Vision Mamba-Based Dual-phase Self-supervised Framework for Neonatal Jaundice Diagnosis

Fati Oiza Salami [1],*, Youssef Mourchid [2], Muhammad Muzammel [1], and Alice Othmani [1]

[1] Laboratoire Images, Signaux et Systémes Intelligents (LiSSi) EA 3956, Université Paris Est Créteil (UPEC),
Vitry sur Seine, France
[2] CESI LINEACT Laboratory, UR 7527, Dijon, France
Email: fati.salami@u-pec.fr (F.O.S.); ymourchid@cesi.fr (Y.M.); muhammad.muzammel@u-pec.fr (M.M.);
alice.othmani@u-pec.fr (A.O.)
*Corresponding author

***Abstract*—Neonatal jaundice is a common and potentially serious condition that, if left undiagnosed or untreated, can lead to severe neurological complications in newborns. Existing diagnostic methods are often invasive and face limitations in accuracy, accessibility, and data availability, especially in resource-constrained environments. This study introduces NeoViM, an adapted MambaVision-based framework for neonatal jaundice classification. The framework adopts a two-stage approach: In the first stage, the adapted MambaVision is used as a deep feature extractor. Self-Relative Clustering (SRC) is then applied to learn discriminative features by organizing images into clinically meaningful clusters, enabling effective unsupervised learning from limited data. To further enhance feature quality, a self-supervised learning strategy based on Linear Kernel Centered Alignment (LCKA) loss is employed to refine the extracted representations. In the second stage, the pre-trained model is fine-tuned on a small labeled dataset, allowing the system to adapt the learned features for accurate jaundice classification. The methodology was evaluated using the publicly available NJN dataset and achieved a classification accuracy of 93.42% and an F1 score of 93.37%, outperforming previous methods applied to the same dataset. This two-step framework ensures high diagnostic performance while maintaining scalability and accessibility across diverse clinical settings.**

***Keywords*—neonatal jaundice diagnosis, non-invasive, Mamba Vision, deep learning, self-supervised learning**

I. Introduction

Neonatal jaundice, also known as hyperbilirubinemia, is a prevalent medical condition affecting newborns worldwide, characterized by elevated bilirubin levels in the bloodstream [1, 2]. Although mild cases often resolve without intervention, severe jaundice can cause irreversible neurological damage and even death if not promptly diagnosed and managed. Several factors are responsible for causing jaundice in neonates, the most common of which are poor performance of the neonate's immature liver, excessive breakdown of red blood cells, and bilirubin build-up [3]. Early diagnosis of neonatal jaundice is critical to prevent severe complications such as kernicterus and other bilirubin-induced neurological damage. However, the gold standard for neonatal jaundice diagnosis currently is Total Serum Bilirubin (TSB) measurement; this method requires invasive blood sampling and laboratory analysis by skilled phlebotomy within a laboratory infrastructure [1]. These steps cause physical discomfort or pain to the neonate as well as introduce delays, particularly in resource-limited settings. While Transcutaneous Bilirubinometers (TcB) offer a noninvasive alternative, their accuracy in jaundice detection varies with skin pigmentation, the gestational age of the baby, and device calibration. Another alternative is visual assessment by clinicians [1, 4], though this approach is widely used, it is highly subjective and prone to inter-observer variability, thereby resulting in incorrect or delayed diagnoses in under-equipped healthcare facilities.

These diagnostic challenges have driven the development of cost-effective, non-invasive diagnostic alternatives to mitigate the risk of bilirubin-induced complications. Recently, advances in Machine Learning (ML) and Deep Learning (DL) have introduced novel, non-invasive screening methods, such as image-based bilirubin prediction using advanced AI-based techniques, smartphone cameras, and an automated risk stratification model trained on clinical datasets [5–7]. These AI-driven techniques, while demonstrating high diagnostic accuracy and reducing dependency on specialized personnel and tools, are also constrained by some challenges, such as a relatively small training dataset, which can affect the generalizability of the model, illumination, and color variability that could cause inconsistent predictions. Hence, limiting its ability to leverage a large unlabeled dataset for better feature extraction. To build upon recent advances and address the persistent challenges in neonatal jaundice detection, this study introduces NeoViM, a non-invasive, image-based diagnostic framework that harnesses self-supervised learning on a MambaVision network to enable fast, accurate identification of neonatal jaundice. The goal

Manuscript received July 16, 2025; revised July 24, 2025; accepted August 28, 2025; published December 19, 2025.

doi: 10.18178/joig.13.6.702-708

is to facilitate timely clinical intervention while reducing the risks associated with delayed or invasive screening procedures. Specifically, this study proposes:

- The adoption of Mamba Vision as a computationally efficient and high-performance backbone network for neonatal jaundice detection.
- The integration of Self-Relative Clustering (SRC) for unsupervised extraction of clinically relevant features, revealing intrinsic spatial structures within unlabeled medical images.
- The application of a Self-Supervised Learning (SSL) strategy, employing Linear Kernel Centered Alignment (LCKA) loss, to refine feature representations and enhance diagnostic accuracy.

The remainder of this study is organized into the following sections; Section II presents related works, Section III details the study methodology, Section IV discusses the results of the study, and Section V details the conclusion.

II. Related Work

Recent advancements in medical image analysis have brought about significant development of various classification frameworks for neonatal jaundice detection. This section presents some recent relevant works categorized into machine learning, deep learning, and self-supervised learning methodologies.

Several studies have adopted the traditional machine learning techniques in the detection of neonatal jaundice. Abdulrazzak *et al.* [8] proposed a real-time neonatal jaundice detection using Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), Random Forest (RF), and Extreme Gradient Boost (XGBoost) on the Normal and Jaundiced (NJN) dataset with 760 images, provided by Abdulrazzak *et al.* [9]. XGBoost outperformed SVM, k-NN, and RF in terms of both classification accuracy and robustness, achieving remarkable performance. Though this study showed great performance, it did not employ cross-validation, and traditional ML techniques require extensive data preprocessing, which results in increased memory and computational requirements. Similarly, Abdulrazzak *et al.* [10] deployed a computer vision-based system by analyzing the skin color variations using a Random Forest classifier trained on 511 neonates' images. The approach demonstrated high performance, however, the reliance of the approach on manual Region-of-Interest (ROI) selection and feature extraction is time-consuming and introduces subjectivity.

Deep learning techniques have also been adopted to automate feature extraction and enhance classification performance in neonatal jaundice detection. In the study by Gupta *et al.* [7], a non-invasive diagnostic framework was presented. It implemented custom Convolutional Neural Network (CNN) models, MobileNet V3, EfficientNet, and Vision Transformer (ViT). With the experiments performed on the NJN dataset, the results showed that the ViT achieved the best performance. Despite the encouraging results achieved, the study highlights the limited dataset size as one of the challenges of model generalization. In another study, Abdulkader *et al.* [5] used deep transfer learning for classifying neonatal jaundice severity using VGG16 and ResNet50 on a dataset of 334 skin images captured from neonates, showcasing the potential of smartphone-based image analysis for early, non-invasive jaundice severity detection. ResNet achieved the highest performance, although its generalizability remains limited.

Self-Supervised Learning (SSL) is revolutionizing medical imaging by leveraging unlabeled data to learn patterns, reducing expert annotations, and enhancing the learning of domain-specific patterns [11].

Recent studies highlight the effectiveness of SSL in medical imaging; studies by Xu *et al.* [12] proposed 3DINO, a state-of-the-art SSL technique adapted to 3D image inputs, deploying the model using a multi-organ dataset containing about 100,000 unlabeled 3D medical images. The 3DINO technique highlighted impressive performance on both segmentation and classification tasks. Chattopadhyay *et al.* [13] and Ali *et al.* [14] also used frameworks such as SimCLR, DCL, VICReg, and SimSiam to demonstrate SSL's ability to extract meaningful features and enhance performance in limited-label settings.

Despite methodological diversity, common limitations encountered in previous works include limited dataset availability, demographic homogeneity, and subtle pathological features, affecting generalization. To address these challenges, this study proposes a novel dual-phase SSL framework for neonatal jaundice detection, enhancing feature extraction for small datasets by leveraging intricate pixel-level pattern representation and ensuring diagnostic accuracy, filling a gap in previous studies, and providing a clinically relevant solution for neonatal jaundice detection.

III. Methodology

A. Overview of the NeoViM Framework

This section introduces the proposed method, NeoViM (Neo-Vision Mamba), a lightweight dual-phase Self-Supervised Learning (SSL) framework for diagnosing neonatal jaundice by learning robust visual representation from a small medical dataset. NeoViM deploys a two-phase training pipeline made up of unsupervised pre-training and supervised fine-tuning. This approach strategically combines Self-Relative Clustering (SRC), an approach used for strong semantic grouping consisting of a single linear layer that maps features to cluster assignments even in low data settings during unsupervised feature extraction, and a strengthened feature alignment via Linear Centered Kernel Alignment (LCKA), a technique to improve feature alignment by quantifying the relationships between the sets of learned features extracted through SRC, to tackle data scarcity and enhance the classification task.

B. Proposed Approach

Fig. 1 represents a progressive learning pipeline of the proposed NeoViM model adopted to overcome the limitation of small labeled dataset. The training process is divided into two subphases; two models are deployed for feature vector extraction, the Self-Supervised Learning model (extracts features, F_1) and the Self-Relative Clustering (SRC) model (extracts features, F_2).

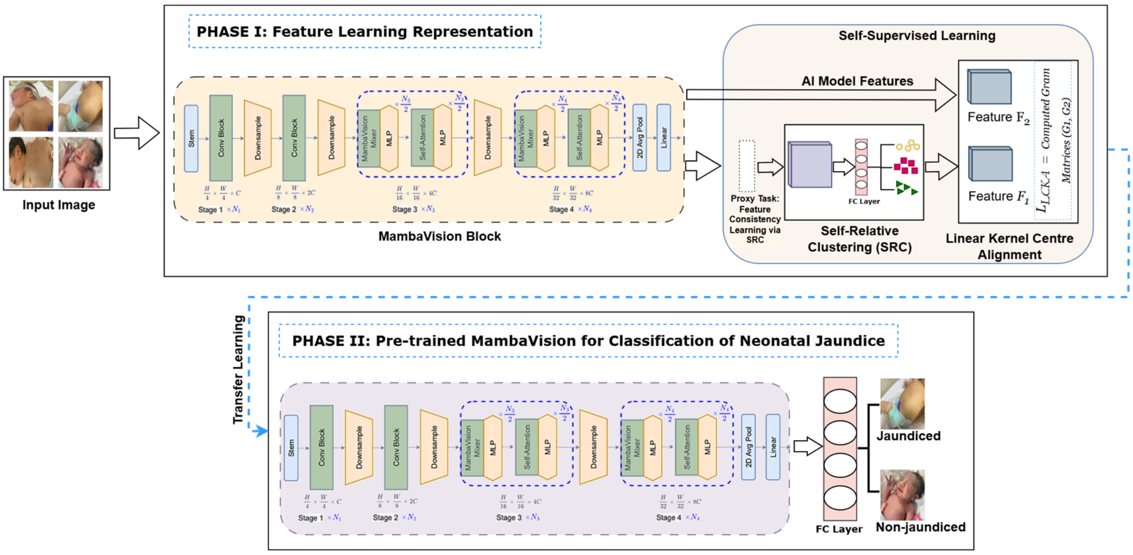
Input Image
PHASE I: Feature Learning Representation
Stem
Conv Block
Downsample
Conv Block
Downsample
MambaVision Mixer
MLP
Self-Attention
MLP
Downsample
MambaVision Mixer
MLP
Self-Attention
MLP
2D Avg Pool
Linear
$\times\frac{N_3}{2}$
$\times\frac{N_4}{2}$
$\frac{H}{4}\times\frac{W}{4}\times C$
Stage 1 $\times N_1$
$\frac{H}{8}\times\frac{W}{8}\times 2C$
Stage 2 $\times N_2$
$\frac{H}{16}\times\frac{W}{16}\times 4C$
Stage 3 $\times N_3$
$\frac{H}{32}\times\frac{W}{32}\times 8C$
Stage 4 $\times N_4$
MambaVision Block
Self-Supervised Learning
AI Model Features
Feature F_2
Feature F_1
L_{LCKA} = Computed Gram Matrices (G_1, G_2)
Proxy Task: Feature Consistency Learning via SRC
FC Layer
Self-Relative Clustering (SRC)
Linear Kernel Centre Alignment
Transfer Learning
PHASE II: Pre-trained MambaVision for Classification of Neonatal Jaundice
FC Layer
Jaundiced
Non-jaundiced

Fig. 1. Proposed dual-phase NeoViM framework. In Phase I, MambaVision used for SSL and SRC model generates feature sets F_1 and F_2 respectively, LCKA then compares the Gram matrices of F_1 and F_2 to enforce consistent representation learning. In Phase II, the pre-trained model is fine-tuned with labeled data to classify neonatal images as jaundiced or not-jaundiced.

The extracted feature vectors are transformed into Gram matrices, and the Linear Centered Kernel Alignment (LCKA) loss is used to compute the similarity between these Gram matrices and update the model accordingly. Hence, enabling the trained model to extract features that mimic the pre-trained model, making it more robust. The following provides a detailed explanation of the various components and steps of the approach used in this paper.

1) Deep feature extraction with backbone network

First, the MambaVision encoder was initialized to extract rich and discriminative deep features that capture essential visual patterns from the input images. MambaVision is a hybrid vision backbone that integrates Mamba-based State Space Models (SSMs) with Transformer self-attention layers. Using a hierarchical design, the early convolutional layers perform local feature extraction, while later hybrid layers efficiently capture global context, enabling linear-time sequence modeling for improved output [15]. The Mamba Vision component's core functionality is based on a discrete linear state-space model that is derived from the original Mamba 1D model [16] shown in Eq. (1), where a continuous input $x(t) \in \mathbb{R}$ is transformed into an output $y(t) \in \mathbb{R}$ through a learnable hidden state $h(t) \in \mathbb{R}^M$.

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t), \\ y(t) &= Ch(t) \end{aligned} \tag{1}$$

The continuous-time parameters $A \in \mathbb{R}^{M\times M}$, $B \in \mathbb{R}^{M\times 1}$ and $C \in \mathbb{R}^{1\times M}$ from Eq. (1) are discretized to obtain a computationally efficient representation, which forms the discrete linear state-space model presented in Eq. (2).

$$\begin{aligned} h(t) &= \bar{A}h(t-1) + \bar{B}x(t), \\ y(t) &= \bar{C}h(t) \end{aligned} \tag{2}$$

MambaVision's linear time complexity and modest computational footprint makes it ideal for small-scale medical imaging tasks.

2) SRC feature learning

In parallel, the unlabeled images are passed through the encoder network (a Vision Mamba backbone) to extract high-level features, where a proxy task based on SRC, introduced in the study of [17], is employed. SRC is used to group images with similar discriminative features, which may not be easily achieved through supervised learning due to data scarcity and inter-class similarity, this enhances feature consistency learning and representation quality from unlabeled data. In this study, an extensive experiment with varying cluster counts was conducted, and a 10-cluster count yielded the best performance in classification tasks. Although the original authors did not specify this cluster count, it was empirically validated in the dataset used in this study. SRC is used to group images with similar discriminative features, which may not be easily achieved through supervised learning due to data scarcity and inter-class similarity.

3) Feature alignment via LCKA loss

After the SRC encoder is pretrained and frozen, its learned representation is subsequently aligned with those from the Mamba encoder using the LCKA loss. First, the feature set F_1, extracted from the MambaVision, and the feature set F_2, extracted from the SRC encoder, are transformed into Gram matrices, G_1 and G_2 respectively, as shown in Eq. (3).

$$G_1 = F_1 \cdot F_1^T, \quad G_2 = F_2 \cdot F_2^T \tag{3}$$

It is essential to convert the feature set into Gram matrices because LCKA operates by comparing similarity information between samples rather than raw features. Gram matrices obtain these similarity relationships in a feature set by measuring how each feature vector relates to others, allowing LCKA to quantify the relationships between the two sets of learned features rather than their exact values.

The LCKA loss computes the similarity between the two Gram matrices as depicted in Eq. (4).

$$L_{\text{LCKA}} = -\frac{\sum(G_1 \cdot G_2)}{\|G_1\| \cdot \|G_2\|} \quad (4)$$

The aim of the LCKA loss function is to refine the extracted feature representations by enabling better alignments between feature distributions.

4) Fine-tuning the SSL model on labeled data

After the pretraining and alignment of the feature extractor, downstream classification was performed using the labeled subset of the dataset. The MambaVision model was fine-tuned for classification using a two-layer ReLU-activated Fully Connected (FC) head, trained on the extracted features. A two-layer FC head was used as opposed to a single layer because it enables non-linear transformation of the learned features, which enables better class separability while being shallow enough to avoid overfitting, making it suitable for limited data settings.

Mixed precision was implemented via a gradient scaler (GradScaler) to accelerate training while preserving stability. It combines 16-bit (FP 16) and 32-bit (FP 32) floating-point operations. During backpropagation, gradients of the scaled loss are computed with respect to FP16 parameters, and unscaled gradients to FP32 precision before optimizer updates. These refined self-supervised features yield an optimized model for the detection of neonatal jaundice.

IV. Result and Discussion

This section presents the datasets utilized, experimental set up, the results obtained, the comparison of the model's results and, the ablation study results.

A. Dataset

In this study, the Normal and Jaundiced Newborns (NJN) dataset curated by Abdulrazak *et al.* (2023) [9] was utilized; the dataset contains 760 RGB neonatal images (560 healthy, 200 jaundiced). The NJN dataset is the only publicly available image dataset for neonatal jaundice; therefore, it serves as a critical resource for non-invasive jaundice detection. This study used 60% of the data set for training, 20% for validation, and the final 20% was used for testing the model. To ensure robust training, standard image pre-processing techniques were applied, including resizing all images to 224×224 pixels and normalization using ImageNet statistics.

B. Experimental Setup

The NeoViM framework implemented on the NJN dataset adopts a structured learning perspective that was designed to progressively extract, refine, and classify neonatal jaundice features. Structured across two main stages: (1) deep feature extraction with the MambaVision backbone that is deployed in parallel with an unsupervised representation learning via SRC terminating with LCKA-based feature alignment, and (2) supervised fine-tuning using a lightweight classifier head. In all stages, the models were trained using the Adam optimizer and a batch size of 64. Configurations, learning rates, and epochs varied by module as detailed below:

1) Mamba vision encoder

The Mamba Vision nvidia/MambaVision-S-1K (pretrained on ImageNet-1K) was used as the backbone. It was modified by replacing the final classification layer with an identity projection. It was trained with a learning rate of 0.0001 and a cross-entropy loss function over 50 epochs.

MambaVision was selected as the backbone model due to its strength in obtaining local and global partial dependencies at relatively low computational cost when compared to the Vision Transformer (ViT) models. ViTs, though, are very efficient in global attention mechanisms; they require significant compute resources to perform optimally. Furthermore, unlike ResNeT, a convolution-based model that relies on locality and may overlook intricate pixel-level features that are critical for jaundice detection, MambaVision demonstrates efficiency in retaining detailed pixel-level information. Hence, the MambaVision architecture achieves the desired balance between model accuracy, speed, and resource usage, making it suitable for deployment in low-resource clinical settings where timely diagnosis is fundamental.

2) SRC module

The SRC module consists of a single linear layer mapping features to 10 clusters; this cluster value was selected based on empirical experiments conducted to determine the best cluster count. This module was trained independently with a learning rate of 0.0003 for 50 epochs. The LCKA loss was used to enforce feature alignment between Gram matrices derived from Mamba Vision features (F_1) and SRC features (F_2).

3) Fine-tuned classifier

This consists of a two-layer MLP appended to a frozen SSL feature extractor. The architectural setup consists of a fully connected layer with 128 units, ReLU activation, and a dropout layer of 0.3, followed by a final classification layer. Trained with a learning rate of 0.0001 and fine-tuned over 200 epochs using mixed precision (FP16/FP32) via PyTorch's Gradscaler to enable faster computation, efficiency, and stability.

4) Regularization techniques

While SSL and SRC strategies were deployed with the intent to offset challenges with data scarcity, regularization played a crucial role in mitigating overfitting in NeoViM and ensuring generalization. During both self-supervised and supervised stages, the study applied data augmentation, including horizontal flips and rotations. A dropout of 0.3 was added between the fully connected layers to reduce overreliance among neurons during training. These techniques ensure that NeoViM remains consistent and generalizes well across different medical samples.

5) Implementation and hardware

The experimental environment includes PyTorch 2.2.1, a GPU Tesla T4, and CUDA 11.8. Each component of the model pipeline trained on average for 50 minutes; the full NeoViM pipeline was completed in approximately 3.5 h.

C. Ablation Study

To validate the choice of the NeoViM model architecture, a systematic ablation study was conducted to evaluate the contribution of specific components within the architecture. To achieve this, the study explored two key aspects: the inclusion of clusters in the SRC module of the proposed NeoViM model and the adoption of the self-supervised module for feature refinement. Each experiment was designed to isolate and evaluate the impact of these components on the overall model performance. Table I presents the ablation results of the NeoViM model, where it evaluates the individual and combined contributions of the Self-Supervised Learning (SSL) and SRC modules to classification performance across four key metrics.

TABLE I. ABLATION STUDY ON THE IMPACT OF SRC CLUSTERING AND SELF-SUPERVISED MODULES ON NEOVIM MODEL PERFORMANCE

Clusters	SSL	Accuracy	Precision	Recall	F1 Score
-	-	85.53%	85.32%	85.53%	85.40%
-	√	89.57%	89.61%	88.16%	88.12%
√	-	92.11%	91.99%	92.11%	91.97%
√	√	**93.42%**	**93.35%**	**93.42%**	**93.37%**

The baseline model without the SRC clustering and SSL achieves an accuracy of 85.53% and a consistent precision (85.32%), recall (85.53%), and F1 score (85.40%), demonstrating a stable but below-optimal performance. On incorporating SSL, the accuracy improved to 89.57%. SSL assists in refining intricate features in neonatal jaundice images, allowing the model to focus more on subtle variations in the input images. Integrating the SRC clustering without SSL resulted in an accuracy increment to 92.11% and a consistent precision and recall (92.11%), emphasizing the ability of SRC to group images with similar discriminative features, which may not be easily achieved through supervised learning due to data scarcity and inter-class similarity. The combination of both SRC clustering and SSL yielded the best performance across all metrics, with accuracy and recall attaining a 93.42% score, precision of 93.35%, and an F1 score of 93.37%. This result shows that a combination of the separability strength of SRC and the deep feature refinement capabilities of SSL is essential for reliable neonatal jaundice detection.

D. Comparison with State-of-the-Art

This study's framework builds upon State-of-the-Art (SOTA) techniques by integrating SRC and SSL strategies with Vision Mamba architectures, aiming to resolve critical challenges in generalization and resource utilization efficiency. This section compares our results against that of the existing study that was conducted using deep learning techniques to classify the NJN dataset.

Table II highlights that our study gained some improvement in performance over the study undertaken by Ref. [7]. This result emphasizes the potential of the proposed NeoViM model as a solution in neonatal jaundice detection. As for the training strategy, Gupta [7] employed a data partitioning strategy that reserved 25% of the original dataset for testing, while the remaining data was augmented and split in a ratio of 75:15 for training and validation, respectively. In contrast, this study adopted a 60:20:20 split because it used a pre-trained Vision Mamba model, which requires fewer training samples for effective fine-tuning, allowing more data to be allocated to validation and testing for better generalization performance assessment. This makes direct comparison of training strategies between both studies difficult.

TABLE II. COMPARISON OF PRIOR STUDIES AND THE PROPOSED NEOVIM FRAMEWORK ON THE NJN DATASET

Approach	Accuracy	Precision	Recall	F1 score
CNN(Custom) [7]	79.00%	75.00%	85.00%	80.00%
MobileNet [7]	64.00%	74.00%	79.00%	76.00%
EfficientNet [7]	82.00%	88.00%	88.00%	88.00%
ViT (ViT-Base) [7]	83.00%	87.00%	90.00%	88.00%
NeoViM [Ours]	**93.42%**	**93.35%**	**93.42%**	**93.37%**

E. External Validation and Inference Testing of the NeoViM Model

To test the generalization of the NeoViM model, an external validation (cross-dataset testing) was carried out using a separate dataset curated by [18] consisting of face images from 45 jaundiced and 55 non-jaundiced neonates.

TABLE III. PERFORMANCE OF NEOVIM MODEL ON EXTERNAL DATASET [18]

Class	Accuracy	Precision	Recall	F1-score
Jaundiced	83.72%	0.78	0.82	0.80
Non-jaundiced	83.72%	0.84	0.85	0.84

As presented in Table III, the external validation metrics yielded a weighted average accuracy of 83.72%. Validation for the jaundiced class resulted in a precision of 0.78, Recall of 0.82, and an F1 score of 0.80. For the non-jaundiced class, slightly higher precision, recall, and an F1 score of 0.84, 0.85, and 0.84, respectively, were recorded. Deducing from the strong and balanced performance of the NeoVim model across both classes on the external dataset, this implies that the model generalizes well to real-world data. In particular, the external dataset could not be validated using the approach proposed by [7] due to the unavailability of their code. As a result, no comparative analysis was performed on the external dataset.

Inference testing was conducted to assess the practical effectiveness of the NeoViM model under real-world conditions, where the deployment data may be acquired under varying conditions such as lighting, camera resolution, and angle of capture. Another dataset obtained from [19] was used as inference data to evaluate the NeoViM model.

Table IV shows that the NeoViM model recorded a high overall accuracy of 93.00% on the inference dataset. With a recall of 1.00 and precision of 0.86, the model displayed remarkable sensitivity to jaundiced cases, indicating its fairly strong potential as a reliable jaundice diagnosis tool, while in the non-jaundiced cases, the model recorded a relatively lower recall of 0.50 and an F1 score of 0.58.

Although the inference results of this study did not match those documented by [7], it achieved a good inference score. Furthermore, this study's inference time is significantly shorter, as shown in Table V.

TABLE IV. COMPARISON OF NEOVIM MODEL AND PRIOR STUDIES ON INFERENCE DATASET

Approach	Class	Accuracy	Precision	Recall	F1-score
Ours	Jaundiced	93.00%	0.86	1.00	0.92
	Non-jaundiced		0.69	0.50	0.58
CNN [7]	Jaundiced	90.00%	1.00	0.86	0.92
	Non-jaundiced		0.75	1.00	0.86
ViT [7]	Jaundiced	100.00%	1.00	1.00	1.00
	Non-jaundiced		0.69	0.62	0.65

TABLE V. COMPUTATIONAL EFFICIENCY COMPARISON OF MODEL PARAMETERS, GFLOPS, AND INFERENCE TIME BETWEEN NOEVIM, CNN AND VIT MODELS

Model	Model Parameters	GFLOPs	Inference Time
CNN [7]	51.48 M	11.4	0.8873 ms
ViT [7]	86.24 M	16.95	1.448 ms
NeoViM (Ours)	**50.14 M**	**7.50**	**0.2336 ms**

F. Discussion

The experimental results establish the proposed NeoViM framework in neonatal jaundice detection as a robust and clinically promising framework for non-invasive neonatal jaundice detection. Trained on the NJN dataset (the only publicly available dataset on neonatal jaundice), the proposed framework leveraged standardized preprocessing and a MambaVision backbone that is enhanced with SSL and SRC. This framework achieved a good accuracy of 93.42% and a 93.37% F1 score. Analysis of the effect of cluster count revealed that a 10-cluster SRC configuration best captured latent features, and a two-layer ReLU-activated fully connected head offered optimal classifier connectivity. Computation-wise, NoeViM deployed using Mamba vision as a base model proved to be more efficient and has potential for generalization and practical deployment. These findings confirm the significance of fine-grained clustering, balanced layer design, and MambaVision's strong potential for deployment in low-resource clinical environments.

G. Computational Efficiency

Classical CNNs and ViTs are heavily impacted by quadratic computational scaling. Both models have been used for similar diagnostic tasks [7] in previous related studies using the NJN dataset; hence, the basis for comparison. As depicted in Table V, the findings from this experiment highlight MambaVision's relative advantage over CNN and ViT models in similar classification tasks, achieving superior computational efficiency with 50.14 million parameters and 7.50 GFLOPs, while maintaining a real-time inference speed of 0.2336 milliseconds per image.

V. CONCLUSION

This study introduced NeoViM, an adapted framework achieved by integrating Vision Mamba architecture with Self-Supervised Learning (SSL) for non-invasive neonatal jaundice detection. This proposed model, NeoViM, achieved a good performance (93.42% accuracy, 93.37% F1 score) by incorporating attention-aware Self-Relative Clustering (SRC) with SSL-driven feature refinement as highlighted in the ablation study. An optimal cluster configuration (10 clusters) and a single 128-dim dense layer further maximized discriminative feature extraction. While this proposed approach demonstrates generalization and clinical viability, the study acknowledges that there are limitations in terms of demographic diversity of the dataset; therefore, the study suggests future research in the area of cross-population validation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Fati Oiza Salami developed the conceptual framework, methodology, and prepared the original draft. Youssef Mourchid contributed to the conceptual framework development, original manuscript review, and editing. Muhammad Muzammel contributed to the conceptualization of the framework and original manuscript review and editing. Alice Othmani contributed to the conceptual framework, revision of the manuscript, and final draft; all authors had approved the final version.

FUNDING

This research was funded by the Nigerian Government, provided through the Tertiary Education Trust Fund (TETFund) 2023/2024 scholarship award, award number TETF/ES/UNI/KOGI/TSAS/2023.

REFERENCES

[1] A. Aune, G. Vartdal, H. Bergseng, L. L. Randeberg, and E. Darj, "Bilirubin estimates from smartphone images of newborn infants' skin correlated highly to serum bilirubin levels," *Acta Paediatr.*, vol. 109, no. 12, pp. 2532–2538, Mar. 2020. https://doi.org/10.1111/apa.15287

[2] B. Sreedha, P. R. Nair, and R. Maity, "Non-invasive early diagnosis of jaundice with computer vision," *Procedia Comput. Sci.*, vol. 218, pp. 1321–1334, 2023. https://doi.org/10.1016/j.procs.2023.01.111

[3] F. O. Salami, M. Muzammel, Y. Mourchid, and A. Othmani, "Artificial intelligence non-invasive methods for neonatal jaundice detection: A review," *Artif. Intell. Med.*, vol. 162, 103088, Feb. 2025. https://doi.org/10.1016/j.artmed.2025.103088

[4] T. S. Leung, F. Outlaw, L. W. MacDonald, and J. Meek, "Jaundice Eye Color Index (JECI): Quantifying the yellowness of the sclera in jaundiced neonates with digital photography," *Biomed. Opt. Express*, vol. 10, no. 3, pp. 1250–1264, Mar. 2019. https://doi.org/10.1364/BOE.10.001250

[5] B. Abdulkader and M. H. Aziz, "Neonatal jaundice severity detection from skin images using deep transfer learning techniques," *J. Electron. Electromed. Eng. Med. Inform.*, vol. 7, no. 1, pp. 92–104, Jun. 2024. https://doi.org/10.35882/jeeemi.v7i1.576

[6] S. Dissaneevate *et al.*, "A mobile computer-aided diagnosis of neonatal hyperbilirubinemia using digital image processing and machine learning techniques," *Int. J. Innov. Res. Sci. Stud.*, vol. 5,

no. 1, pp. 10–17, Jan. 2022. https://doi.org/10.53894/ijirss.v5i1.334

[7] K. Gupta, V. Sharma, and S. S. Kathait, "Smart screening: Non-invasive detection of severe neonatal jaundice using computer vision and deep learning," *Int. J. Comput. Appl.*, vol. 186, no. 35, pp. 35–43, Aug. 2024. https://doi.org/10.5120/ijca2024923924

[8] A. Y. Abdulrazzak, S. L. Mohammed, A. Al-Naji, and J. Chahl, "Real-time jaundice detection in neonates based on machine learning models," *BioMedInformatics*, vol. 4, no. 1, pp. 623–637, Feb. 2024. https://doi.org/10.3390/biomedinformatics4010034

[9] A. Y. Abdulrazzak, S. L. Mohammed, and A. Al-Naji, "NJN: A dataset for the normal and jaundiced newborns," *BioMedInformatics*, vol. 3, no. 3, pp. 543–552, Jul. 2023. https://doi.org/10.3390/biomedinformatics3030037

[10] A. Y. Abdulrazzak, S. L. Mohammed, A. Al-Naji, and J. Chahl, "Computer-aid system for automated jaundice detection," *J. Techn.*, vol. 5, no. 1, pp. 8–15, Mar. 2023. https://doi.org/10.51173/jt.v5i1.1128

[11] V. Rani *et al.*, "Self-supervised learning: A succinct review," *Arch. Comput. Methods Eng.*, vol. 30, no. 4, pp. 2761–2775, Jan. 2023. https://doi.org/10.1007/s11831-023-09884-2

[12] T. Xu *et al.*, "A generalizable 3D framework and model for self-supervised learning in medical imaging," arXiv Print, arXiv:2501.11755, Jan. 2025. https://arxiv.org/abs/2501.11755

[13] S. Chattopadhyay, S. Ganguly, S. Chaudhury, and S. Nag, "Exploring self-supervised representation learning for low-resource medical image analysis," arXiv Print, arXiv:2303.02245, Jun. 2023. https://doi.org/10.48550/arXiv.2303.02245

[14] Y. Ali, A. Taleb, M. M. C. Höhne, and C. Lippert, "Self-supervised learning for 3D medical image analysis using 3D SimCLR and Monte Carlo dropout," arXiv Print, arXiv:2109.14288, Mar. 2021. https://doi.org/10.48550/arXiv.2109.14288

[15] A. Hatamizadeh and J. Kautz, "MambaVision: A hybrid Mamba-transformer vision backbone," arXiv Print, arXiv:2407.08083, Mar. 2025. https://doi.org/10.48550/arXiv.2407.08083

[16] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," arXiv Print, arXiv:2312.00752, May 2024. https://doi.org/10.48550/arXiv.2312.00752

[17] J. Zhu, Y. Li, L. Ding, and S. K. Zhou, "Aggregative self-supervised feature learning from limited medical images," in *Proc. Med. Image Comput. Comput. Assist. Interv—MICCAI 2022*, vol. 13438, 2022, pp. 57–66. https://doi.org/10.1007/978-3-031-16452-1_6

[18] 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, p. 7038, Oct. 2021. https://doi.org/10.3390/s21217038

[19] A. Sardana. (2017). Neonatal jaundice detection/dataset/face. GitHub repository [Online]. Available: https://github.com/AshishSardana/jaundice-detection