

Enhancing Conjunctival Image-Based Anemia Classification Using Coordinate Attention on MobileNetV3-Small

Arnanda Asshaffa Mahendra¹, Aris Marjuni², Affandy Affandy³, Muhammad Naufal⁴

¹ Faculty of Intelligent Technology, Postgraduate Program, Universitas Dian Nuswantoro, Semarang, Indonesia

^{2,3,4} Faculty of Intelligent Technology, Universitas Dian Nuswantoro, Semarang, Indonesia

Received:

June 29, 2026

Revised:

July 20, 2026

Accepted:

September 2, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Arnanda Asshaffa
Mahendra

Email*:

arnanda.asshaffa@adm.din
us.ac.id

DOI:

10.63158/journalisi.v8i5.1796

© 2026 Journal of
Information Systems and
Informatics. This open
access article is distributed
under a (CC-BY License)



Abstract. This study evaluates the integration of attention mechanisms (Efficient Channel Attention [ECA] and Coordinate Attention [CoordAtt]) into MobileNetV3-Small for preliminary conjunctival anemia screening. Using a single public dataset of 710 pediatric images from Ghana (CP-AnemiC dataset), the model was evaluated retrospectively via stratified 10-fold cross-validation (80%) and tested on a hold-out test set (20%). CoordAtt achieved the highest performance (Accuracy 78%, Sensitivity 84%, Specificity 70%, AUC 0.85), outperforming the Baseline (Accuracy 75%, Sensitivity 80%, Specificity 67%, AUC 0.80) and ECA (Accuracy 76%, Sensitivity 83%, Specificity 66%, AUC 0.81). A 95% Bootstrap BCa confidence interval test confirmed that the CoordAtt improvement over the Baseline is statistically significant for Accuracy, F1-Macro, Sensitivity, and AUC. Grad-CAM visualizations further showed more focused conjunctival localization. This improvement came with only a minimal increase in theoretical computational footprint (1.06M parameters, 0.0584 GFLOPs), keeping a lightweight profile similar to the Baseline (1.03M parameters, 0.0581 GFLOPs). These results show that CoordAtt improves classification performance without sacrificing computational efficiency, suggesting potential suitability for resource-constrained devices prior to clinical validation.

Keywords: Anemia Classification, Conjunctival Imaging, MobileNetV3-Small, Coordinate Attention, Grad-CAM, Bootstrap BCa, Lightweight Medical Imaging.

1. INTRODUCTION

Anemia remains a pressing global health problem, with no meaningful decline in prevalence, falling hardest on women of reproductive age and children under five [1]. Standard diagnosis via Complete Blood Count (CBC) requires venous blood collection, chemical reagents, and automated analyzers operated by trained personnel, a setup routinely unavailable at primary healthcare facilities in low and middle income countries (LMICs) [2], [3]. Cloud based diagnostics and telemedicine offer little relief where internet connectivity is too unstable to sustain clinical data transmission [4]. Noninvasive preliminary screening that runs offline on low cost edge devices, without any laboratory dependency, is what these settings actually need [5], [6], [7]. The conjunctiva suits this purpose well, its capillary dense tissue and low melanin content mean that visible pallor often correlates with systemic hemoglobin concentration, although this relationship is highly sensitive to ambient lighting, image acquisition quality, and underlying clinical conditions [8]. Computer vision can read that pallor objectively, replacing subjective clinical judgment with deep learning models trained on color and texture features for early triage and screening [8], [9].

Conjunctival image-based anemia screening has attracted growing attention from the deep learning community. Using the public CP-AnemiC dataset, Appiahene et al. [8] benchmarked ResNet50 on 710 pediatric images from Ghana, reporting 84.79% accuracy, but included no attention mechanism and disclosed no computational cost. Ramos-Soto et al. [10] achieved 98.47% accuracy with a Vision Transformer (ViT-B/16) impressive on paper, but at 86 million parameters the model cannot run offline on constrained hardware. Lin et al. [11] added Channel-Spatial Attention to a ResNet50 backbone in their BPANet, reaching 84.9% accuracy; even with the attention module, the architecture is heavy and FLOPs were never reported. Chen et al. [6] built a lightweight 1.08-million-parameter pipeline combining EGE-Unet and DHA for smartphone-based real-time inference, though the system estimates hemoglobin as a continuous value rather than classifying anemia, and relies on a proprietary closed dataset. Fuentes-Beingolea et al. [12] addressed a different problem entirely illumination variability using LAB color space normalization and CLAHE preprocessing to lift accuracy to 88.28%, without touching the classifier architecture or introducing any attention module. While these prior studies established important baselines on the CP-AnemiC dataset, the balance between clinical

screening sensitivity and computational efficiency on lightweight architectures remains underexplored.

Four gaps in the literature still stand between current research and practical AI based anemia screening in resource-limited settings [8], [10]. First, the field has leaned heavily on large-parameter architectures, leaving lightweight models suitable for edge deployment underexplored [8], [13]. Second, compressing a model to run on constrained hardware often blunts its sensitivity to subtle clinical features [11], [14], yet no study has specifically tested whether integrating an attention mechanism into a backbone as light as MobileNetV3-Small can recover that sensitivity for conjunctival anemia classification. Third, MobileNetV3 already contains built-in Squeeze and Excitation (SE) channel attention in its bottleneck blocks [15], so adding an external attention module risks computational redundancy unless the injection point is carefully validated. Therefore, this study specifically tests whether supplementing the built-in SE with external attention still adds clinical value. Fourth, prior evaluations have prioritized accuracy while leaving computational footprint unreported and statistical significance testing largely absent [6], [13].

This study compares two attention mechanisms integrated into MobileNetV3-Small: Efficient Channel Attention (ECA) and Coordinate Attention (CoordAtt). Both were selected over alternatives such as CBAM or Triplet Attention based on parameter efficiency and structural fit. ECA preserves full cross-channel interaction without the dimensionality reduction found in standard SE or CBAM, adding virtually no parameter overhead [16]. CoordAtt encodes horizontal and vertical positions separately, which is expected to suit conjunctival ROI images better than channel-only attention, as the anatomical shape of the conjunctiva requires precise spatial localization of pallor within distinct boundaries [17].

Two hypotheses guide the evaluation: (1) Integrating CoordAtt or ECA will significantly improve the preliminary screening performance (Sensitivity and Specificity) of MobileNetV3-Small compared to the baseline model with only built-in SE. (2) The integration of these modules will not substantially increase the theoretical computational footprint, maintaining suitability for edge devices.

Most prior work stops at reporting mean accuracy. This study does not. Performance is assessed on two dimensions simultaneously: clinical screening metrics and theoretical computational footprint. Differences between architectures are tested with Bootstrap BCa confidence intervals to confirm that the performance gains are statistically robust rather than artifacts of sampling luck [11], [18]. The result is a theoretically lightweight architecture that can serve as an empirical baseline for future noninvasive anemia screening systems prior to formal external clinical validation and edge deployment.

2. METHODS

This study investigates whether integrating external attention mechanisms into MobileNetV3-Small can recover the clinical sensitivity that model compression tends to reduce. In conjunctival anemia screening, pallor gradients are the primary diagnostic features, and lightweight architectures may lack the representational capacity to localize these fine-grained spatial cues while maintaining low parameter counts and FLOPs suitable for edge hardware in resource-limited settings.

To address this, three variants are compared: the baseline MobileNetV3-Small, and two attention-augmented variants using Efficient Channel Attention (ECA) and Coordinate Attention (CoordAtt). ECA targets cross-channel interactions, while CoordAtt integrates direction-aware spatial encoding suited to the conjunctiva's geometry. The evaluation measures clinical screening metrics alongside computational footprint to quantify both the screening performance gains and the associated computational trade-offs.

Figure 1 outlines the experimental pipeline across three phases: data preparation, model training, and final evaluation, implemented in Python with TensorFlow and Keras with a fixed global random seed of 511 for exact reproducibility. Evaluation goes beyond accuracy alone. Each architecture is assessed on clinical screening metrics, raw computational footprint, and Bootstrap BCa significance tests to confirm that the observed performance differences are statistically robust.

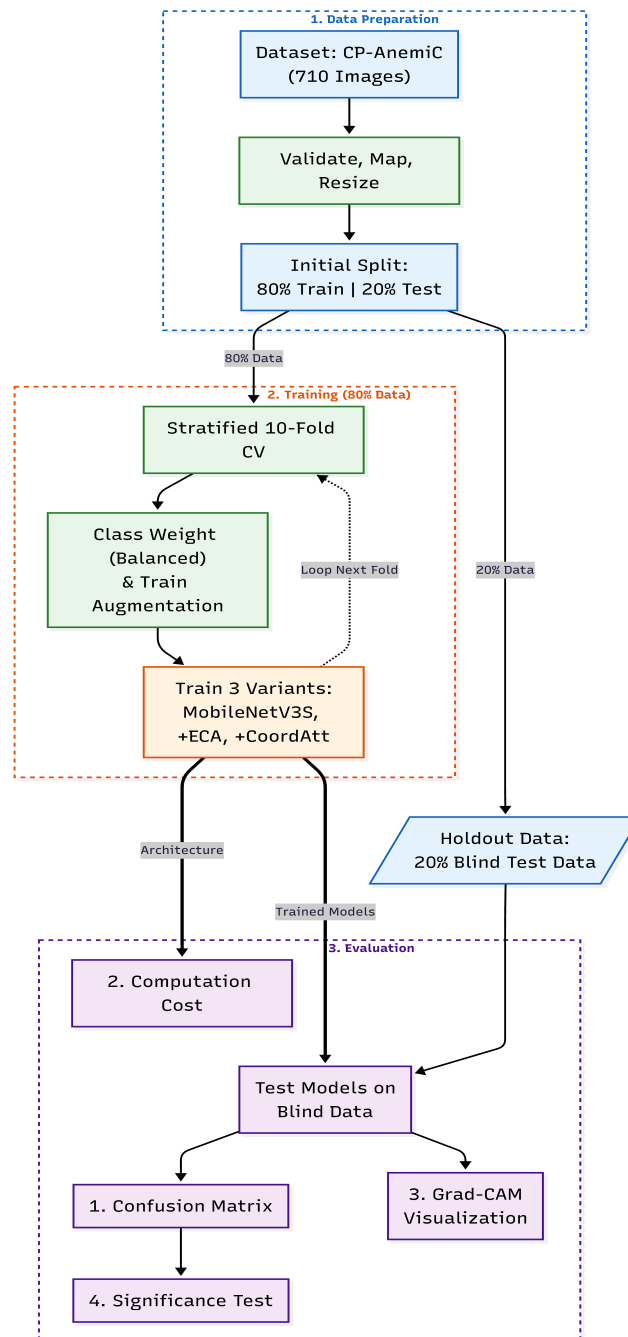


Figure 1. Research Method Flow

2.1. Dataset and Image Acquisition

All experiments use CP-AnemiC, a publicly released conjunctival image dataset on Mendeley Data [19]. The dataset comprises 710 unique individuals, with each image linked to a distinct patient via a unique Image ID [8], [19]. The dataset contains 710 JPG images of conjunctival regions of interest (ROI) from children aged 6 to 59 months, collected at

10 hospitals across Ghana. Images were acquired with a 12-megapixel tablet camera under natural lighting; flash was deliberately avoided to prevent specular reflection on the ocular surface. ROIs were extracted automatically using a combination of triangle thresholding and entropy-based grayscale segmentation. Patient identities were fully anonymized, and the study protocol received ethics approval from the University of Energy and Natural Resources, Ghana [8], [19]. Representative conjunctival ROI samples for both classes are shown in Figure 2.



Figure 2. Sample Conjunctival ROI Image

2.2. Data Preparation

Data preparation started with file integrity checks and class label verification. All images were resized to 224×224 pixels, resizing and pixel normalization were handled on-the-fly through an ImageDataGenerator pipeline, using MobileNetV3-Small built-in preprocessing function to remap raw pixel values from [0, 255] to [-1, 1]. Running these steps dynamically rather than storing preprocessed copies reduces memory overhead and stabilizes gradient updates during training [15]. The dataset was then split using a nested holdout strategy to prevent data leakage [20]. Specifically, 142 images (20%) were set aside as a blind test set before any modeling began and were not touched again until final evaluation. The remaining 568 images (80%) formed the development pool for the training phase [21]. Because the dataset maintains a one-to-one correspondence between images and patients [8], [19], this partitioning prevents patient crossover between the development and test splits. The full dataset of 710 images (424 anemic, 286 non anemic) was divided into a development set of 568 images (339 anemic, 229 non anemic) and a hold-out blind test set of 142 images (85 anemic, 57 non anemic). Within the stratified 10-fold cross-validation on the development set, each fold used approximately 511 images for training (90%) and 57 for validation (10%), with slight per-fold variation inherent to the stratified partitioning.

2.3. Model Architecture and Attention Mechanism

This study evaluated three model variants built on a MobileNetV3-Small backbone pre-trained on ImageNet. The backbone was fully frozen during training to preserve the feature representations learned from ImageNet and to avoid overfitting on the limited dataset size. A preliminary experiment that unfroze the two final blocks consistently degraded blind test performance across all variants, with Accuracy falling to 0.54 to 0.59 and the CoordAtt variant collapsing in six of the ten folds. Full freezing was therefore retained for all reported experiments. The backbone's final feature map has spatial and channel dimensions of $7 \times 7 \times 576$ [15].

Three experimental setups were defined. Setup A (No Attention) passes the feature map directly to the classification head without any attention module. Setup B integrates an Efficient Channel Attention (ECA) module [16], and Setup C inserts a Coordinate Attention (CoordAtt) module [17]. In all cases, the external attention mechanism is injected at the neck position to avoid functional redundancy with the Squeeze-and-Excitation (SE) blocks already built into the MobileNetV3-Small architecture [15]. The overall architecture is shown in Figure 3.

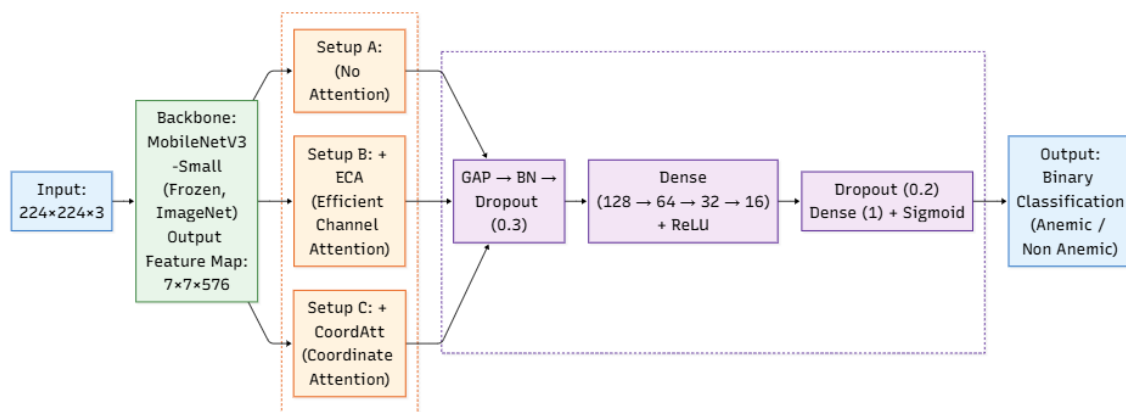


Figure 3. Model Architecture

Following the attention module, a custom classification head maps the extracted feature representations to the final binary output. The head applies Global Average Pooling 2D to reduce spatial dimensions, followed by Batch Normalization and Dropout (rate = 0.3) for regularization. Four fully connected layers then progressively reduce dimensionality from 128 to 64, 32, and 16 units, each using ReLU activation. A second Dropout layer (rate

= 0.2) precedes the output layer, which consists of a single unit with Sigmoid activation to produce the anemia probability. All model variants were optimized using AdamW with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4} , with binary cross-entropy as the loss function.

The ECA module extracts inter-channel relationships by applying Global Average Pooling (GAP) to produce a single spatial vector, followed by a 1D convolution with an adaptive kernel size k (a logarithmic function of the total channel count, $C = 576$) to generate channel probability weights ω_c :

$$\omega_c = \sigma \left(\frac{1}{k} \sum_{j=1}^k w_j y_c^j \right) \quad (1)$$

where σ denotes the Sigmoid activation function, w_j represents the 1D convolution weight parameters, and y_c is the aggregated information for channel c [16]. CoordAtt, by contrast, splits the pooling process into horizontal and vertical directions to preserve awareness of absolute coordinate position, then fuses the result through a 1×1 convolution to reduce dimensionality, before projecting it into a pair of independent spatial weight matrices $Y_c(i, j)$:

$$Y_c(i, j) = X_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (2)$$

where $X_c(i, j)$ is the original input feature matrix, $g_c^h(i)$ and $g_c^w(j)$ denote the directional attention weights along height (vertical) at position i and width (horizontal) at position j [17]. Feature maps from all three variants are fed into an identical Custom Classifier Head block, which progressively reduces dimensionality through Global Average Pooling 2D, Batch Normalization, Dropout (0.3), four ReLU-activated Dense layers $128 \Rightarrow 64 \Rightarrow 32 \Rightarrow 16$ Dropout (0.2), and finally a Sigmoid-activated Dense (1) layer for binary classification.

2.4. Training and Configuration

The training phase was executed on 80% of the data (568 images) using a Stratified 10-Fold Cross-Validation scheme [21]. Class imbalance in this training subset was addressed through a Cost-Sensitive Learning approach based on balanced class weights. This technique was chosen specifically to avoid oversampling or SMOTE, which are prone to overfitting, and undersampling, which risks discarding clinically relevant information in a limited-size medical dataset [22], [23]. The weight coefficients were computed

automatically at each fold iteration based on the actual sample distribution of the training data. These coefficients were then dynamically injected into the Binary Cross-Entropy (BCE) loss function. This modification produced a weighted loss function that imposes a higher mathematical penalty whenever the model misclassifies the minority class [23].

In addition to receiving class weight adjustments, the training subset in each cross-validation fold was also subjected to a conservative image augmentation pipeline [22], while the validation and test sets were kept in their original condition. These spatial transformations included random rotation of $\pm 15^\circ$, horizontal and vertical shifts of 10%, and zoom variation of $\pm 15\%$, with empty areas resulting from the transformation filled using a constant pixel value of zero (black). In contrast, manipulative procedures such as random crop, color distortion, and horizontal flip were deliberately avoided to preserve the integrity of the pallor gradient, the primary diagnostic marker.

Each of the three model variants was trained independently using AdamW with a learning rate of 1×10^{-4} , a batch size of 16, and a maximum of 200 epochs. Early Stopping monitored validation loss with a patience of 20 epochs and a minimum improvement threshold of 0.001, restoring the best weights automatically at the end of each fold. Model Checkpoint saved only the weights corresponding to the lowest validation loss. Before initializing each new model, the backend session was cleared to prevent memory leakage between folds.

2.5. Performance Evaluation

The clinical validity of each model was evaluated on the 20% blind test set. Relying solely on overall accuracy can be misleading in imbalanced medical datasets [24], so clinical discrimination performance was primarily assessed using Sensitivity and Specificity [24]. Sensitivity measures the model's ability to correctly identify anemic cases, while Specificity measures its ability to correctly rule out non anemic individuals. These metrics were derived from a confusion matrix of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), using standard formulations:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (4)$$

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

$$F1_{Macro} = \frac{1}{2} \left(\frac{2TP}{2TP+FP+FN} + \frac{2TN}{2TN+FP+FN} \right) \quad (6)$$

The arithmetic mean over both classes is used to ensure fair evaluation on imbalanced data [11], [22]. Model calibration was additionally assessed because discrimination metrics alone do not reflect the reliability of predicted probabilities [25]. ROC curves were extracted to obtain Area Under Curve (AUC) values, with Anemia as the positive class [26]. Final metrics are reported as the mean across ten separately trained models, each evaluated independently on the same blind test set, rather than selecting a single best fold or applying ensemble voting.

2.6. Computation Cost

The second evaluation dimension measured the feasibility of offline model implementation for edge deployment. This computational complexity profile was assessed theoretically by extracting three key metrics from the model architecture: total parameters, estimated computational load in Floating Point Operations (FLOPs), and model file size footprint in float32 binary format [13]. To complement this theoretical profile, the three trained models were exported to TensorFlow Lite float32 format and profiled on a commercial smartphone (Samsung Galaxy A55 5G, Exynos 1480, 8 GB RAM) through the Android Debug Bridge. Inference latency was measured with the native TensorFlow Lite benchmarking tool using four CPU threads, a two second warm-up, and a ten second measurement window repeated three times per model. Model size, initialization time, and approximate memory footprint were recorded from the same tool. All measurements were taken on a single device unit (SM-A556E) running Android 16 with One UI 8.0. During testing, the device was connected to the host computer through USB debugging and operated with no applications running in the background, in an air-conditioned room at room temperature, and the battery level was at 55%. Because the benchmark tool reports per-iteration percentiles, mild CPU frequency scaling was observable in the third consecutive run; the reported means and standard deviations therefore include this variability rather than excluding it.

2.7. Model Explainability Visualization (Grad-CAM)

To address the black-box limitation that constrains clinical transparency [26], [27], this study implemented Gradient-weighted Class Activation Mapping (Grad-CAM) [28] to visually validate the model's prioritization of relevant diagnostic features over background noise [10], [29]. Grad-CAM extracts spatial gradients from the last convolutional layer before the attention block and converts them into a heatmap superimposed on the original image. This visualization facilitates a comparative observation of how well the Efficient Channel Attention (ECA) and Coordinate Attention (CoordAtt) modules sharpen the model's focus on the conjunctival pallor gradient compared to the MobileNetV3-Small [10]. This visualization offers qualitative support for assessing whether the model's focus corresponds to clinically relevant conjunctival features [26], [30].

2.8. Statistical Significance Test

To demonstrate that the performance improvement was not the result of random fluctuation, this study adopted the non-parametric Bootstrap Bias-Corrected and Accelerated (BCa) significance test [18], [31]. The procedure was carried out by resampling with replacement from the paired metric differences between model variants across the ten folds when predicting the blind test set, replicated over 10,000 computational iterations [32]. BCa was chosen to correct for skewness and bias arising from the small sample size ($n=10$) [33]. The performance advantage was considered statistically significant when the 95% Confidence Interval (CI) did not include zero at all, mathematically indicated by a lower bound greater than 0 [31].

3. RESULTS AND DISCUSSION

3.1. Training and Validation Convergence

Learning dynamics from the 10-Fold Cross-Validation scheme were aggregated into a single validation curve for accuracy and loss metrics in Figure 4. The x-axis was deliberately truncated at the median of the highest stopping epoch across folds (out of a maximum of 200 epochs) to mitigate visual bias from a small number of anomalous folds. Folds that triggered Early Stopping earlier had their values forward-filled to maintain consistency in the aggregated array. The main line represents the mean across the 10 folds, while the shaded area shows the standard deviation. The curve was

smoothed using an exponential moving average (factor 0.7) to reduce fluctuation noise and clarify the trend.

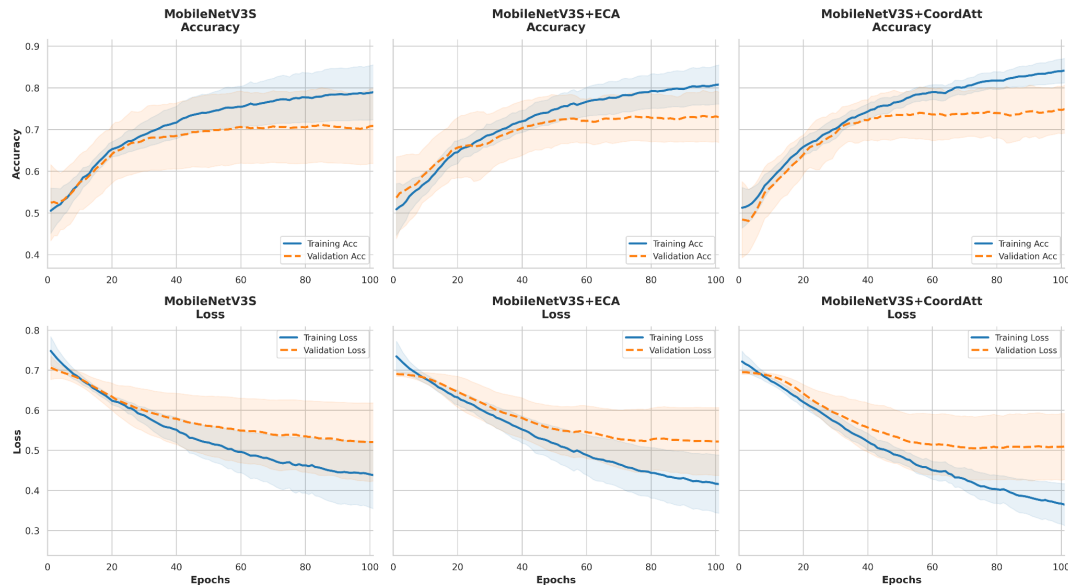


Figure 4. Learning Curve 10-Fold

Based on the resulting plot, the CoordAtt architecture achieved the highest average final Validation Accuracy, approaching 0.75, surpassing both ECA (around 0.73) and the Baseline (around 0.71). Although it started with the lowest accuracy, its improvement trajectory crossed over and overtook both other models between roughly epoch 20 and epoch 30. Correspondingly, CoordAtt's Validation Loss consistently recorded the lowest final value, around 0.50. A sharp drop in loss and steep rise in accuracy occurred up to epoch 60. From epoch 60 onward to the point where the curve was truncated, the trajectories began to flatten, with a stable deviation band width, indicating that all architectures had reached model convergence. The CoordAtt variant's convergence with the lowest loss value and a narrowing deviation band suggests that the spatial information retained by the Coordinate Attention mechanism contributes to a more stable optimization process compared to both the ECA variant and the MobileNetV3-Small.

3.2. Confusion Matrix and Clinical Performance Evaluation

Table 1 presents the average confusion matrix across the 10-fold cross-validation for each model evaluated on the 142 images in the blind test set, where the MobileNetV3-Small row denotes the Baseline model.

Table 1. Confusion Matrix Mean 10 Fold

Model	TP	FN	FP	TN
MobileNetV3-Small	67.9	17.1	18.9	38.1
+ ECA	70.9	14.1	19.5	37.5
+ CoordAtt	71.3	13.7	17.2	39.8

Across all ten folds, the CoordAtt model averaged the highest True Positives (71.3) and the lowest False Negatives (13.7), outperforming the Baseline (17.1 FN) and ECA (14.1 FN). In a preliminary screening context, minimizing False Negatives is the primary clinical concern, as missed anemia cases delay further laboratory confirmation. CoordAtt successfully reduced average False Negatives while also lowering average False Positives compared to the Baseline (17.2 vs. 18.9). However, it is worth noting that any remaining False Positives (averaging 17.2 per fold) can increase unnecessary referrals, potentially burdening already limited primary-care facilities. The clinical performance of the three architecture variants was evaluated using Accuracy, F1-Macro, Sensitivity, Specificity, and AUC, with 95% confidence intervals (CI) computed via t-distribution ($df = 9$) across the ten folds on the blind test set. Results are presented in Table 2.

Table 2. Clinical Performance Comparison on Blind Test Set (Mean and 95% CI)

Model	Acc	F1-Macro	Sens	Spec	AUC
MobileNetV3-Small	0.75 [0.68,0.81]	0.73 [0.65,0.81]	0.80 [0.72,0.87]	0.67 [0.54,0.80]	0.80 [0.72,0.88]
+ ECA	0.76 [0.72,0.81]	0.73 [0.64,0.83]	0.83 [0.79,0.88]	0.66 [0.49,0.83]	0.81 [0.74,0.88]
+ CoordAtt	0.78 [0.76,0.81]	0.77 [0.75,0.80]	0.84 [0.81,0.87]	0.70 [0.68,0.72]	0.85 [0.83,0.87]

CoordAtt recorded the highest performance across all five metrics (Table 2), with the narrowest confidence intervals among the variants. Integrating coordinate attention into MobileNetV3-Small appears to improve spatial localization of diagnostic features, which may contribute to the performance advantage over the channel-only ECA variant.

CoordAtt also produced the narrowest confidence intervals across all five metrics (widths of 0.04 to 0.06), compared with considerably wider intervals for ECA, particularly for Specificity (width 0.34) and F1-Macro (width 0.18), and intermediate variability for the Baseline (Specificity width 0.27). The narrower intervals suggest that the spatial attention

module reduces sensitivity to training data variation across folds, particularly for Specificity. Positive Predictive Value (PPV) and Negative Predictive Value (NPV) were also computed as supplementary clinical indicators, with CoordAtt achieving an average PPV of 0.81 and NPV of 0.75. These values reflect a screening trade-off: high sensitivity reduces missed anemia cases, but a Specificity of 0.70 means some healthy individuals will be flagged, adding potential referral burden on primary-care facilities, while an NPV of 0.75 indicates that negative predictions carry reasonable confidence, though not sufficient for standalone clinical decisions without further laboratory confirmation.

3.3. Computation Cost Evaluation

Details on the number of operational parameters, estimated computational load (Floating Point Operations), and memory footprint of the three models are compared descriptively in Table 3.

Table 3. Computation Cost Comparison

Model	Total Parameters	Trainable	GFLOPs	Model Size
MobileNetV3-Small	1,026,161	85,889	0.0581	3.91 MB
+ ECA	1,026,166	85,894	0.0581	3.91 MB
+ CoordAtt	1,057,337	117,029	0.0584	4.03 MB

The ECA module adds only five trainable parameters without changing the computational load or memory footprint, while CoordAtt adds 31,176 parameters and raises the load only marginally, from 0.0581 to 0.0584 GFLOPs. Both modified variants remain within a four megabyte footprint, indicating that injecting attention at the neck position balances the trade-off between accuracy and network complexity. This theoretical profile was confirmed empirically on a Samsung Galaxy A55 5G through TensorFlow Lite Float32 profiling, as summarized in Table 4. All variants achieved real time inference, with the proposed CoordAtt model reaching a mean latency of 2.25 ± 0.48 ms with a 4.21 MB model file, and the TFLite conversion matched the original Keras models with no classification changes across the 142 test images.

Table 4. On-device inference profiling on a Samsung Galaxy A55 5G

Model	Model Load	Memory Footprint	Latency	Model Size
MobileNetV3-Small	25.42 ms	15.54 MB	2.11 ± 0.52 ms	4.08 MB

Model	Model Load	Memory Footprint	Latency	Model Size
+ ECA	21.20 ms	15.67 MB	2.16±0.52 ms	4.08 MB
+ CoordAtt	17.00 ms	15.91 MB	2.25±0.48 ms	4.21 MB

3.4. Grad-CAM Visualization

Spatial activation patterns of MobileNetV3-Small, ECA, and CoordAtt were visualized using Gradient-weighted Class Activation Mapping (Grad-CAM) derived from the best-performing fold checkpoint across four diagnostic outcome categories on the blind test set: True Positive, True Negative, False Positive, and False Negative, as shown in Figure 5. Unlike the averaged clinical metrics, Grad-CAM maps represent a single fold checkpoint. Heatmaps were qualitatively assessed against anatomical conjunctival landmarks.

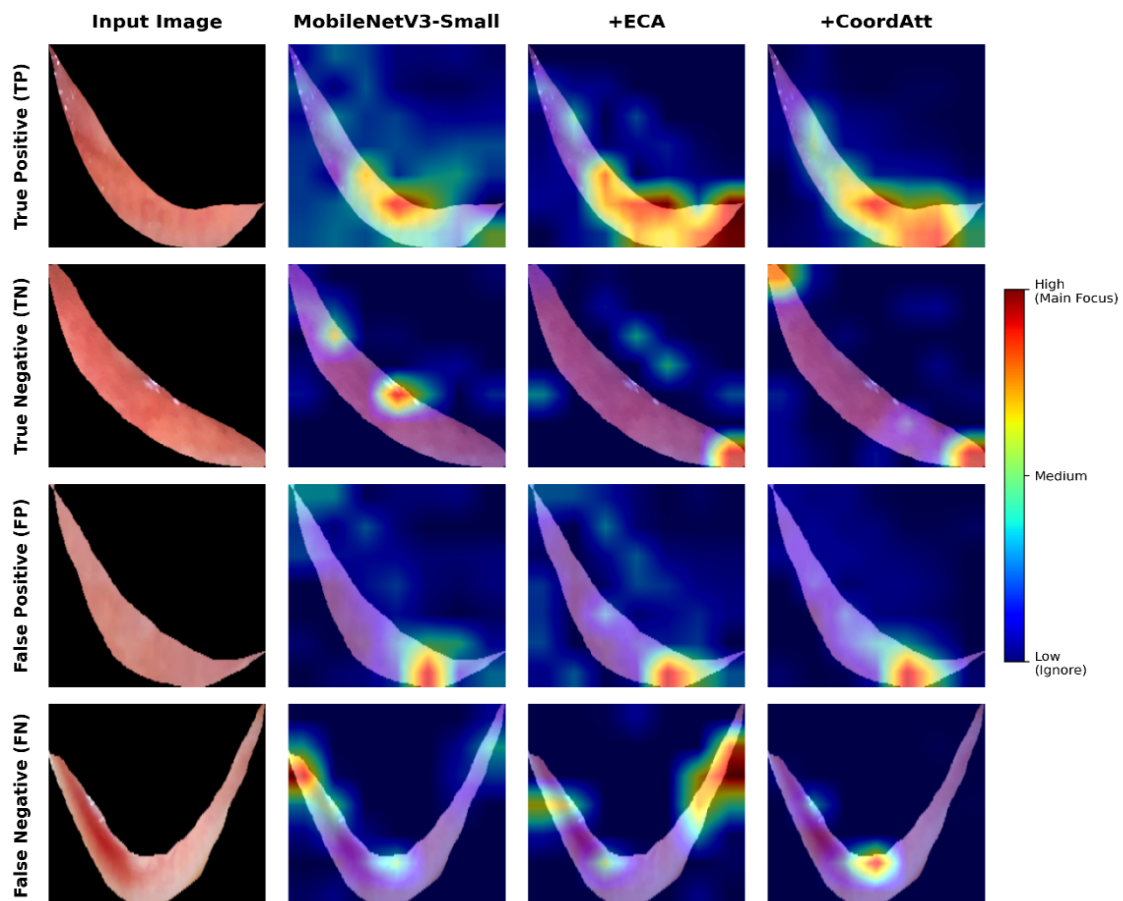


Figure 5. Grad-CAM heatmaps for TP, TN, FP, and FN cases across evaluated models

In correctly classified cases (TP and TN), CoordAtt shows closer anatomical alignment with the conjunctiva compared to the Baseline and ECA. For TP cases, CoordAtt generates

a concentrated high-intensity activation zone (red-to-yellow spectrum) centered on the central, while ECA shows activation extending beyond tissue boundaries into the background. For TN cases, CoordAtt suppresses activation across healthy tissue (blue spectrum), whereas the Baseline produces scattered hotspots on non anemic mucosal regions. Analysis of failure cases (FP and FN) reveals the following patterns. FP All three architectures show occasional warm activations at tissue margins or areas affected by specular lighting, suggesting residual sensitivity to acquisition artifacts rather than clinical anemia indicators. FN The Baseline and ECA show activation spreading into background regions. CoordAtt re-centers focus on the lower conjunctival fold but produces activation below the positive classification threshold. These Grad-CAM patterns are qualitative indicators of gradient-based saliency and should not be interpreted as direct evidence of causal decision pathways.

3.5. Statistical Significance

In the ECA versus Baseline comparison, the 95% confidence intervals included zero for all five metrics, indicating no statistically significant difference. In contrast, the CoordAtt versus Baseline comparison excluded zero for Accuracy, F1-Macro, Sensitivity, and ROC-AUC, confirming that its improvements are statistically significant, while Specificity remained non-significant with an interval that included zero. These statistical validation results are presented in Table 5.

Table 5. Bootstrap BCa Statistical Significance Results (95% CI)

Metrics	ECA (Δ)	95% CI (ECA)	CoordAtt (Δ)	95% CI (CoordAtt)
Test Accuracy	+0.0169	[-0.0141, +0.0676]	+0.0359	[+0.0043, +0.0894]
Test F1	+0.0048	[-0.0566, +0.0882]	+0.0429	[+0.0046, +0.1317]
Sensitivity	+0.0353	[-0.0270, +0.2064]	+0.0400	[+0.0071, +0.0977]
Specificity	-0.0105	[-0.2088, +0.1368]	+0.0298	[-0.0368, +0.2053]
Test ROC-AUC	+0.0163	[-0.0264, +0.1088]	+0.0497	[+0.0085, +0.1504]

The delta (Δ) values represent the performance difference of the attention-based models compared to the Baseline (MobileNetV3-Small).

The resampled BCa distributions shifted toward positive differences for CoordAtt across all metrics, with more compact spreads than ECA, consistent with the interval results presented in Table 5. This improvement is not the result of random fluctuation in the

training data distribution, in contrast to the ECA variant, whose confidence interval still includes zero. The exclusion of zero from the lower bound for the majority of metrics therefore provides clear statistical support for CoordAtt's performance advantage over the Baseline model.

Because screening deployment depends not only on discrimination but also on the reliability of predicted probabilities, model calibration was assessed on the blind test set. Figure 6 shows the reliability diagram pooled from the ten fold-trained models. All three variants track the diagonal closely in the mid and high probability ranges, while the largest deviations occur in the low probability bins, where predictions are slightly less reliable. The proposed CoordAtt model achieved the lowest Brier score of 0.1573 (95% CI, 0.1469 to 0.1677), compared with 0.1667 (95% CI, 0.1451 to 0.1883) for ECA and 0.1709 (95% CI, 0.1448 to 0.1970) for the Baseline, indicating that the improvement of CoordAtt extends beyond classification accuracy to probabilistic reliability.

Sensitivity, specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) were examined across decision thresholds from 0.1 to 0.9 (Figure 6). At the default threshold of 0.5, which was fixed during training and not tuned on the test set, the proposed CoordAtt model achieved a sensitivity of 0.8388 and a specificity of 0.6982, compared with 0.7988 and 0.6684 for the Baseline and 0.8341 and 0.6614 for ECA. Lowering the threshold increased sensitivity at the cost of specificity: at a threshold of 0.4, CoordAtt reached a sensitivity of 0.8871 while retaining a specificity of 0.6211, and its sensitivity decreased more gradually across thresholds than that of the Baseline and ECA. This behavior is consistent with a screening design in which missed anemia cases carry a higher clinical cost than false referrals, while the steadily increasing PPV of CoordAtt at higher thresholds, reaching 0.9430 at a threshold of 0.9, supports its use for confirming positive predictions when the operating point is adjusted toward specificity.

Left: calibration curves (reliability diagram) pooled from ten folds for the Baseline, ECA, and CoordAtt models, with the dotted diagonal indicating perfect calibration. Right: sensitivity, specificity, PPV, and NPV across decision thresholds from 0.1 to 0.9 for each model; the dashed vertical line marks the default threshold of 0.5.

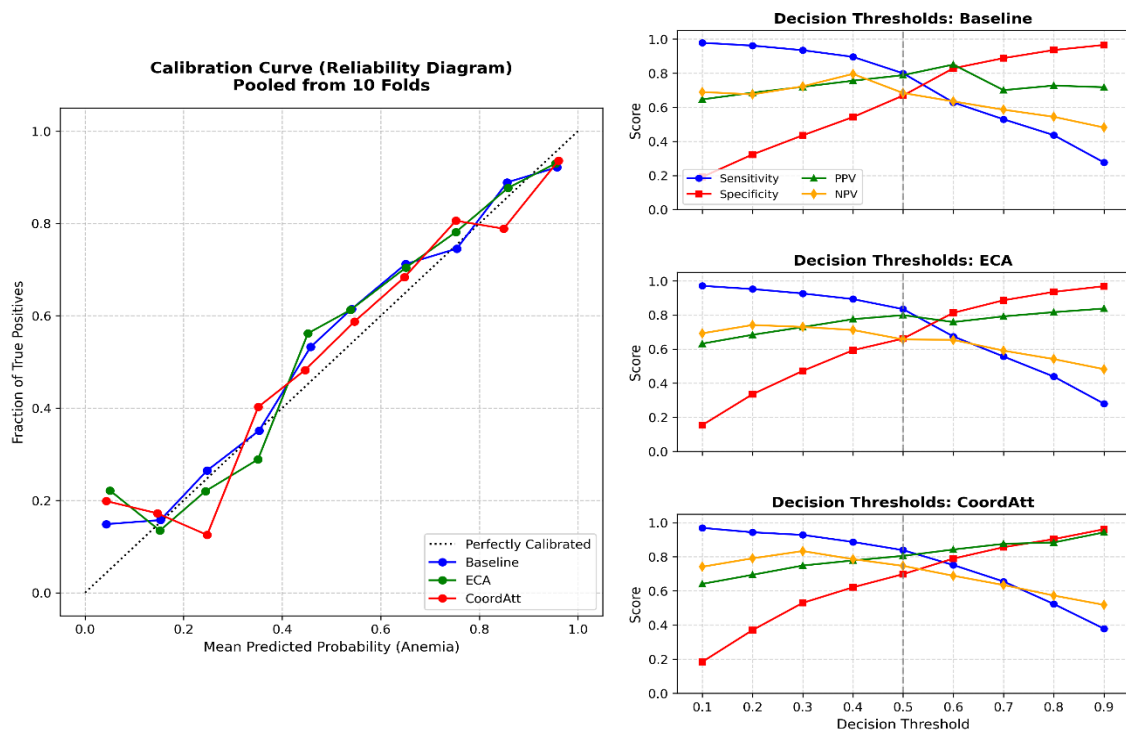


Figure 6. Calibration and decision threshold analysis.

3.6. Discussion

3.6.1. Interpretation of Results

A comprehensive synthesis of the full testing sequence on conjunctival image-based classification architectures confirms that the Coordinate Attention (CoordAtt) mechanism delivers the strongest performance among the architectures evaluated in this study. Integrating CoordAtt into the lightweight MobileNetV3-Small backbone consistently produced the highest clinical performance across all evaluation metrics (Table 3). This diagnostic advantage was accompanied by the lowest inter-fold fluctuation among the variants, indicating consistent performance across data splits despite variability in lighting and acquisition angle.

The Bootstrap BCa results confirm a performance advantage for CoordAtt over the Baseline across training splits. Because the 10 folds checkpoints are evaluated on a fixed blind test set, the confidence intervals primarily reflect fold-wise training stability rather than test-sample variance; the exclusion of zero from the BCa intervals nonetheless indicates that CoordAtt's advantage is consistent rather than incidental. The absolute gains are modest, approximately 3.6 percentage points in Accuracy (0.75 to 0.78) and 5.0

percentage points in ROC-AUC (0.80 to 0.85), but may carry practical value in a screening setting by reducing missed anemia cases prior to confirmatory laboratory testing. The improvement should therefore be interpreted as promising but not clinically sufficient without external validation. This gain is achieved through spatial coordinate attention with minimal parameter and computational overhead, which is relevant for resource-constrained deployment.

The BCa confidence intervals show that the improvement in Specificity was not statistically significant; CoordAtt's higher average Specificity notwithstanding, its confidence interval included zero. In resource-constrained primary-care settings, this Specificity level means a proportion of healthy individuals will be flagged as anemic, adding burden to limited laboratory testing capacity. This trade-off should be considered when determining deployment thresholds, which may need to be adjusted based on local facility capacity and prevalence, as illustrated by the threshold analysis showing sensitivity rising to 0.8871 at a threshold of 0.4.

CoordAtt embeds direction-aware and position-sensitive information into channel features, enabling the model to capture spatial dependencies along the narrow geometry of the conjunctiva. In this dataset, this formulation appears to support more precise feature localization compared to channel-only attention. Grad-CAM observations provide qualitative context: the Baseline and ECA show dispersed activation with spillover into background regions, while CoordAtt consistently generates localized activation centered on the conjunctival region. This localization pattern is consistent with the anatomical area used clinically to evaluate pallor, suggesting that spatial coordinate encoding helps suppress background regions. These findings are specific to the CP-AnemiC dataset. They should not be generalized to other conjunctival imaging settings without further empirical validation. Failure mode analysis reveals persistent challenges across all architectures related to specular highlights and edge illumination at tissue margins.

Image acquisition conditions affect the reliability of noninvasive screening models in practice. Variations in ambient lighting, camera hardware, ROI extraction quality, and inter-patient differences in skin and ocular pigmentation can introduce domain shifts that affect generalizability. Because this evaluation relies on a single dataset, how well

the observed activation patterns transfer to heterogeneous acquisition environments requires further empirical validation.

From a computational engineering perspective, all of CoordAtt's gains in predictive localization were achieved without sacrificing efficiency. This substantial performance improvement came at the cost of only a relatively small increase in parameter count and computational load (FLOPs), meaning the model retains the fundamental lightweight character of MobileNetV3-Small, with a total parameter count of approximately 1.06 million. A direct efficiency comparison with prior studies was not possible, however, since none reported computational complexity metrics such as parameter count, FLOPs, or model size. Beyond efficiency, this study extends the depth of evaluation through a more comprehensive framework combining classification performance, statistical significance testing, inter-fold stability, visual interpretability, and computational efficiency. The main contribution of this study therefore does not rest on achieving the highest absolute accuracy, but on this combined evaluation of the Coordinate Attention mechanism within a lightweight architecture. It supports deployment on resource-constrained edge devices, enabling non-invasive anemia screening without full reliance on cloud network connectivity, and is further supported by on-device profiling on a Samsung Galaxy A55 5G, where the proposed model achieved a mean inference latency of 2.25 ms with a 4.21 MB TensorFlow Lite model file (Table 5). Clinical validation remains a necessary step before real-world deployment. Prospective, multi-center field trials are needed to validate diagnostic thresholds across diverse patient populations, assess integration into primary healthcare workflows, and meet regulatory and ethical requirements for AI-assisted medical screening.

3.6.2. Comparison with Previous Studies

Table 6 summarizes the comparison with previous studies. The study by Appiahene et al. [8] is the most relevant comparison, as it evaluated various deep learning architectures using the same CP-AnemiC dataset used in this study.

Table 6. Performance Comparison with Previous Studies on the CP-AnemiC Dataset

Model	Accuracy	Sens	Spec	AUC	Parameters
ResNet50	0.8479	0.830	N/R	0.835	N/R

Model	Accuracy	Sens	Spec	AUC	Parameters
ViT	0.8408	0.835	N/R	0.841	N/R
VGG16	0.8056	0.785	N/R	0.791	N/R
DenseNet121	0.7958	0.788	N/R	0.793	N/R
MobileNetV3-Small + CoordAtt (Proposed)	0.7824	0.8388	0.6982	0.8463	~1.06M

In terms of performance comparison, the proposed MobileNetV3-Small with Coordinate Attention integration did not reach the highest absolute accuracy previously achieved by larger-capacity models such as ResNet50 (0.8479) and ViT (0.8408). However, the proposed architecture achieved a Sensitivity of 0.8388 and a ROC-AUC of 0.8463, both competitive with and comparable to the highest values reported in the Appiahene et al. study. In an early-stage medical screening system for anemia detection, these high Sensitivity and ROC-AUC values carry critical clinical value, as they relate directly to the system's ability to identify positive anemia cases and minimize the risk of false negatives during initial screening, where a failure to detect a positive case could delay further laboratory testing and medical intervention. Although the reference study [8] did not report computational metrics explicitly, standard architecture literature [34] and recent reviews of lightweight models [13] note that models such as ResNet50 and ViT-Base [10] require substantially larger parameter loads of approximately 25.6 million and 86 million, respectively. Therefore, the minor gap in absolute accuracy represents a rational and acceptable computational trade-off.

3.6.3. Limitations and Future Work

Several methodological limitations apply to this study. First, evaluation relied on the single CP-AnemiC dataset from one demographic population in Ghana with a predominance of pediatric patients; generalizability to populations with different racial characteristics, skin pigmentation, age ranges, or geographic regions remains uncertain. Second, although partial fine-tuning of the two final blocks was evaluated and found to degrade performance consistently, a more systematic search of fine-tuning strategies, including layer-wise learning rates and quantization-aware fine-tuning, remains unexplored. Third, on-device evaluation was limited to inference latency, initialization time, and approximate memory footprint on a single Android device using float32

TensorFlow Lite models; post-training INT8 quantization degraded classification performance and was therefore not deployed, leaving quantization-aware training unexplored. Fourth, the model's Specificity of 0.70 produces a clinical trade-off: while high Sensitivity reduces missed cases, the resulting false positives may increase referral burden on primary-care facilities. Fifth, statistical validation via Bootstrap BCa relied on evaluating fold checkpoints against a fixed blind test set; while confirming fold-wise training stability, this scheme does not directly quantify test-sample population variance.

Future work should address these limitations across several directions. Dataset expansion should include multi-population cohorts, varying acquisition devices, and diverse lighting conditions, with external validation on independent data. Future statistical benchmarking should incorporate test-image level resampling or nested cross-validation to capture population-level sampling variability. Architectural exploration should compare CoordAtt against CBAM, Triplet Attention, and SimAM, with partial fine-tuning of the backbone. For deployment readiness, building on the float32 profiling reported in this study, FP16 or INT8 deployment should pursue quantization-aware training, together with energy profiling across diverse device classes. Prospective clinical testing in primary healthcare settings remains an essential step toward a deployable noninvasive anemia screening system.

4. CONCLUSION

This study concludes that integrating Coordinate Attention (CoordAtt) into MobileNetV3-Small improves spatial feature localization for conjunctival image-based anemia classification. Grad-CAM visualizations show more focused activation on the conjunctival region, and Bootstrap BCa testing confirms statistically significant gains in Sensitivity and ROC-AUC over the Baseline, with minimal increase in computational footprint. As its primary contribution, this work provides preliminary evidence for a lightweight architecture that may support future edge-based anemia screening in primary healthcare facilities with limited computational resources. The model is not yet clinically validated and is limited to retrospective pediatric images from the CP-AnemiC dataset; external validation across diverse cohorts is required before clinical application. On-device profiling of TensorFlow Lite float32 models on a Samsung Galaxy A55 5G confirmed real-time inference, with the proposed model achieving a mean latency of 2.25 ms at

approximately 445 frames per second and a 4.21 MB model file, with conversion verified as numerically lossless against the original Keras models. Future work should prioritize quantization-aware training for integer-only deployment, clinician-reviewed explainability, and prospective clinical evaluation in primary healthcare settings.

REFERENCES

- [1] World Health Organization, "Global Anaemia Estimates 2025 Edition," World Health Organization, 2025. Accessed: Oct. 12, 2025. [Online]. Available: https://www.who.int/data/gho/data/themes/topics/anaemia_in_women_and_children
- [2] V. Haggemüller *et al.*, "Smartphone-based point-of-care anemia screening in rural Bihar in India," *Commun. Med.*, vol. 3, no. 1, p. 38, Mar. 2023, doi: 10.1038/s43856-023-00267-z.
- [3] H. Yadav, D. Shah, S. Sayed, S. Horton, and L. F. Schroeder, "Availability of essential diagnostics in ten low-income and middle-income countries: results from national health facility surveys," *Lancet Glob. Health*, vol. 9, no. 11, pp. e1553–e1560, Nov. 2021, doi: 10.1016/S2214-109X(21)00442-3.
- [4] A. Al-Ganad, A. Al-Shahdhi, O. Al-Dhaifi, E. Hajeb, H. Hajeb, and A. Al-Motarreb, "Deploying medical AI in low-resource settings: a scoping review of challenges and strategies," *Front. Digit. Health*, vol. 8, p. 1743634, Apr. 2026, doi: 10.3389/fdgth.2026.1743634.
- [5] S. Ghosal, D. Das, V. Udutalapally, and P. N. Wasnik, "iNAP: A Hybrid Approach for NonInvasive Anemia-Polycythemia Detection in the IoMT," *ACM Trans. Comput. Healthc.*, vol. 3, no. 3, pp. 1–28, Jul. 2022, doi: 10.1145/3503466.
- [6] Y. Chen, X. Hu, Y. Zhu, X. Liu, and B. Yi, "Real-time non-invasive hemoglobin prediction using deep learning-enabled smartphone imaging," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, p. 187, Jul. 2024, doi: 10.1186/s12911-024-02585-1.
- [7] Muljono, S. A. Wulandari, H. A. Azies, M. Naufal, W. A. Prasetyanto, and F. A. Zahra, "Breaking Boundaries in Diagnosis: Non-Invasive Anemia Detection Empowered by AI," *IEEE Access*, vol. 12, pp. 9292–9307, 2024, doi: 10.1109/ACCESS.2024.3353788.
- [8] P. Appiahene, K. Chaturvedi, J. W. Asare, E. T. Donkoh, and M. Prasad, "CP-AnemiC: A conjunctival pallor dataset and benchmark for anemia detection in children," *Med. Nov. Technol. Devices*, vol. 18, p. 100244, Jun. 2023, doi: 10.1016/j.medntd.2023.100244.

- [9] A. S. Babu and S. Parthasarathy, "Computer-Vision-Based Non-Invasive Detection of Anemia from Medical Images: A Scoping Review of Artificial Intelligence with Applications in Drug Delivery and Therapeutic Monitoring," *Int. J. Drug Deliv. Technol.*, vol. 16, no. 2, Feb. 2026, doi: 10.25258/ijddt.16.2.23.
- [10] O. Ramos-Soto, J. Ramos-Frutos, E. Perez-Zarate, D. Oliva, S. J. Mousavirad, and S. E. Balderas-Mata, "Non-invasive anemia detection from conjunctiva and sclera images using vision transformer with attention map explainability," *Sci. Rep.*, vol. 15, no. 1, p. 44142, Dec. 2025, doi: 10.1038/s41598-025-32343-w.
- [11] E.-T. Lin *et al.*, "Deep Learning-Based Model for Non-invasive Hemoglobin Estimation via Body Parts Images: A Retrospective Analysis and a Prospective Emergency Department Study," *J. Imaging Inform. Med.*, vol. 38, no. 2, pp. 775–792, Aug. 2024, doi: 10.1007/s10278-024-01209-4.
- [12] J. H. Fuentes-Beingolea, F. Palomino-Quispe, J. C. Herrera-Levano, W. Vargas-Mateos, R. Florez, and A. B. Alvarez, "Illumination-Robust Conjunctival Image Preprocessing for Accurate Segmentation and Anemia Detection Using Deep Learning," *Int. J. Online Biomed. Eng. IJOE*, vol. 21, no. 07, pp. 106–124, Jun. 2025, doi: 10.3991/ijoe.v21i07.54439.
- [13] H.-I. Liu *et al.*, "Lightweight Deep Learning for Resource-Constrained Environments: A Survey," *ACM Comput. Surv.*, vol. 56, no. 10, pp. 1–42, Oct. 2024, doi: 10.1145/3657282.
- [14] J. H. Fuentes-Beingolea and F. Palomino-Quispe, "Adaptive Contrastive Learning in Vision Transformer for Non-Invasive Hemoglobin Estimation From Conjunctival Images," *IEEE Access*, vol. 13, pp. 210257–210270, 2025, doi: 10.1109/ACCESS.2025.3638640.
- [15] A. Howard *et al.*, "Searching for MobileNetV3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.
- [16] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 11531–11539. doi: 10.1109/CVPR42600.2020.01155.
- [17] Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 13708–13717. doi: 10.1109/CVPR46437.2021.01350.

- [18] F. García Fernández, P. De Palacios, A. García-Iruela, and L. G. Esteban, "Using Bootstrapping to Determine Artificial Neural Network Confidence Intervals—Case Study of Particleboard Internal Bond Determined from Production Data," *Appl. Sci.*, vol. 15, no. 8, p. 4554, Apr. 2025, doi: 10.3390/app15084554.
- [19] Justice Williams Asare, "CP-AnemiC (A Conjunctival Pallor) Dataset from Ghana." Mendeley, Mar. 02, 2023. doi: 10.17632/M53VZ6B7FX.1.
- [20] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging," *Radiol. Artif. Intell.*, vol. 5, no. 4, p. e220232, Jul. 2023, doi: 10.1148/ryai.220232.
- [21] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Front. Nanotechnol.*, vol. 4, p. 972421, Aug. 2022, doi: 10.3389/fnano.2022.972421.
- [22] E. Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 12561–12605, Nov. 2023, doi: 10.1007/s10462-023-10453-z.
- [23] L. Cui *et al.*, "Towards Reliable Healthcare Imaging: A Multifaceted Approach in Class Imbalance Handling for Medical Image Segmentation," *Interdiscip. Sci. Comput. Life Sci.*, vol. 17, no. 3, pp. 614–633, Sep. 2025, doi: 10.1007/s12539-025-00726-2.
- [24] S. H. Park and K. Han, "Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction," *Radiology*, vol. 286, no. 3, pp. 800–809, Mar. 2018, doi: 10.1148/radiol.2017171920.
- [25] G. S. Collins *et al.*, "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, Apr. 2024, doi: 10.1136/bmj-2023-078378.
- [26] X. Chen *et al.*, "Recent advances and clinical applications of deep learning in medical image analysis," *Med. Image Anal.*, vol. 79, p. 102444, Jul. 2022, doi: 10.1016/j.media.2022.102444.
- [27] B. H. M. Van Der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, p. 102470, Jul. 2022, doi: 10.1016/j.media.2022.102470.
- [28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

- [29] K. Borys *et al.*, "Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches," *Eur. J. Radiol.*, vol. 162, p. 110786, May 2023, doi: 10.1016/j.ejrad.2023.110786.
- [30] H. Chen, C. Gomez, C.-M. Huang, and M. Unberath, "Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review," *Npj Digit. Med.*, vol. 5, no. 1, p. 156, Oct. 2022, doi: 10.1038/s41746-022-00699-2.
- [31] W. Du, "When +1% is Not Enough: A Paired Bootstrap Protocol for Evaluating Small Improvements," in *Advances in Knowledge Discovery and Data Mining*, vol. 16597, R. C.-W. Wong, H. Tong, H. Lu, J. Kwok, F. Salim, Y. Song, and M. L. Yiu, Eds., in *Lecture Notes in Computer Science*, vol. 16597, Singapore: Springer Nature Singapore, 2026, pp. 519–531. doi: 10.1007/978-981-92-1300-9_41.
- [32] T. C. Hesterberg, "What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum," *Am. Stat.*, vol. 69, no. 4, pp. 371–386, Oct. 2015, doi: 10.1080/00031305.2015.1089789.
- [33] J. Carpenter and J. Bithell, "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians," *Stat. Med.*, vol. 19, no. 9, pp. 1141–1164, May 2000, doi: 10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F.
- [34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.