Comparative Study of CNN Backbones for Hemoglobin Estimation from Fingernail Images

Rudokaite, J.[1], Ong, L.L.[1], van Eijkelenburg, J[1], van den Hurk, K[2], and Huis in 't Veld, E[1]

[1] Department of Cognitive Sciences and Artificial Intelligence, Tilburg University, Warandelaan 2, 5037 AB, Tilburg, the Netherlands
j.rudokaite@tilburguniversity.edu

[2] Donor Health, Sanquin, Sanquin Bloedvoorziening, Plesmanlaan 125, 1066 CX Amsterdam, the Netherlands

Abstract. Non-invasive hemoglobin (Hb) screening from smartphone images could reduce reliance on finger-prick tests in blood donation and outpatient clinics. While previous studies have shown feasibility on small datasets, few have systematically examined how methodological choices affect performance. We compared four frozen convolutional backbones—ConvNeXt-Tiny, ResNet-50, EfficientNetV2-S, and MobileNet V3 Large—using fingernail crops under a shared pipeline. Deep embeddings and color features were extracted and combined through early and late fusion, evaluated with 5-fold stratified cross-validation. The best-performing model, ConvNeXt-Tiny, achieved an out-of-fold MAE of 0.6 mmol/L. Robustness analyses showed stable performance under illumination changes (MAE shift < 0.02 mmol/L), minimal subgroup differences by sex or donation type (0.04 mmol/L), and well-calibrated uncertainty estimates (interval width 1.0–1.2 mmol/L, 0.9 mmol/L with normalization). These findings demonstrate that Hb-related information can be extracted from self-captured fingernail images using lightweight, reproducible pipelines. While current performance does not yet meet clinical standards, the results support the feasibility of image-based Hb estimation and motivate larger, multi-site studies for clinical validation.

Keywords: Hemoglobin measurement · Machine learning · Fingernails

1 Introduction

Hemoglobin (Hb) is essential for maintaining physiological balance by delivering oxygen to vital organs such as the brain, heart, and kidneys. When hemoglobin levels fall at or below healthy thresholds (men ≤ 8.1 mmol/L, women ≤ 7.5 mmol/L), organ function may be affected, particularly in individuals with underlying health issues or nutritional deficiencies [1]. Given hemoglobin's vital physiological role, disruptions in its levels are associated with a wide range of acute and chronic conditions, including anemia, cardiovascular disease, stroke, and pregnancy-related complications [2].

Anemia, a condition described by insufficient hemoglobin levels, affects almost 30% of people worldwide [3]. It is especially prevalent in children and individuals in low- and middle-income countries, where nutritional deficiencies and chronic blood loss are common [3]. Even in the United States, 5–9% of the population is anemic, with a significant subset suffering from moderate to severe forms [4, 5]. In addition to public health issues, the financial burden of related disorders is substantial. For example, hemoglobin monitoring is crucial for managing conditions such as anemia or sickle cell disease, where regular testing is needed to prevent complications [6].

Despite the clinical importance of regular hemoglobin monitoring, traditional Hb testing methods remain invasive, requiring venous or capillary blood draws, trained personnel, and specialized laboratory equipment or devices [7, 8]. These factors not only limit scalability but also introduce other barriers such as pain [9, 10], needle phobia [11], and prolonged diagnostic processing times. In resource-limited settings, or for large-scale population screenings, these limitations severely restrict access to timely diagnosis and care [12].

This is particularly critical in the context of blood donation centers and transfusion services, where pre-donation hemoglobin screening is essential to safeguard both donor health and the quality of the blood supply, but ideally, it should be fast, accurate, and painless [13, 14]. A non-invasive, rapid alternative to finger-pricks would enhance the donor experience and could improve retention rates.

Emerging tools, especially those integrated in everyday technologies such as smartphones, offer the potential for scalable, user-friendly, and affordable solutions that can reach even the most remote communities. In addition to reducing financial and logistical burdens, non-invasive techniques also minimize biohazard waste and improve operational efficiency. However, for such tools to gain clinical acceptance, their accuracy must approach that of standard laboratory methods. Clinical hemoglobin measurement typically achieves measurement errors below 0.3 mmol/L and correlation coefficients $r > 0.98$ when compared to laboratory reference methods [7, 8]. Non-invasive or optical devices are generally considered clinically acceptable if their mean absolute error remains within ± 0.62 mmol/L of reference values [15, 16]. These thresholds provide a practical benchmark for evaluating the performance and potential deployment readiness of image-based hemoglobin estimation methods.

2 Methods and Materials

2.1 Data collection

The study was completed at the national blood bank (Sanquin) location in Amersfoort, the Netherlands. The participants received a QR code with an information letter about the study. Each participant was asked to submit photos of their fingernails (see Fig. 1 for an example of the instructions) and report hemoglobin (Hb) measurements performed by a physician at the donation site.

For every subject, we manually cropped the fingernail regions from the submitted images and selected 3 to 5 image crops per participant (some crops had to be excluded due to poor quality or missing images). These crops were grouped into a single sample (bag), each associated with one Hb value. Any participants who had missing Hb values were excluded from the sample. The study was reviewed and approved by Sanquin's medical board. All participants provided written informed consent prior to participation.

Fig. 1: The images show the instructional example given to participants on how to photograph their fingernails. Participants were asked to take one photo of their thumb and one of the remaining four fingers, for both hands.

2.2 Participants and hemoglobin distribution

The study included 159 participants (88 males, 63 females, 8 others) with an average age of 44.3 years (SD = 13.1, age range: 18–75 years). Most participants donated either whole blood (N = 100) or plasma (N = 51). Hemoglobin (Hb) values ranged from 6.5 to 11.2 mmol/L (median = 8.5, mean = 8.62, SD = 0.79; Figure 2). The mean Hb levels for female participants were 8.09 mmol/L (min = 6.5, max = 9.5) and for male participants 9.08 mmol/L (min = 8.1, max = 11.2).

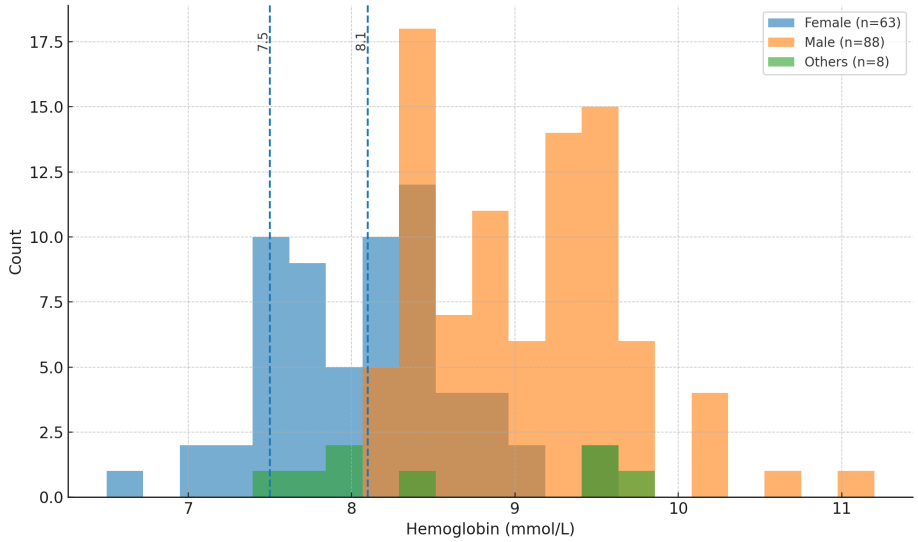

Fig. 2: Distribution of hemoglobin values (N = 159). Colors represent donor sex, and dashed lines indicate hemoglobin donation thresholds. Donors with hemoglobin values below these thresholds are not eligible to donate. Note: hemoglobin is measured in mmol/L at Sanquin, which is the standard clinical unit used in the Netherlands. Donation thresholds are higher for whole blood (men = 8.4 mmol/L, women = 7.8 mmol/L) than for plasma (men = 8.1 mmol/L, women = 7.5 mmol/L).

2.3 Image processing

Images were aligned using orientation tags (EXIF) and center-cropped around the nail bed. To mitigate illumination variability, we applied computational color constancy using the generalized Gray-World model (Minkowski p-norm, p=6) [26]. After illumination correction, images used to extract deep embeddings were resized to 224×224, then normalized with ImageNet statistics [27].

2.4 Feature Extraction

We constructed two main feature sets:

1. **Deep visual embeddings.** We extracted fixed-length representations using four frozen convolutional backbones from the timm library [28]: ConvNeXt-Tiny [29], ResNet-50 [30], EfficientNetV2-S (pretrained on ImageNet-21k) [31], and MobileNetV3-Large-100 [32]. This set of backbones covers a spectrum of architectures, from the classical ResNet-50 baseline to modern compact CNNs (ConvNeXt, EfficientNetV2) and a lightweight mobile-optimized model (MobileNetV3), enabling us to assess whether recent architectural advances improve Hb estimation. Each network was initialized with ImageNet-1k pretrained weights [27] and kept frozen. Freezing the convolutional layers

ensured consistent feature extraction across backbones and prevented overfitting given the limited dataset size. Although these models were originally trained for general image classification, the early convolutional layers are known to capture low-level color and texture representations that remain informative for subtle chromatic variations relevant to hemoglobin estimation. This design choice allows for reproducible comparisons while isolating architectural differences without the confound of task-specific fine-tuning. The final classification head was removed, and global average pooling was applied to the last convolutional feature maps. Embedding dimensionality was 768, 2048, and 1280 for ConvNeXt-Tiny, ResNet-50, and both EfficientNetV2-S and MobileNetV3-Large-100, respectively. For each participant, up to 3-5 crop embeddings were aggregated by element-wise mean and standard deviation, yielding 2×d–dimensional participant-level vectors. This design captures both central tendency and variability across crops, improving robustness to image quality variation.

2. **Color features.** From each crop we computed a set of descriptors capturing pigmentation and illumination: RGB channel means and standard deviations, chromaticity statistics for $r/(r+g+b)$, $g/(r+g+b)$, and $b/(r+g+b)$, grayscale mean and standard deviation, grayscale 10th/50th/90th percentiles, and clipped channel ratios (red/green, red/blue). For each participant, descriptors were aggregated across crops by both mean and median, producing fixed-length vectors of approximately 38 features.

We evaluated three feature sets: (1) color-only; (2) deep-only; and (3) combined color+deep. In the multimodal setup, we compared early fusion (feature concatenation) with late fusion (stacked regression on backbone-specific predictions). Additionally, in *fusion experiments*, we combined predictions from multiple backbones using Ridge regression [37] as a stacking model.

2.5 Regression models

We treated hemoglobin estimation as a participant–level regression task. Input feature vectors (color, deep embeddings, or fused) were standardized to zero mean and unit variance based on the training set, and the same transformation was applied to the validation set. This ensures all features are on a comparable scale for the following regression models used in this study:

1. **Partial Least Squares (PLS) [33].** Because the deep feature vectors are high-dimensional and exhibit substantial inter-feature correlation, we used PLS regression, which is designed for problems with many, potentially collinear predictors and limited samples [33, 34]. The number of latent components was tuned by 5–fold inner cross–validation, searching from 1 up to 20. The final PLS model was refit on the training split with the selected number of components.
2. **Support Vector Regression (SVR, RBF kernel) [35].** To capture nonlinear effects, we trained SVR models with radial basis function kernels.

Hyperparameters were selected by 3–fold inner cross–validation over $C \in \{0.5, 1, 2, 4\}$ and $\gamma \in \{\text{scale}, 10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$.

3. **Calibrated blending.** Each regressor's predictions were first corrected for bias using isotonic regression [36], ensuring they matched the true Hb scale. The two calibrated outputs were then combined into a weighted average, with the weight chosen to minimize error on the training folds. This blending may produce more accurate and stable predictions than either model alone.

2.6 Model Training and Evaluation

We evaluated all models with participant-level 5-fold stratified cross-validation [38, 39], ensuring that all images from a given donor stayed in the same fold. To preserve the hemoglobin distribution, Hb values were stratified into 5 quantile bins before splitting. In each outer fold, 80% of participants formed the training split and 20% the validation split. With $N = 159$ participants, this corresponds to about 127 training and 32 validation donors per fold (exact counts varied slightly by stratification). Within each outer training split, model hyperparameters were tuned by an inner cross-validation loop, after which the final models were refit on the full outer training split and evaluated on the held-out validation fold.

For each feature set, we generated out–of–fold (OOF) predictions across all participants. PLS and SVR predictions were further calibrated using isotonic regression [36] applied only within each training fold to avoid data leakage from the test folds, and then created a weighted average of the two models' predictions to minimize mean absolute error (MAE) on the training folds. This yielded the final blended OOF predictions for evaluation.

Model performance was summarized on the concatenated OOF predictions across all folds using the following metrics:

1. **Mean Absolute Error (MAE):** the average absolute difference between predicted and true Hb (mmol/L). For example, an MAE of 1.0 means that on average the model's predictions are 1 mmol/L away from the actual value.
2. **Root Mean Squared Error (RMSE):** the square root of the mean squared error, which gives more weight to larger deviations. For example, an RMSE of 1.3 means that typical prediction errors are around 1.3 mmol/L, with large errors contributing more strongly.
3. **Calibration error and slope/intercept:** we assessed whether the predicted Hb values were systematically biased. Expected calibration error (ECE) measures how closely the average predictions match the observed values. A slope of 1 and intercept of 0 indicate perfect calibration. For example, if the slope is < 1, extreme predictions are too conservative; if the intercept is positive, the model tends to overestimate Hb levels. Model outputs were calibrated using isotonic regression with sample-size weighting, which enforces a monotonic mapping between predicted and true values. For visualization, we additionally fitted simple linear regression lines to illustrate global calibration trends, but all reported metrics (e.g., MAE, AUROC) are based on isotonic-calibrated predictions.

4. **AUROC at clinical thresholds:** the ability of the model to rank participants below vs. above clinically relevant cut–offs (i.e. females 7.8 mmol/L, males 8.4 mmol/L). AUROC uses the continuous regression outputs as decision scores, without fixing a threshold. An AUROC of 0.5 means the model performs no better than random guessing, while 1.0 means perfect discrimination. For example, an AUROC of 0.80 at the 7.8 mmol/L threshold means that in 80% of randomly chosen donor pairs (one below and one above the cut–off) the model correctly ranks the lower–Hb participant as having the lower predicted Hb.

In addition, we further conducted analyses to evaluate prediction robustness and reliability:

1. **Heteroscedastic residual spread:** whether prediction errors varied systematically across the Hb range.
2. **Prediction intervals:** confidence prediction intervals providing coverage guarantees (e.g. 90% of true Hb within the predicted bounds).
3. **Subgroup robustness:** error metrics stratified by gender (male/female) and donation type (blood/plasma).
4. **Photometric stress tests:** testing robustness both to different illumination corrections (none vs. Shades–of–Gray) and to simulated brightness/contrast jitter (±20%).

3 Results

3.1 Hemoglobin estimation from fingernail images

Across all experiments, the lowest average absolute error was consistently around ≈0.60 mmol/L. Although sex was not included as an explicit input covariate, we evaluated model discrimination using the sex-specific donation thresholds (7.8 mmol/L for females, 8.4 mmol/L for males) to reflect clinical practice. Discrimination performance was modest but consistent, with AUROC values up to 0.646 at 8.0 mmol/L and 0.645 at 10.0 mmol/L for the late fusion model, while deep-only models performed slightly lower. Although current performance does not yet meet clinical accuracy standards, these results are promising given the small dataset size and demonstrate stability across folds.

Models using deep embeddings outperformed those relying solely on color features (Table 1, Table 2). Among fusion strategies, early concatenation and late fusion achieved comparable error metrics, but late fusion consistently showed stronger discrimination at clinically relevant thresholds. Per-fold analyses confirmed this hierarchy: in nearly every fold, deep or fused models yielded lower MAE values than color-only baselines.

Table 1: OOF performance aggregated across all folds (MAE, RMSE, and AUROC at Hb thresholds). Best overall values are **bold**.

Model	MAE	RMSE	$\text{AUROC}_{\geq 8.0}$	$\text{AUROC}_{\geq 9.0}$	$\text{AUROC}_{\geq 10.0}$
Color features*	0.639	0.794	0.514	0.451	0.376
Deep embeddings**	**0.606**	**0.765**	0.609	0.575	0.462
Concat (Early fusion)***	0.610	0.769	0.610	0.573	0.479
Blend (Late fusion)****	0.614	0.767	**0.646**	**0.589**	**0.645**

* Color features: hand-crafted color statistics from fingernail regions.
** Deep embeddings: frozen ConvNeXt-Tiny backbone pre-trained on ImageNet.
*** Concat (Early fusion): concatenation of color features and ConvNeXt-Tiny embeddings before regression.
**** Blend (Late fusion): ridge stacking of predictions from ConvNeXt-Tiny, ResNet-50, EfficientNetV2-S, and MobileNetV3-Large.

Overall, these results indicate that frozen convolutional embeddings carry the primary predictive signal, and simple fusion schemes can add modest but consistent performance improvements. Stacking the isotonic-calibrated PLS and SVR predictions improved discrimination at clinical thresholds with comparable error, specifically, AUROC at 8.0 mmol/L increased from 0.609 to 0.646, while MAE differed by 0.01 mmol/L (Table 1). This aligns with prior evidence that stacking diverse learners enhances generalization. [40–43]

Table 2: Per-fold performance (MAE, RMSE, and AUROC at Hb≥9.0 mmol/L). "Deep" uses ConvNeXt-Tiny embeddings; "Concat" = color + ConvNeXt-Tiny (early fusion); "Blend" stacks all four backbones (late fusion). Best per-fold values are **bold**.

Fold	Color only*			Deep (ConvNeXt)**			Merged models					
							Concat (Early)***			Blend (Late)****		
	MAE	RMSE	AUROC	MAE	RMSE	AUROC	MAE	RMSE	AUROC	MAE	RMSE	AUROC
01	0.648	0.761	0.514	0.593	0.690	0.648	0.597	0.695	0.664	**0.587**	**0.688**	**0.676**
02	**0.565**	**0.688**	**0.483**	0.616	0.778	0.362	0.627	0.788	0.394	0.623	0.776	0.411
03	0.700	0.854	0.565	0.629	0.785	**0.702**	**0.624**	**0.780**	0.690	0.703	0.845	0.577
04	0.579	0.753	0.455	0.546	**0.701**	**0.727**	**0.545**	0.704	0.673	0.550	0.706	0.718
05	0.705	0.899	0.448	0.650	0.864	0.583	0.656	0.869	0.567	**0.605**	**0.811**	**0.648**
Average	0.639	0.791	0.493	0.607	0.764	0.604	0.610	0.767	0.598	**0.614**	**0.765**	**0.646**

* Hand-crafted color statistics. ** Frozen ConvNeXt–Tiny embeddings (ImageNet).
*** Feature concatenation: color and ConvNeXt–Tiny. **** Ridge stacking of predictions from ConvNeXt–Tiny, ResNet-50, EfficientNetV2-S, and MobileNetV3-Large.

Error analysis showed no large systematic bias across the Hb range (Fig. 3; error analyses for all backbones are shown in Appendix, Fig. 5 and Fig. 6), but indicate a tendency toward underestimation at higher and overestimation at lower Hb levels.

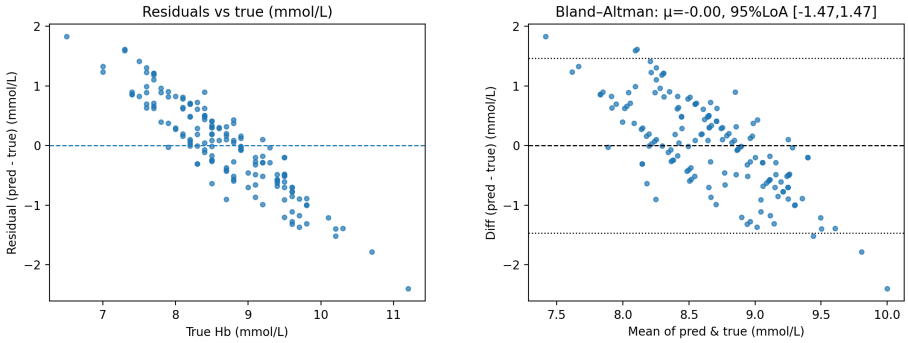

Fig. 3: Out-of-fold error analysis for the blended model (best model; ConvNeXt-Tiny backbone): (a) residuals vs. true Hb (dashed line = zero error); (b) Bland–Altman plot (dashed = mean bias, dotted = 95% limits).

3.2 Backbone comparison

We compared four frozen convolutional backbones using the same feature extraction and regression pipeline. Results are summarized in Table 3. ConvNeXt-Tiny achieved the best overall performance (MAE = 0.603 mmol/L; RMSE = 0.748), followed by EfficientNetV2-S (MAE = 0.613; RMSE = 0.760).

In practice, this indicates that modern compact architectures (ConvNeXt, EfficientNet, MobileNet) provide state-of-the-art performance while remaining computationally efficient, an important factor for deployment on edge devices such as smartphones.

Table 3: Backbone-level OOF performance (MAE, RMSE, and AUROC at Hb $\geq$ 8.4 mmol/L). Best values are in **bold**.

Backbone	MAE	RMSE	$\text{AUROC}_{\geq 8.4}$
ConvNeXt-Tiny	**0.603**	**0.748**	**0.661**
ResNet-50	0.644	0.799	0.496
EfficientNetV2-S	0.613	0.760	0.637
MobileNetV3-Large	0.630	0.794	0.574

3.3 Calibration analysis

We assessed calibration using linear recalibration (slope, intercept) and expected calibration error (ECE) (Table 4; full plots in Appendix, Fig.7). ConvNeXt-Tiny and EfficientNetV2-S achieved the best calibration, with slopes of 0.76 and moderate intercept shifts (+2.1 mmol/L), indicating a consistent but moderate overestimation of Hb. MobileNetV3-Large showed weaker calibration (slope 0.38, intercept +5.4), while ResNet-50 performed poorly (slope –0.09, intercept

+9.5). ECE values ranged from 0.12–0.19 across models, suggesting moderate miscalibration overall, with ConvNeXt and EfficientNet performing best.

Table 4: Calibration results across backbones. Slope and intercept are from a linear fit of predicted vs. true Hb. Best values are in **bold**.

Backbone	Slope	Intercept	ECE
ConvNeXt-Tiny	0.76	2.12	0.15
ResNet-50	–0.09	9.45	0.19
EfficientNetV2-S	**0.76**	**2.11**	**0.12**
MobileNetV3-Large	0.38	5.38	0.18

3.4 Learning curves

Learning curves were constructed by progressively increasing the number of training participants from 20 to 159 and computing out-of-fold error at each subsample size (Fig. 4). Both MAE and RMSE decreased as more participants were included, with the largest gains between 20 and 60 participants. Beyond 80 participants, improvements leveled off, suggesting diminishing returns at the current scale. Extrapolation from these curves suggests that substantially larger cohorts would be required for further gains, although doubling the sample size could still reduce errors by an estimated 10–15%.

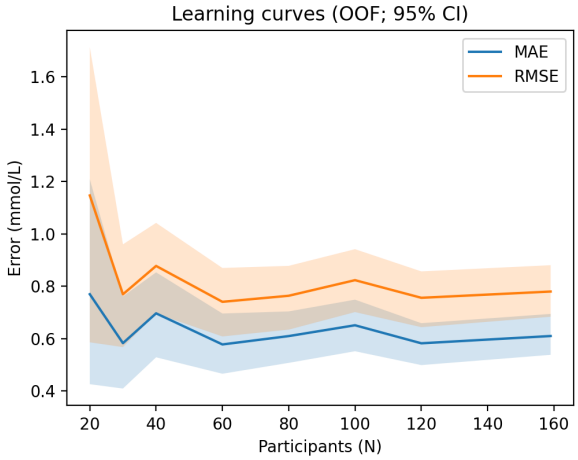

Fig. 4: Learning curves showing out-of-fold MAE and RMSE with 95% confidence intervals as a function of dataset size (ConvNeXt-Tiny backbone).

3.5 Subgroup robustness

We evaluated model performance across gender and donation type (Table 5). Errors were similar between males and females (MAE difference = 0.04 mmol/L) and across whole-blood vs. plasma donors (MAE difference = 0.03 mmol/L). Since there were no major differences between subgroups, the model seems to work consistently across donor groups.

Table 5: Subgroup robustness analysis (ConvNeXt-Tiny backbone, deep embeddings). Best values are in **bold**.

Subgroup	MAE	RMSE
Male (N=88)	**0.602**	**0.745**
Female (N=63)	0.642	0.770
Blood donors (N=100)	0.610	0.752
Plasma donors (N=51)	0.637	0.761

3.6 Photometric stress tests

To simulate uncontrolled smartphone conditions, we applied both alternative illumination correction and synthetic brightness/contrast perturbations (Table 6, Appendix Fig. 8). Results showed negligible differences compared to baseline: shaded illumination yielded MAE = 0.610 mmol/L vs. 0.612 at baseline, and ±20% brightness/contrast jitter produced shifts of < 0.01 mmol/L. This indicates robustness to modest changes in lighting.

Table 6: Performance under photometric stress tests (ConvNeXt-Tiny backbone). Best values are in **bold**.

Condition	MAE	RMSE
Baseline	0.612	0.773
Shaded illumination	**0.610**	**0.769**
Brightness/contrast ±20%	0.615	0.776

3.7 Heteroscedasticity, confidence intervals, and label noise

Residual analysis revealed slightly higher error variance at mid-range Hb values (8–9 mmol/L) but no strong heteroscedastic trend (see Appendix, Fig. 9). The uncertainty intervals contained the true value 90% of the time with mean interval width of 1.2 mmol/L; normalized uncertainty intervals reduced average width to 0.9 mmol/L.

Because Hb measurements themselves are not perfectly precise, we tested how sensitive the models are to noisy labels. We added Gaussian noise with standard deviations $\sigma = 0.1$–0.3 mmol/L to the reference Hb values. Model performance

was only slightly affected (MAE increased from 0.612 to 0.623 mmol/L; RMSE from 0.773 to 0.788), suggesting that the models are robust to moderate levels of measurement uncertainty.

In addition, we evaluated potential rule-out utility at the clinical threshold of 7.8 mmol/L (Appendix, Fig. 10). When applying a conservative safety margin of 0.5 mmol/L, the model correctly identified nearly 70% of donors as being safely above the threshold, while only 5–10% of true low-Hb cases would have been incorrectly classified as safe.

In conclusion, these analyses confirm that errors remain small and stable across the Hb range, that the model can provide calibrated uncertainty estimates, and that results are not overly sensitive to label noise.

4 Discussion

This study aimed to examine how different methodological choices such as backbone architecture, feature fusion strategies, calibration, and robustness testing affect the performance of hemoglobin estimation from fingernail images. Across all folds, models based on frozen image embeddings consistently outperformed colour-only features, and both early and late multimodal fusion offered further improvements. The ConvNeXt-Tiny backbone achieved the lowest overall error (OOF MAE = 0.603 mmol/L, RMSE = 0.748 mmol/L), whereas the late-fusion ensemble had slightly higher error (MAE = 0.614 mmol/L; RMSE = 0.767 mmol/L) but provided the strongest discrimination at higher clinical thresholds.

When compared to earlier work, our error rates are higher. For example, Yılmaz et al. [19] reported RMSE = 0.35 mmol/L (converted from 0.56 g/dL) using multimodal inputs (fingernail images combined with demographics such as age or sex), while Das et al. [21] achieved RMSE = 0.39 mmol/L (converted from 0.63 g/dL) using nail images alone. Our best RMSE of 0.748–0.767 mmol/L ($\approx$ 1.21–1.24 g/dL) is therefore higher, but our study used a smaller dataset (N = 159) and relied solely on frozen backbones without fine-tuning or personalized calibration. For instance, Mannino et al. [20] demonstrated improved performance through personalization and continuous self-monitoring, achieving clinical readiness only with thousands of participants and adaptive calibration.

These comparisons suggest that while our models are not yet at clinical accuracy, they demonstrate feasibility under more constrained and reproducible settings. This shows that hemoglobin-related information can be extracted from uncontrolled smartphone images, even without demographic variables or standardized capture conditions. Importantly, our focus is on overall hemoglobin monitoring rather than classification at clinical thresholds. The models were trained and evaluated as continuous regressors, and threshold-based AUROC values are included only to provide a complementary perspective. Because the dataset consists predominantly of healthy donors with a narrow hemoglobin range, limited discriminative performance at predefined cut-offs is somewhat expected. The goal is not to achieve clinical utility at this stage, but to assess

whether methodological choices meaningfully affect prediction quality and robustness under realistic conditions.

Therefore, beyond absolute error, our analyses focused on model reliability and design trade-offs. Backbone comparison showed that modern compact architectures (ConvNeXt-Tiny, EfficientNetV2-S) performed best. In contrast, many earlier studies achieved lower error rates by relying on more resource-intensive strategies. For example, Yılmaz et al. combined deep image features with demographic data (age, height, weight, sex) and reported lower error (RMSE = 0.35 mmol/L) [19], while Navarro-Cabrera et al. used fingernail images with a DenseNet-169 backbone, achieving moderate classification performance (accuracy = 0.70, AUC = 0.74) in a controlled university sample [18]. These examples highlight that combining image data with personalization or richer models can improve accuracy but might also require higher complexity, larger datasets, or controlled imaging environments. Thus, our choice of frozen deep embeddings without additional metadata is deliberately minimalist and reproducible. Despite slightly higher error compared to multimodal systems, our pipeline shows that using a self-captured fingernail image can still extract a substantial hemoglobin signal. In other words, we traded some accuracy for simplicity and broader applicability. While fine-tuning models could potentially capture subtle chromatic changes associated with hemoglobin concentration even better, our goal was to establish a controlled baseline to evaluate backbone and fusion effects independently of training, ensure reproducibility, and avoid overfitting due to the small dataset.

Looking at calibration, our models showed moderate miscalibration and sometimes underestimated hemoglobin at higher levels and overestimated at lower levels. This is in line with other studies, where recalibration was also needed before clinical use (e.g., Mannino et al. who performed post-deployment recalibration) [20]. While frozen embeddings capture the main visual signal, aligning predictions with clinical thresholds will be essential for safe application.

We also tested how robust the models were under different conditions, and the results suggest they are fairly stable. Changes in illumination had almost no effect on accuracy, and performance was similar across subgroups such as sex or donation type. The confidence intervals captured the true Hb values at the expected rate and were reasonably narrow, meaning the uncertainty estimates are informative. When we added simulated noise to the reference Hb values, model errors increased only slightly, indicating that the predictions are not highly sensitive to small variations in the ground-truth measurements. Learning curves indicated diminishing returns beyond 80 participants, with accuracy stabilizing around 0.61 mmol/L MAE; however, extrapolation suggests that larger datasets would allow further improvements.

Most earlier studies focus narrowly on accuracy metrics without testing robustness. For instance, Yılmaz et al. [19] and Das et al. [21] reported promising RMSE values but did not assess sensitivity to lighting changes, subgroup differences, or label noise. Even the work by Mannino et al. [20], which demonstrates a

personalized hemoglobin app, reported post-deployment calibration rather than systematic stress testing of model inputs.

Taken together, these findings suggest that fingernail imaging could provide a reliable, low-cost method for pre-screening Hb, with robustness to device, lighting, and subgroup variation. However, this study has several limitations. First, the cohort size was modest and from a single site; second, capture conditions were heterogeneous but not systematically evaluated; third, manual cropping may have introduced selection bias; and fourth, personalization or demographic features were not included, even though prior work shows these can reduce error in larger studies [20]. Finally, all performance estimates are out-of-fold rather than from an external test set.

We are currently collecting larger, multi-site datasets to address these limitations. Future work will focus on automated nail detection and quality control, improved calibration strategies, and fairness analysis across broader demographic and device subgroups. Additional robustness tests could also include artifacts more common in real-world mobile use, such as motion blur, image compression, or low-light capture, and fairness evaluation across skin tone, age, and device type. Limited fine-tuning of convolutional backbones could be explored to quantify trade-offs between task-specific adaptation and computational efficiency, while minimal multimodal extensions (e.g., adding age or sex) may help clarify the added value of demographic information.

From a clinical safety perspective, the consistent overestimation at low Hb levels may risk missing borderline cases if not properly calibrated. Incorporating threshold-aware calibration or conservative safety margins could mitigate this in deployment. Finally, integrating explainability methods such as Grad-CAM or feature attribution could help visualize which nail regions contribute most to predictions, improving interpretability and clinical trust. With such extensions, image-based Hb estimation could evolve into a practical and scalable tool for donor pre-screening and outpatient monitoring.

Acknowledgments. This study was funded by Tilburg University, Growth Project (DSFS 2023020) and NWO SGW Open Competitie 406.XS.01.109. We would also like to thank all blood donors who participated in the study and the Sanquin staff for their support.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Data Availability The dataset analyzed during the current study are not publicly available due to donor privacy but are available from the corresponding author upon reasonable request.

Bibliography

[1] van Swelm, R.P.L., Wetzels, J.F.M., Swinkels, D.W.: The multifaceted role of iron in renal health and disease. *Nat. Rev. Nephrol.* **16**(2), 77–98 (2020). https://doi.org/10.1038/s41581-019-0216-7

[2] Ataga, K.I., Gordeuk, V.R., Agodoa, I., Colby, J.A., Gittings, K., Allen, I.E.: Low hemoglobin increases risk for cerebrovascular disease, kidney disease, pulmonary vasculopathy, and mortality in sickle cell disease: A systematic review and meta-analysis. *PLoS One* **15**(4), e0229959 (2020). https://doi.org/10.1371/journal.pone.0229959

[3] Gardner, W., Kassebaum, N.: Global, regional, and national prevalence of anemia and its causes in 204 countries and territories, 1990–2019. *Current Developments in Nutrition* Abstract presented at *Nutrition 2020 Live Online*.

[4] Le, C.H.H.: The prevalence of anemia and moderate-severe anemia in the US population (NHANES 2003–2012). *PLoS One* **11**(11), e0166635 (2016). https://doi.org/10.1371/journal.pone.0166635

[5] Williams, A.M., Ansai, N., Ahluwalia, N., Nguyen, D.T.: Anemia prevalence: United States, August 2021–August 2023. *NCHS Data Brief* No. 490, National Center for Health Statistics (2024).

[6] Krishnamurti, L., Ross, D., Sinha, C., Leong, T., Bakshi, N., Mittal, N., et al.: Comparative effectiveness of a web-based patient decision aid for therapeutic options for sickle cell disease: Randomized controlled trial. *J. Med. Internet Res.* **21**(12), e14462 (2019). https://doi.org/10.2196/14462

[7] Marn, H., Critchley, J.A.: Accuracy of the WHO haemoglobin colour scale for the diagnosis of anaemia in primary health care settings in low-income countries: A systematic review and meta-analysis. *Lancet Glob. Health* **4**(4), e251–e265 (2016). https://doi.org/10.1016/S2214-109X(16)00048-5

[8] Whitehead, R.D., Mei, Z., Mapango, C., Jefferds, M.E.D.: Methods and analyzers for hemoglobin measurement in clinical laboratories and field settings. *Ann. N. Y. Acad. Sci.* **1450**(1), 147–171 (2019). https://doi.org/10.1111/nyas.14124

[9] Bisogni, S., Dini, C., Olivini, N., Ciofi, D., Giusti, F., Caprilli, S., Gonzalez Lopez, J.R., Festini, F.: Perception of venipuncture pain in children suffering from chronic diseases. *BMC Res. Notes* **7**, 735 (2014). https://doi.org/10.1186/1756-0500-7-735

[10] Courtois, E., Cimerman, P., Dubuche, V., Goiset, M.F., Orfèvre, C., Lagarde, A., Sgaggero, B., et al.: The burden of venipuncture pain in neonatal intensive care units: EPIPPAIN 2, a prospective observational study. *Int. J. Nurs. Stud.* **57**, 48–59 (2016). https://doi.org/10.1016/j.ijnurstu.2016.01.016

[11] McLenon, J., Rogers, M.A.: The fear of needles: A systematic review and meta-analysis. *J. Adv. Nurs.* **75**(1), 30–42 (2019). https://doi.org/10.1111/jan.13818

[12] Ahn, H., Lenet, T., Gilbert, R.W.D., Mallick, R., Shaw, J.L.V., Fergusson, D.A., McIsaac, D.I., Martel, G.: Accuracy of point-of-care testing devices for haemoglobin in the operating room: Meta-analysis. *BJS Open* **8**(1), zrad148 (2024). https://doi.org/10.1093/bjsopen/zrad148

[13] De Clippel, D., Van Heddegem, L., Vandewalle, G., Vandekerckhove, P., Compernolle, V.: Hemoglobin screening in blood donors: A prospective study assessing the value of an invasive and a noninvasive point-of-care device for donor safety. *Transfusion* **57**(4), 938–945 (2017). https://doi.org/10.1111/trf.14013

[14] Sümnig, A., Hron, G., Westphal, A., Petersmann, A., Kohlmann, T., Greinacher, A., Thiele, T.: The impact of noninvasive, capillary, and venous hemoglobin screening on donor deferrals and the hemoglobin content of red blood cell concentrates: A prospective study. *Transfusion* **55**(12), 2847–2854 (2015). https://doi.org/10.1111/trf.13252

[15] World Health Organization.: *Blood Donor Selection: Guidelines on Assessing Donor Suitability for Blood Donation.* World Health Organization, Geneva (2012)

[16] U.S. Food and Drug Administration.: *Guidance for Industry and FDA Staff: Noninvasive Blood Pressure and Hemoglobin Measurement Devices.* U.S. Department of Health and Human Services (2013). Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/noninvasive-blood-pressure-and-hemoglobin-measurement-devices

[17] Stoltzfus, R.J., Edward-Raj, A., Dreyfuss, M.L., Albonico, M., Montresor, A., Thapa, M.D., et al.: Clinical pallor is useful to detect severe anemia in populations where anemia is prevalent and severe. *J. Nutr.* **129**(9), 1675–1681 (1999). https://doi.org/10.1093/jn/129.9.1675

[18] Navarro-Cabrera, J.R., Valles-Coral, M.A., Farro-Roque, M.E., Reátegui-Lozano, N., Arévalo-Fasanando, L.: Machine vision model using nail images for non-invasive detection of iron deficiency anemia in university students. *Front. Big Data* **8**, 1557600 (2025). https://doi.org/10.3389/fdata.2025.1557600

[19] Yılmaz, H., Şen, B.: Deep learning-based noninvasive hemoglobin estimation using fingernail images and patient metadata. *IEEE Journal of Biomedical and Health Informatics* **26**(11), 5464–5474 (2022). https://doi.org/10.1109/JBHI.2022.3189760

[20] Mannino, R.G., Sullivan, J., Frediani, J.K., George, P., Whitson, J., Tumlin, J., Lyon, L.A., Tyburski, E.A., Lam, W.A.: Smartphone-based hemoglobin monitoring with personalized calibration. *Proceedings of the National Academy of Sciences of the United States of America* **122**(3), e2315453122 (2025). https://doi.org/10.1073/pnas.2424677122

[21] Das, A., Goswami, S., Jana, S.: Smartphone-based hemoglobin estimation using deep neural networks trained on fingernail images. *IEEE Access* **11**, 7065–7078 (2023). https://doi.org/10.1109/ACCESS.2023.3238672

[22] Amrutha, S., Joseph, P., Rani, R., Anitha, G., Gopinath, P.: Spectroscopic analysis and machine learning for non-invasive hemoglobin estimation from nail beds. *Sensors* **24**(6), 2390 (2024). https://doi.org/10.3390/s24062390

[23] Peksi, A., Özcan, T., Doğan, A., Aytaç, İ.: Non-invasive anemia detection using Naive Bayes classification of fingernail images. In: *Proceedings of the IEEE International Conference on Biomedical and Health Informatics (BHI 2021)*, pp. 275–279 (2021). https://doi.org/10.1109/BHI50953.2021.9508525

[24] Roychowdhury, S., Das, D., Das, S., Roy, M.: Color feature-based hemoglobin estimation from nail images. In: *Proceedings of the IEEE International Conference on Biomedical and Health Informatics (BHI 2017)*, pp. 335–338 (2017). https://doi.org/10.1109/BHI.2017.7897240

[25] Hasan, A., Al-Zubaidi, S., Hashim, M.M., AlRasheed, M.M., Hameed, S., Hashim, F.A.: Machine learning-based hemoglobin estimation using nail bed images captured by smartphones. *JMIR mHealth and uHealth* **9**(3), e24136 (2021). https://doi.org/10.2196/24136

[26] Finlayson, G.D., Hordley, S.D., Hubel, P.M.: Shades of gray and colour constancy. In: *Proceedings of the 12th Color Imaging Conference: Color Science and Engineering Systems, Technologies, and Applications*, pp. 37–41 (2004). Society for Imaging Science and Technology (IS&T). https://doi.org/10.2352/CIC.2004.12.1.037

[27] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848

[28] Wightman, R.: *PyTorch Image Models (timm)* (2019). https://github.com/rwightman/pytorch-image-models

[29] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pp. 11976–11986 (2022). https://doi.org/10.1109/CVPR52688.2022.01168

[30] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[31] Tan, M., Le, Q.V.: EfficientNetV2: Smaller models and faster training. In: *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, pp. 10096–10106. PMLR (2021).

[32] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for MobileNetV3. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, pp. 1314–1324 (2019). https://doi.org/10.1109/ICCV.2019.00148

[33] Wold, H.: Partial least squares. In: *Encyclopedia of Statistical Sciences*, Vol. 6, pp. 581–591. Wiley, New York (1984). https://doi.org/10.1002/0471667196.ess1914.pub2

[34] Rosipal, R., Kramer, N.: Overview and recent advances in partial least squares. In: Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (eds.) *Subspace, Latent Structure and Feature Selection.* Lecture Notes in Computer Science, vol. 3940. Springer, Berlin, Heidelberg (2006), pp. 34–51.

[35] Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. In: *Advances in Neural Information Processing Systems (NeurIPS 1996)*, Vol. 9, pp. 155–161 (1997). MIT Press.

[36] Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, pp. 694–699. ACM Press (2002). https://doi.org/10.1145/775047.775151

[37] Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67 (1970). https://doi.org/10.1080/00401706.1970.10488634

[38] Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 1995)*, Vol. 2, pp. 1137–1145. Morgan Kaufmann (1995).

[39] Sechidis, K., Tsoumakas, G., Vlahavas, I.: On the stratification of multilabel data. In: *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2011)*, pp. 145–158. Springer, Berlin, Heidelberg (2011).

[40] Wolpert, D.H.: Stacked generalization. *Neural Networks* **5**(2), 241–259 (1992). https://doi.org/10.1016/S0893-6080(05)80023-1

[41] Breiman, L.: Stacked regressions. *Machine Learning* **24**(1), 49–64 (1996). https://doi.org/10.1007/BF00117832

[42] Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A.: Ensemble selection from libraries of models. In: *Proceedings of the 21st International Conference on Machine Learning (ICML 2004)*, pp. 137–144. ACM Press (2004). https://doi.org/10.1145/1015330.1015432

[43] Zhou, Z.-H.: *Ensemble Methods: Foundations and Algorithms.* Chapman and Hall/CRC, Boca Raton, FL (2012). https://doi.org/10.1201/b12207

A Appendix

A.1 Bland–Altman plots by backbone

Figure 5 shows Bland–Altman plots for the four CNN backbones. All models exhibited small mean bias ($\mu \approx 0$) with 95% limits of agreement around ± 1.5 mmol/L. This indicates no systematic over- or under-estimation, though individual errors can exceed 1 mmol/L.

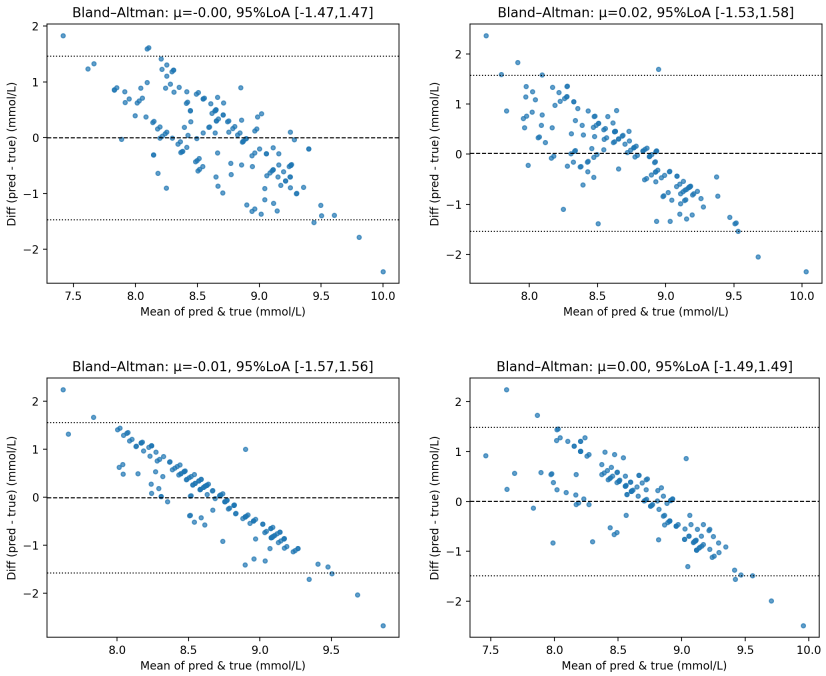

Fig. 5: Bland–Altman plots for ConvNeXt-Tiny, MobileNetV3-Large, ResNet-50, and EfficientNetV2-S backbones.

A.2 Residual plots by backbone

Figure 6 shows residual error plots. All models tended to underestimate high Hb values and overestimate low values.

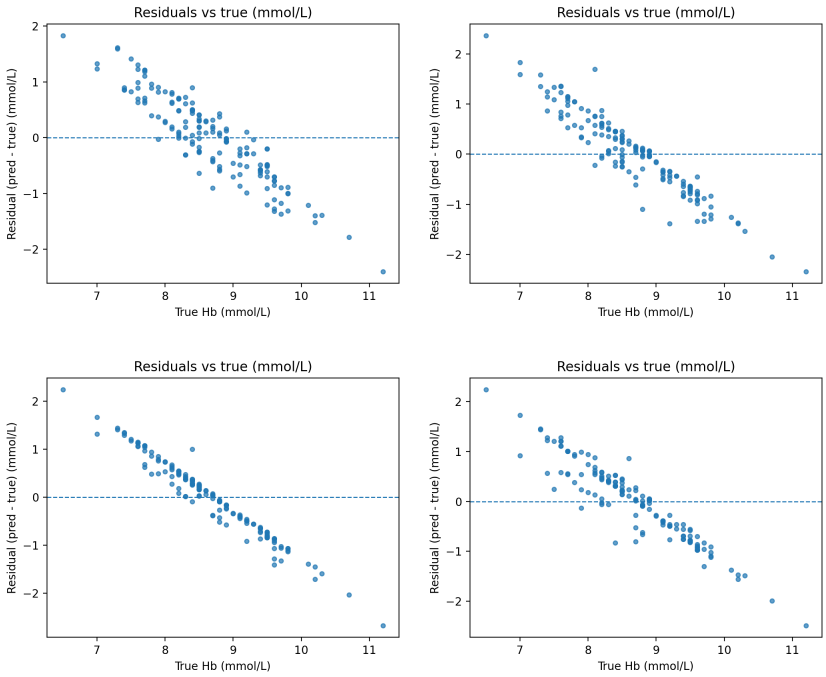

Fig. 6: Residual plots for ConvNeXt-Tiny, MobileNetV3-Large, ResNet-50, and EfficientNetV2-S.

A.3 Calibration curves

Figure 7 shows calibration plots. ConvNeXt-Tiny and EfficientNetV2-S were closest to the diagonal (slope ≈ 0.75), while MobileNetV3-Large and ResNet-50 showed stronger underestimation at high Hb (slope < 0.5). ECE ranged from 0.12–0.19, suggesting moderate miscalibration.

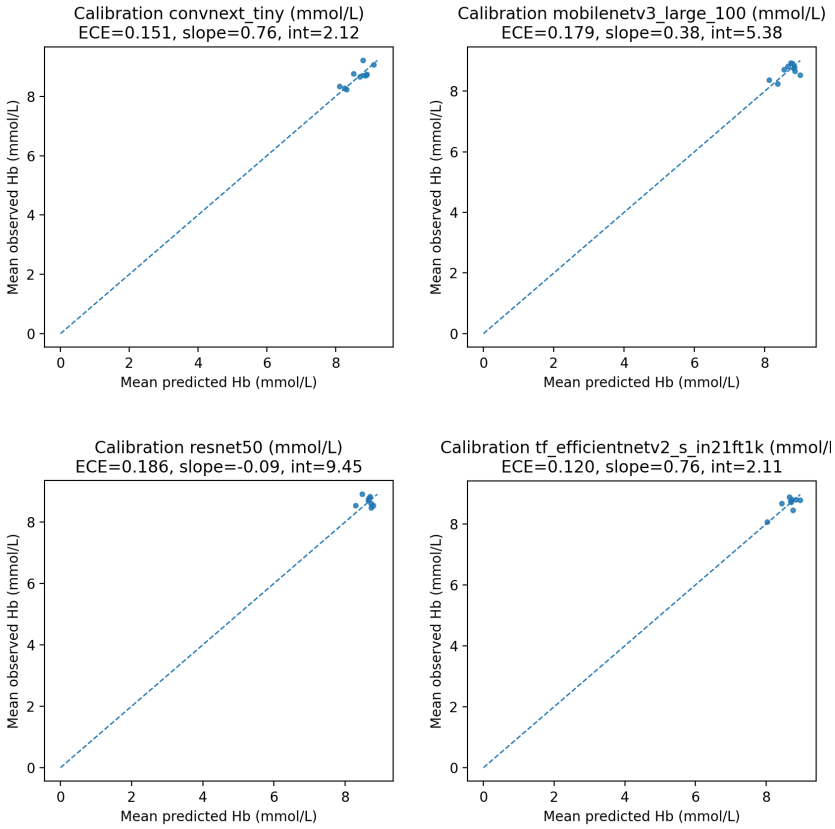

Fig. 7: Calibration analysis for the four backbones. Reported ECE, slope, and intercept quantify prediction bias.

A.4 Photometric stress test

Figure 8 shows robustness to illumination changes. Brightness/contrast perturbations of ±20% changed predictions by less than 0.2 mmol/L (typically < 0.1), suggesting resilience to moderate lighting variation.

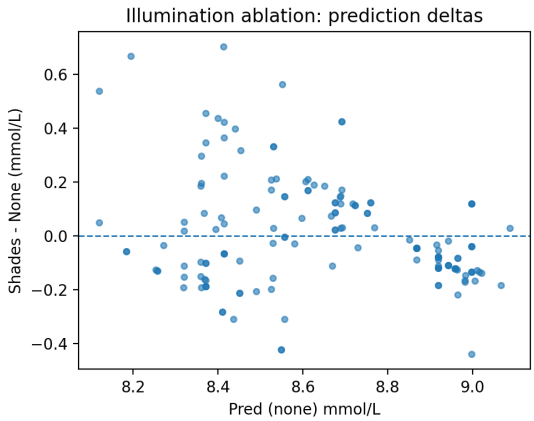

Fig. 8: Prediction deltas under ±20% illumination perturbations.

A.5 Residual spread and heteroscedasticity

Figure 9 shows the spread of residuals. Error variance was lowest near 8.5–9.0 mmol/L and increased at extremes, indicating modest heteroscedasticity.

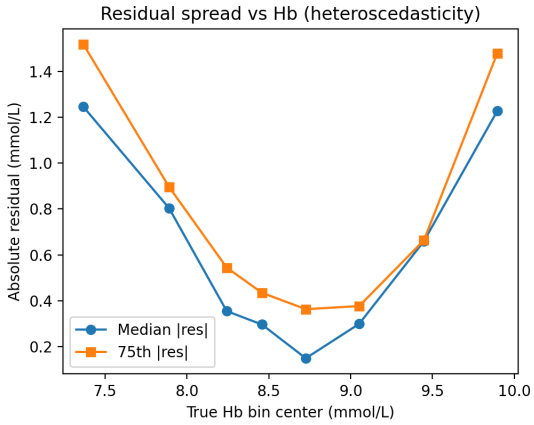

Fig. 9: Residual spread by true Hb (heteroscedasticity analysis).

A.6 Rule-out utility

Figure 10 shows rule-out analysis at the 7.8 mmol/L threshold. With a 0.5 mmol/L safety margin, ∼70% of donors could avoid invasive measurement while missing only 5–10% of true low-Hb cases.

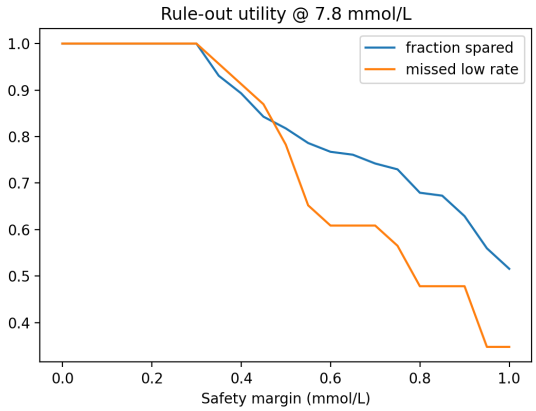

Fig. 10: Rule-out analysis at 7.8 mmol/L, showing trade-off between spared participants and missed low-Hb cases.