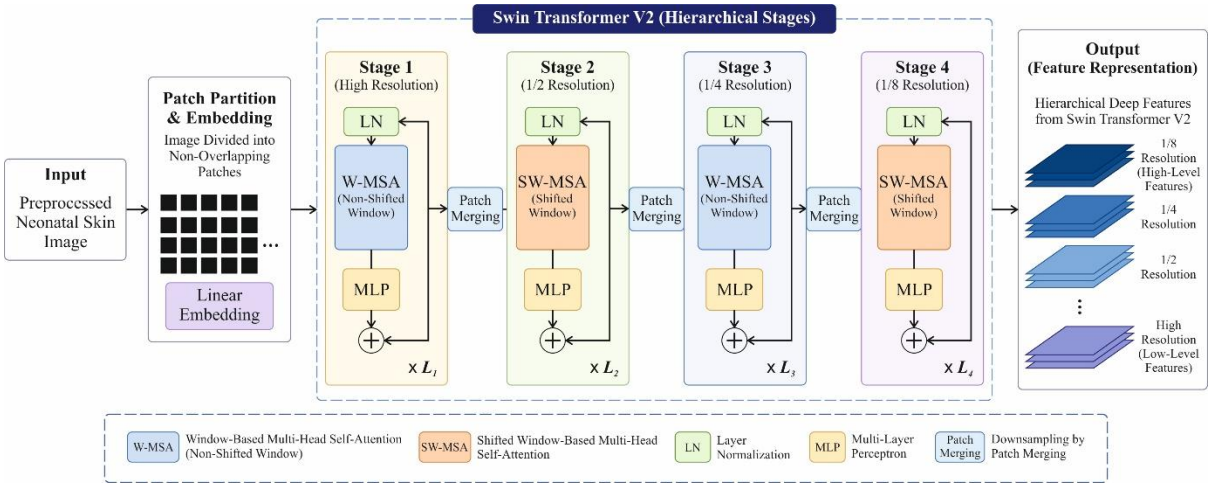


Fig. 5. Architecture of the Swin Transformer V2 used in N-BEAM, illustrating hierarchical feature extraction through patch partitioning, window-based self-attention, shifted-window attention, and patch merging

The hierarchical, shifted window-based self-attention capability of Swin Transformer V2 enables effective multi-scale feature extraction at a reasonable computational cost. In addition, the improved training stability is beneficial for medical image applications that typically involve relatively limited datasets and variations in image quality and characteristics. Recent studies have demonstrated the utility of vision-transformer-based networks for feature learning from neonatal jaundice images, which further supports the use of transformer architectures for this application [27]. The normalized neonatal skin ROI is processed through the Swin Transformer V2 method to yield a set of hierarchical deep features, which are subsequently passed to the regression module. The overall feature representation can be expressed as:

$$F_{swin} = SwinV2(I_{ROI}) \tag{3}$$

where I_{ROI} represents the preprocessed neonatal skin ROI and F_{swin} represents the hierarchical deep feature representation obtained from Swin Transformer V2. GAP is then applied to the final spatial feature representation obtained from Swin Transformer V2 to aggregate the learned features, which are subsequently fed into the FC regression layers for continuous TSB prediction.

3.5. Bilirubin Regression and TSB-Based Assessment

The hierarchical features generated by Swin Transformer V2 are then used for the prediction of continuous TSB. Different from directly estimating and classifying images into predefined jaundice classes, the proposed N-BEAM model formulates bilirubin prediction as a continuous regression task and converts the predicted continuous bilirubin value into an assessment category [28, 29]. The extracted features are then aggregated into a single feature vector using GAP so that the spatial feature representations from the final feature layer are aggregated into a compact representation.

The pooled vector is then passed through a dropout layer with a rate of 0.30 and fed into the first FC layer with 256 neurons and ReLU activation.

A second dropout layer with a rate of 0.20 is then added to minimize the risk of overfitting, followed by a second FC layer with 64 neurons and ReLU activation. Finally, a one-neuron FC layer with linear activation is used to regress the continuous TSB value (mg/dL), which is expressed as:

$$\hat{y} = f_{reg}\big(GAP(F_{Swin})\big) \tag{4}$$

where F_{Swin} is the hierarchical feature representation obtained from Swin Transformer V2, $f_{reg}(\cdot)$is the FC regression head, and y represents the predicted TSB value in mg/dL.

The MSE loss is utilized for training the regression network. The loss is computed as the mean squared difference among the predicted bilirubin concentration and the ground-truth bilirubin concentration, as follows:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{5}$$

where N signifies the total samples, y_i is the ground-truth TSB value, and $\hat{y}_i$ is the predicted TSB value for the i-th sample, respectively.

By formulating bilirubin estimation as a regression task, the proposed framework preserves the continuous nature of bilirubin measurements and does not restrict the predicted value to a single predefined jaundice category.

After the prediction of continuous bilirubin, it is transformed into the four predefined assessment categories without a dedicated classification network. The Normal category is assigned to any predicted bilirubin value < 5 mg/dL, Mild to 5 ≤ predicted bilirubin < 10 mg/dL,

Moderate to 10 ≤ predicted bilirubin < 15 mg/dL, and Severe to predicted bilirubin ≥ 15 mg/dL. The categorization function is expressed as:

$$C(\hat{y}) = \begin{cases} Normal, & \hat{y} < 5 \\ Mild, & 5 \leq \hat{y} < 10 \\ Moderate, & 10 \leq \hat{y} < 15 \\ Severe, & \hat{y} \geq 15 \end{cases} \quad (6)$$

where y is the predicted TSB value in mg/dL and $C(y)$ is the corresponding assessment category. Therefore, a two-level output is produced by N-BEAM: continuous TSB estimates in mg/dL and the corresponding discrete assessment categories derived from the predicted TSB values.

3.6. Grad-CAM++-Based Explainability

To enhance the interpretability of the N-BEAM method, Grad-CAM++ is employed as a post-hoc explainability technique [30]. Unlike the conventional Grad-CAM method, Grad-CAM++ generates more fine-grained spatial explanations by exploiting higher-order gradient information. In the proposed N-BEAM framework, Grad-CAM++ explanation is applied to identify the image regions of neonatal skin that contribute to the prediction of continuous TSB. A distinction is made between classification-based applications, where Grad-CAM++ typically explains the prediction of a particular class, and the proposed framework, where the continuous regression output produced by the bilirubin prediction head is used as the target of explanation. Let Y denote the predicted bilirubin concentration, and A^k represent the k-th spatial feature representation captured by the Swin Transformer V2 backbone. The Grad-CAM++ localization map for the predicted TSB value Y is expressed as:

$$L^Y_{Grad-CAM++} = ReLU\left(\sum_k \alpha_k^Y A^k\right) \quad (7)$$

where $L^Y_{Grad-CAM++}$ represents the heatmap corresponding to the predicted TSB value Y, A^k represents the k-th spatial feature representation, and α_k^Y represents the importance coefficient associated with the k-th feature representation with respect to Y. The importance coefficient can be expressed as:

$$\alpha_k^Y = \frac{\frac{\partial Y}{\partial A^k}}{\sum_{i,j} \frac{\partial Y}{\partial A_{ij}^k} + \epsilon} \quad (8)$$

where A_{ij}^{k} represents the activation value of the k-th spatial feature representation at the spatial position (i, y), and $\in$ represents a small constant introduced to ensure numerical stability during computation. Using these weights, the feature representations are aggregated across the channels to generate the explanatory heatmap through a weighted summation followed by a ReLU operation, where regions with higher activation indicate image regions with a stronger positive contribution to the prediction, while regions with lower activation indicate image regions with a relatively weaker contribution to the prediction.

In the N-BEAM framework, the Grad-CAM++ visualization is superimposed onto the neonatal ROI corresponding to the input image used for prediction. The region of images considered for yielding the bilirubin prediction can also be highlighted. Brighter image regions contribute more toward the prediction compared to lower-intensity regions in darker image portions. Through visual inspection, it is possible to assess whether the network has learned relevant neonatal skin regions and avoided focusing on unrelated elements such as artifacts or background structures.

Because a continuous medical-related measurement is generated from an image in the proposed bilirubin estimation framework, understanding the regions contributing to the prediction is particularly important. The generated explanations therefore help assess whether the learned representations are associated with relevant neonatal skin regions and consequently enhance the interpretability of the proposed model. However, Grad-CAM++ maps should not be considered definitive medical evidence, as they represent explanations of the model's behavior rather than independent clinical evidence.

4. Results and Discussion

The proposed N-BEAM method is implemented using Python with the PyTorch deep learning library. Experiments are carried out on a Windows 11 workstation equipped with an Intel Core Ultra 7 255HX processor, 32 GB RAM, and an NVIDIA GeForce RTX 5060 GPU with 8 GB VRAM. The performance of the proposed framework is assessed using MSE, RMSE, MAE, and R^2 for continuous TSB estimation, while accuracy, precision, recall, F1-score, and MCC are used for TSB-based neonatal jaundice severity assessment.

4.1. Cross-Dataset Experimental Evaluation

The presented N-BEAM system was assessed using a cross-dataset experimental setup to analyze its performance and generalizability across independent neonatal image datasets. The NeoJaundice dataset was used as the training dataset, and the NJN dataset was used as an independent testing dataset. Hence, during the training process of the proposed framework, no images from the test set were considered, and its ability to generalize to an independent neonatal image collection with different image characteristics and image acquisition conditions was evaluated. All the experimental assessments include continuous TSB estimation, TSB-based neonatal jaundice severity assessment, comparison with baseline methods, and visual explainability of the results through Grad-CAM++, and the corresponding results are presented below.

4.2. Continuous TSB Estimation Performance

Table 1 presents the continuous TSB estimation performance of the N-BEAM approach with the training and independent testing datasets. The proposed N-BEAM method obtained R^2 scores of 0.7496 and 0.7562 for the training and independent testing datasets, respectively, indicating its consistent performance in predicting TSB values.

Table 1 Continuous TSB estimation outcomes of the N-BEAM method under cross-dataset evaluation

split	MSE	RMSE	MAE	R2 Score
Training Set	0.0103	0.1016	0.0818	0.7496
Testing Set	0.0099	0.0997	0.0821	0.7562

4.3. Prediction Behavior and Training Convergence

Fig. 6. illustrates the actual and predicted TSB values of the N-BEAM framework for the training and independent testing datasets. The predicted values are consistent with the actual values in these datasets, representing the model's capability to capture the relevant data patterns. Even though a few deviations are perceived, especially at specific peaks and lower values, the overall prediction pattern remains reliable. It indicates that the N-BEAM model offers consistent TSB estimation performance on the training and independent testing datasets.

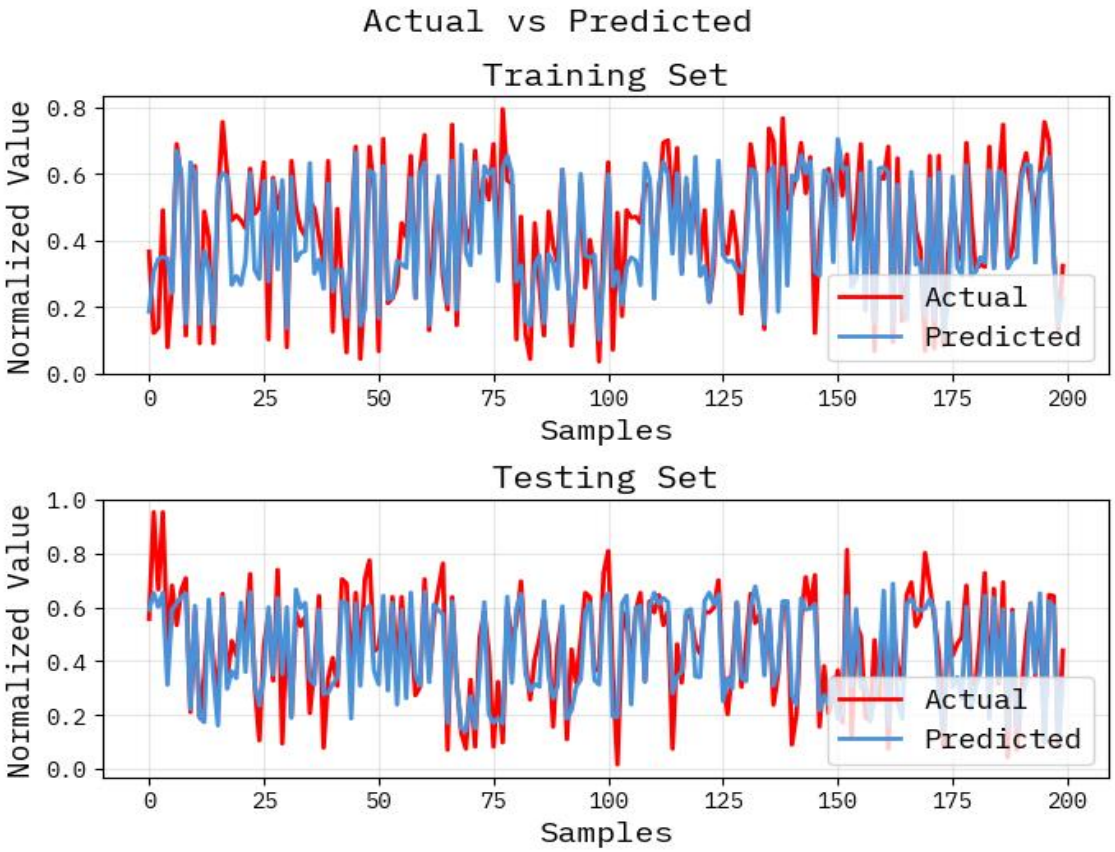


Fig. 6. Actual vs predicted TSB values of the N-BEAM method under cross-dataset evaluation

Fig. 7 illustrates the training and independent testing loss curves of the N-BEAM approach with 25 epochs using MSE, RMSE, MAE, and R^2 metrics. The MSE, RMSE, and MAE values decrease rapidly in the early epochs and progressively stabilize, demonstrating stable model convergence and learning performance. In contrast, the R^2 score rises abruptly from negative values and remains at almost 0.75, indicating improved prediction performance. The relatively close alignment between the training and independent testing curves denotes stable learning and suggests limited performance degradation on the independent dataset.

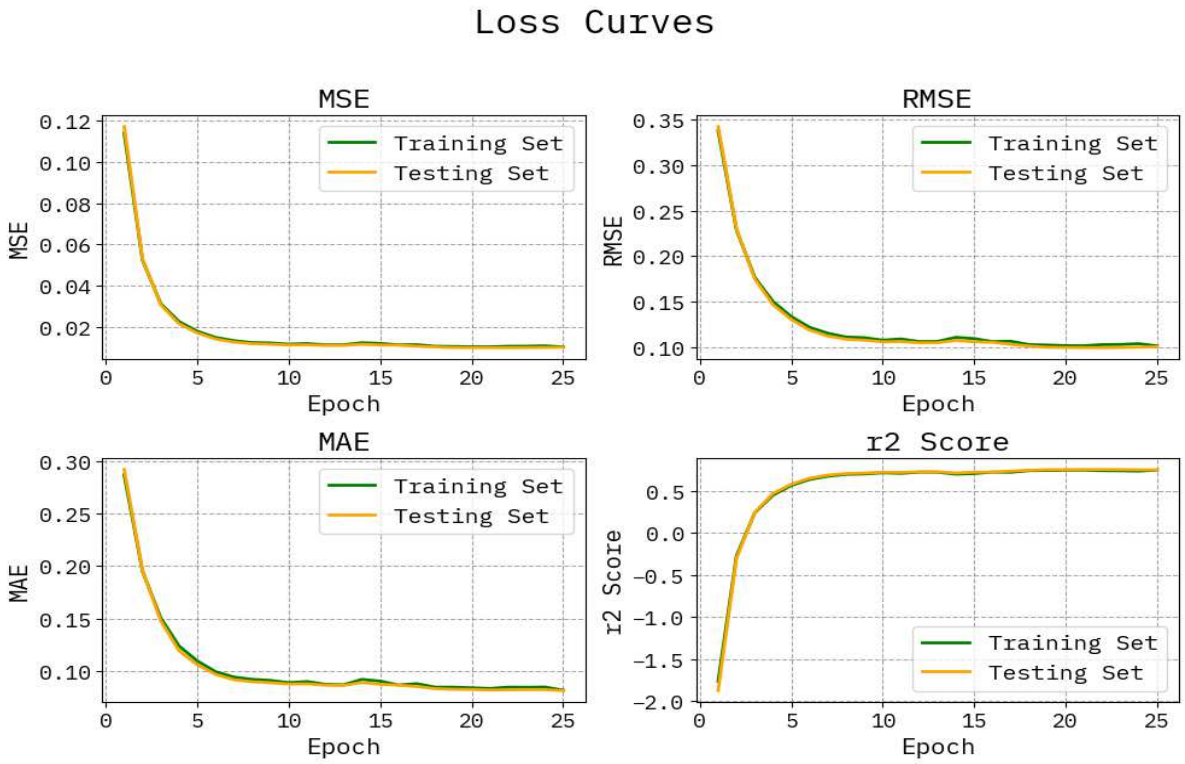


Fig. 7. Training and independent testing curves of the N-BEAM method with various metrics

4.4. TSB-Based Neonatal Jaundice Severity Assessment

Fig. 8 exhibits the confusion matrices of the N-BEAM approach with the training and testing datasets. Fig. 8a presents the training matrix, where 697 moderate, 616 severe, 478 mild, and 323 normal samples are accurately classified. Also, Fig. 8b shows the testing matrix, where 62 moderate, 42 severe, 58 mild, and 532 normal samples are correctly identified, demonstrating the robustness and effectiveness of the N-BEAM method for TSB-based neonatal jaundice severity assessment.

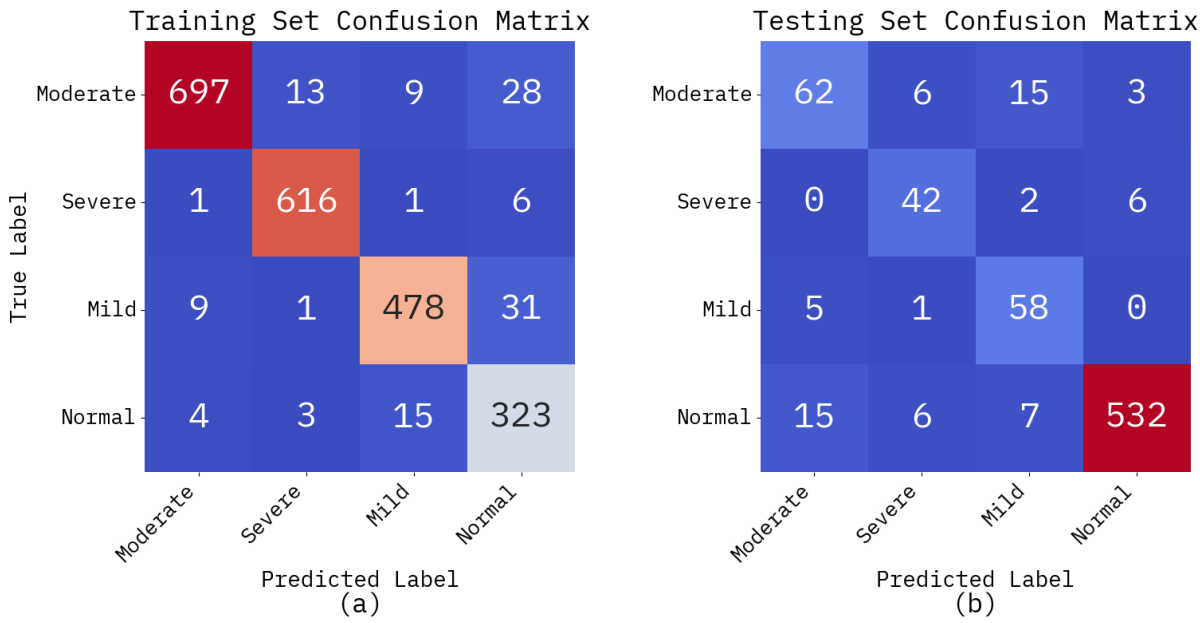


Fig. 8. Confusion matrices of the N-BEAM method with (a-b) training and independent testing datasets

Table 2 shows the neonatal jaundice severity assessment outcomes of the N-BEAM method with the training and independent testing datasets. The proposed N-BEAM method achieves average accuracies of 97.29% and 95.66% along with standard deviations of ±1.00 and ±1.12 for the training and independent testing datasets, representing the predictive capability and robustness of the presented jaundice severity assessment process.

Table 2 Neonatal jaundice severity assessment results of the N-BEAM approach under cross-dataset evaluation

Class	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**MCC**
Training Set					
Moderate	97.14	98.03	93.31	95.61	93.55
Severe	98.88	97.31	98.72	98.01	97.24
Mild	97.05	95.03	92.10	93.54	91.65
Normal	96.11	83.25	93.62	88.13	86.02
Average	**97.29**	**93.41**	**94.44**	**93.82**	**92.12**
Standard Deviation	**±1.00**	**±5.97**	**±2.54**	**±3.65**	**±4.05**
Testing Set					
Moderate	94.21	75.61	72.09	73.81	70.58
Severe	97.24	76.36	84.00	80.00	78.62
Mild	96.05	70.73	90.62	79.45	78.03
Normal	95.13	98.34	95.00	96.64	87.99
Average	**95.66**	**80.26**	**85.43**	**82.48**	**78.81**
Standard Deviation	**±1.12**	**±10.66**	**±8.64**	**±8.53**	**±6.18**

4.5. Comparative Analysis with Existing Methods

Table 3 presents the comparative results of the N-BEAM system with existing methodologies using the reported evaluation metrics [31, 32]. The outcome indicated that the 1DCNN model obtained lower performance compared with standard approaches such as LightGBM, RF, and SVR methods. In addition, the DT and KNN systems performed comparatively closer results. However, the proposed N-BEAM method achieved the highest R^2 score of 0.7562, indicating improved predictive ability and explanatory power of the proposed neonatal jaundice assessment system.

Table 3 Comparison outcomes of the N-BEAM approach with other methods

Methods	**MSE**	**RMSE**	**MAE**	**R2-Score**
DT [31]	9.6350	3.1040	2.4010	0.7340
KNN [31]	9.3660	3.0600	2.3560	0.7420
RF [31]	6.8160	2.6110	2.0180	0.7070
SVR [31]	7.0540	2.6560	2.0740	0.6980
LightGBM [31]	6.9460	2.6360	2.0410	0.7280
1DCNN [32]	1.2770	1.1300	0.8730	0.6540
N-BEAM	**0.0099**	**0.0997**	**0.0821**	**0.7562**

4.6. Visual Explainability Analysis Using Grad-CAM++

Fig. 9 reveals the Grad-CAM++ visualization of the original ROI images. The heatmaps show the areas contributing strongly to the model prediction, with yellow or red indicating higher activation and blue or purple denoting lower activation. The overlay images denote how these distinctive areas correspond to the ROI structures and textures. This visualization offers an explainable view of the image areas influencing the TSB prediction.

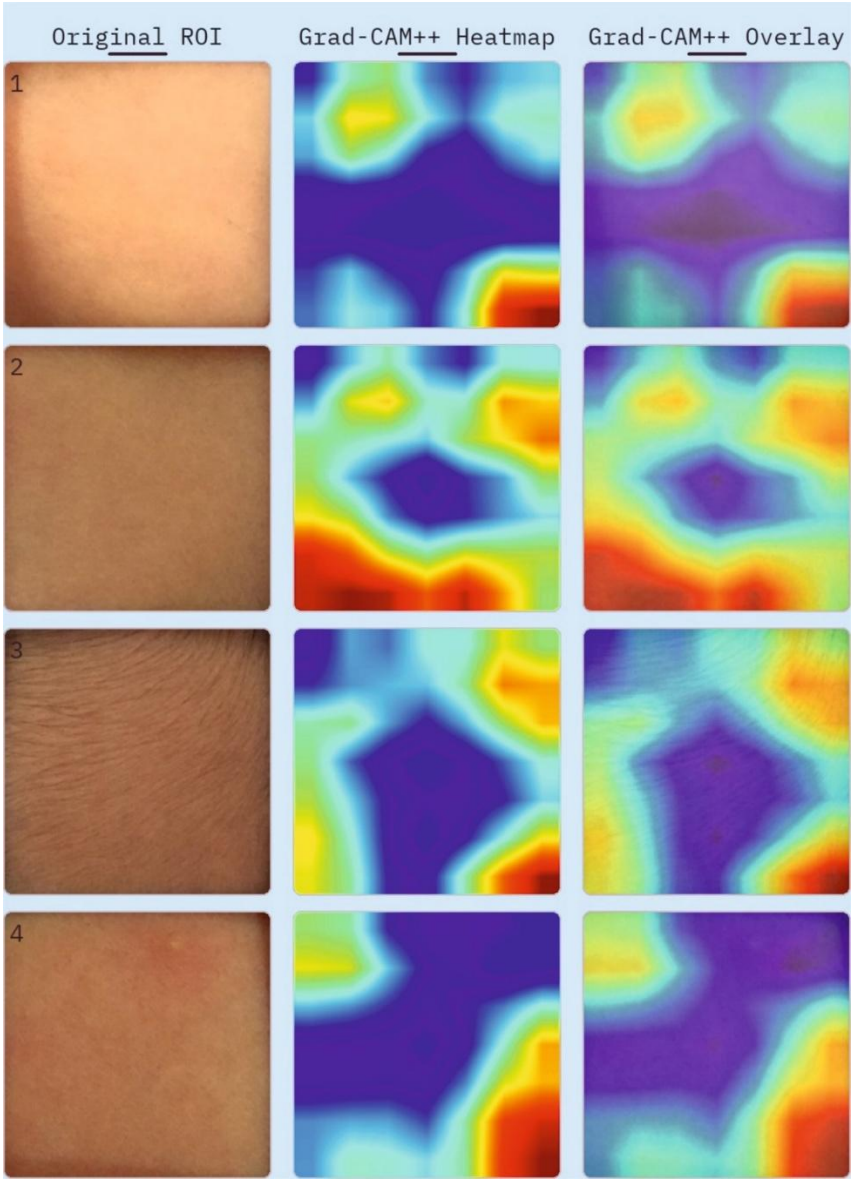


Fig. 9. Grad-CAM++ visualization of the original ROI images

4.7. Discussion

Overall, the proposed N-BEAM system achieved promising performance for non-invasive TSB estimation and TSB-based assessment of neonatal jaundice from images. Unlike traditional methods that primarily rely on binary or multi-class classification of jaundice severity, the proposed framework uses regression to directly estimate the continuous TSB value and subsequently maps the prediction into different assessment levels. This preserves the quantitative nature of bilirubin estimation while also providing an interpretable severity assessment. With the use of Swin Transformer V2, the N-BEAM framework effectively incorporates both local skin characteristics and global contextual information for bilirubin estimation. The cross-dataset evaluation using the NeoJaundice and NJN datasets, with one

dataset utilized for training and the other used as an independent testing dataset, further enables evaluation of the framework's generalization capability across heterogeneous neonatal image datasets. The comparison with other models and the ablation study demonstrate the contribution of image preprocessing and transformer-based feature extraction to bilirubin estimation. Meanwhile, the post-hoc visualizations using Grad-CAM++ provide an interpretation mechanism in which the contributing regions are highlighted in the images used for TSB estimation.

Though it is not appropriate to consider the post-hoc visualization (using Grad-CAM++) as direct clinical evidence, the display still helps interpret the learning process of the model and can clearly portray the visual cues that the model identified. However, collectively, these methods serve as an effective platform to determine the bilirubin level as well as assess the overall neonatal jaundice status of a newborn without the need for a blood test for each prediction. However, larger and more diverse multi-center datasets are still required for more comprehensive evaluation of the model across variations in skin pigmentation, illumination conditions, imaging devices, and clinical environments. Furthermore, the proposed image-based TSB estimation should be considered a supportive non-invasive assessment approach rather than a replacement for established clinical bilirubin measurements until further clinical validation is performed.

5. Conclusion and Future Directions

The importance of rapid and regular monitoring of bilirubin levels in newborns is associated with the prevention of potentially life-threatening complications. Thus, an image-based AI framework capable of estimating bilirubin levels from neonatal images can provide a useful non-invasive option for jaundice assessment. This work proposes N-BEAM as a framework for neonatal jaundice assessment and TSB estimation. The N-BEAM framework extracts the neonatal skin region, preprocesses the image, and uses Swin Transformer V2 for deep feature learning from visual representations relevant to bilirubin levels through hierarchical feature extraction. These learned features are directly utilized for continuous bilirubin regression to retain the quantitative nature of the measurement rather than using fixed disease-category-based classification. The predicted TSB value is then mapped to predefined severity levels. Additionally, a Grad-CAM++ component is incorporated to reveal the regions used by the model for TSB estimation, thereby increasing the interpretability of the model and enabling

assessment of whether the model concentrates on relevant regions for predicting the TSB value rather than irrelevant portions of the images.

The evaluation on two neonatal image datasets using a cross-dataset training and independent testing protocol further indicates the feasibility of N-BEAM across different image characteristics and acquisition conditions. The promising results show that N-BEAM provides a strong alternative non-invasive image-based approach for TSB prediction and jaundice estimation. Nonetheless, for model validation and confirmation of robustness, extensive examination using larger, diverse, and multi-center datasets incorporating various skin phototypes, illumination conditions, lighting, camera models, and hospital environments is necessary. Additional investigations may include validation on other publicly available datasets, quantitative evaluation of explainability, clinical validation against laboratory-measured TSB values, and deployment of the proposed framework on portable devices or smartphones.

Declarations

Data Availability Statement: The data that support the findings of this study are openly available in Kaggle repository at https://springernature.figshare.com/articles/dataset/NeoJaundice_Neonatal_Jaundice_Evaluation_in_Demographic_Images/22302559?file=39672982 and https://zenodo.org/records/7825810, reference number [22, 23].

Author Contribution: "Conceptualization, methodology, software, validation, writing—original draft preparation: X, Y, Z; formal analysis, investigation, writing—review and editing, supervision, project administration: X, Y, Z. All authors have read and agreed to the published version of the manuscript."

Conflict of Interest: The authors declare no conflict of interests.

Competing interests: The authors declare no competing interests.

Ethics Approval: Not Applicable

Consent to Participate: Not applicable.

Informed Consent: Not applicable.

Clinical Trial Registration: Not applicable.

Funding Information: No funding is provided for the preparation of the manuscript.

References

[1] Ma, H., Chen, X., Zhang, P., Dong, X., Wu, B., Chen, G., Zhou, W. and Wang, M., 2026. Machine learning-based genome-wide association analysis to construct a clinical decision model for severe neonatal jaundice. Translational Pediatrics, 15(5), p.181.

[2] Achaliya, S., Agarwal, S. and Ghate, S., 2026. Neonatal jaundice detection: A comprehensive review of traditional, machine learning, deep learning and emerging approaches. Intelligence-Based Medicine, 15, p.100392.

[3] Gupta, S. and Kovuri, K., 2026. Artificial Intelligence-Based Approaches for the Detection of Neonatal Jaundice. Indian J Mod Res Rev, 4(5), pp.291-297.

[4] Poshitha, M. and Pandi, J.M., 2026, April. Optimized Machine Learning Approach for Neonatal Jaundice Detection Using HOG-SVM Classification. In 2026 International Conference on Recent Advancement in Electrical, Computer and Communication Technologies (IECCT) (pp. 1-6). IEEE.

[5] Miranda, L. and Ticona, W., 2026. Model for Detecting Liver Disease Based on Jaundice in Newborns Using Convolutional. Focus on Artificial Intelligence in Intelligent Systems Design: Proceedings of 9th Computational Methods in Systems and Software 2025, Volume 3, 3, p.257.

[6] Shen, G.T., Liu, Y.H., Lin, J.Y., Hsu, K.C. and Chou, H.H., 2025. Development and validation of neonatal jaundice detection using smartphone application. Computers in Biology and Medicine, 196, p.110845.

[7] Liang, Y., Zou, L., Goh, M.M.R., Ngeow, A.J.H., Tan, N.C., Ta, A.W.A. and Goh, H.L., 2025. Multimodal feature analysis for automated neonatal jaundice assessment using machine learning. JAMIA open, 8(6), p.ooaf165.

[8] Oni, F., Olamide, O. and Daramola, N., 2026, April. LumiCare: A Low-Cost Machine Learning–Based Optical Imaging Framework Integrating Skin and Sclera Features for Non-Invasive Neonatal Hyperbilirubinemia Detection. In Optical Tomography and Spectroscopy (pp. JM4A-31). Optica Publishing Group.

[9] Xiong, Y., Jia, C., Yao, Y., Xue, T., Zhang, Y. and Xiang, L., 2026. Identifying payment-driven patient selection using machine-learning-predicted cost risk: evidence from neonatal care in China. BMC Medicine.

[10] P., Raghuwanshi, M. and Bagde, A., 2025. Noninvasive way for the detection of neonatal jaundice using GBR. Chulalongkorn Medical Journal, 69(5).

[11] Mangayarkarasi, T., Abitha, V.S., Jayasree, R., Kalyani, V. and Varshini, S.V., 2025, April. Machine Learning Based Jaundice Detection for Neonates. In 2025 International Conference on Computing and Communication Technologies (ICCCT) (pp. 1-7). IEEE.

[12] Karthikeyan, P., Al Mohamed Bilal, A., Raja Muthaiah, C., Shivani Suvatheka, S., Karthikeyan, B. and Vinoth, S., 2025, March. Determining Jaundice Severity in Neonates using Optical Spectroscopy. In 2025 5th International Conference on Expert Clouds and Applications (ICOECA) (pp. 941-945). IEEE.

[13] Kumari, K., Pahuja, S.K. and Kumar, S., 2025. Neonatal Jaundice Detection and Classification Using Deep Convolutional Neural Network. In Research on Engineering Applications in Multidisciplinary Sectors (Part 2) (pp. 85-93). Bentham Science Publishers.

[14] Lotfi, M., Rabiee, M., Nazarpak, M.H., Sangesari, R., Alishahi, N. and Azimi, M.S., 2026. Neonatal jaundice detection using a vision transformer-based deep learning model. Scientific Reports, 16(1), p.9243.

[15] Akcan, A., Hanci, N.B., Erdem, O. and Duran, R., 2026, July. Detection of Neonatal Jaundice with Hybrid Models of ViT and Machine Learning. In 2026 34th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[16] Patel, R. and Jarpula, N., 2026. NeoJaundice-AI: Smartphone-Based Neonatal Jaundice Detection Using Dual-Input Deep Learning and Synthetic Augmentation. arXiv preprint arXiv:2606.20689.

[17] Gogulakrishnan, R., Rameshvar, A. and Kannan, R.G., 2026, April. Comparitive Analysis of Deep Learning Architectures for Automated Neonatal Jaundice Detetction. In 2026 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI) (pp. 1-7). IEEE.

[18] Akarapanuvitaya, N. and Pintavirooj, C., 2025, July. Real Time Diagnosis of Neonatal Jaundice Using Machine Learning. In 2025 17th Biomedical Engineering International Conference (BMEiCON) (pp. 1-5). IEEE.

[19] Krishnaraj, S., Subramaniam, S. and Krishnamoorthy, U., 2026. Class-Imbalance Aware Learning for Enhanced Noninvasive Neonatal Jaundice Diagnosis–Using Convnext-Tiny with Focal Loss (CNXT-FL). International Journal of Industrial Engineering: Theory, Applications and Practice, 33(4).

[20] Premalatha, P., Deepa, T.B. and Madhumithaa, S., 2026, March. A Lightweight Deep Learning Model for Non-Invasive Bilirubin Grading by Multi-Colour Space Images and

Metadata Fusion. In 2026 Twelfth International Conference on Bio Signals, Images, and Instrumentation (ICBSII) (pp. 1-6). IEEE.

[21] Salami, F.O., Mourchid, Y., Muzammel, M. and Othmani, A., 2025, October. Non-Invasive Neonatal Jaundice Detection via Two-Phase Self-Supervised Learning and Vision Transformer. In 2025 Fourteenth International Conference on Image Processing, Theory, Tools & Applications (IPTA) (pp. 1-6). IEEE.

[22] https://springernature.figshare.com/articles/dataset/NeoJaundice_Neonatal_Jaundice_Evaluation_in_Demographic_Images/22302559?file=39672982

[23] https://zenodo.org/records/7825810

[24] Makhloughi, F., 2025. Artificial intelligence-based non-invasive bilirubin prediction for neonatal jaundice using 1D convolutional neural network. Scientific Reports, 15(1), p.11571.

[25] Lotfi, M., Rabiee, M., Nazarpak, M.H., Sangesari, R., Alishahi, N. and Azimi, M.S., 2026. Neonatal jaundice detection using a vision transformer-based deep learning model. Scientific Reports, 16(1), p.9243.

[26] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L. and Wei, F., 2022, June. Swin transformer v2: Scaling up capacity and resolution. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 11999-12009). IEEE.

[27] Lotfi, M., Rabiee, M., Nazarpak, M.H., Sangesari, R., Alishahi, N. and Azimi, M.S., 2026. Neonatal jaundice detection using a vision transformer-based deep learning model. Scientific Reports, 16(1), p.9243.

[28] Liang, Y., Zou, L., Goh, M.M.R., Ngeow, A.J.H., Tan, N.C., Ta, A.W.A. and Goh, H.L., 2025. Multimodal feature analysis for automated neonatal jaundice assessment using machine learning. JAMIA open, 8(6), p.ooaf165.

[29] Phattraprayoon, N., Ungtrakul, T., Kummanee, P., Tavaen, S., Pirunnet, T. and Fuangrod, T., 2025. Feasibility study of texture-based machine learning approach for early detection of neonatal jaundice. Scientific Reports, 15(1), p.6481.

[30] Soomro, S., Niaz, A. and Choi, K.N., 2024. Grad++ ScoreCAM: enhancing visual explanations of deep convolutional networks using incremented gradient and score-weighted methods. IEEE Access, 12, pp.61104-61112.

[31] Yuan, Y., Huang, J., Yu, J., Tan, J.K.S., Chng, K.Z., Lee, J. and Kim, S., 2024. Application of machine learning algorithms for accurate determination of bilirubin level on in vitro engineered tissue phantom images. *Scientific Reports*, *14*(1), p.5952.

[32] Makhloughi, F., 2025. Artificial intelligence-based non-invasive bilirubin prediction for neonatal jaundice using 1D convolutional neural network. *Scientific Reports*, *15*(1), p.11571.