SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Multimodal Deep Learning for Automated Anemia Screening and Hemoglobin Estimation Using Conjunctival Images

[1]Dev Kumar, [2]Prof. Anamika Chaudhary, [3]Anushika Chaubey

[1]M.Tech Student, Department of Artificial Intelligence Noida Institute of Engineering and Technology (NIET), Greater Noida, Uttar Pradesh, India. E-mail: dev339880@gmail.com
[2]Department of Computer Science & Engineering (Data Science) Noida Institute of Engineering and Technology (NIET), Greater Noida, Uttar Pradesh, India. E-mail: anamika.cse@niet.co.in
[3]Software Engineer, Capgemini Airoli, Navi Mumbai, Maharashtra, India. E-mail: chaubeyanushika@gmail.com
Corresponding Author: Dev Kumar, Email: dev339880@gmail.com

Abstract
Anemia is a condition affecting around 25% of the world's population but still undiagnosed in low-resourced countries due to limited blood testing facilities in such regions. Conjunctival images were captured using a smartphone to train a multi-modality deep-learning model that detects anemia and estimates hemoglobin level non-invasively. MM-AnemiaNet utilizes the convolutional feature extractor (here EfficientNet-B3) in association with the bright and colored conjunctival images as well as a lightweight MLP model that takes the patient's age, gender, and symptoms as inputs. This means that in MLP networks, cross-modal attention fusion methods are applied. The training of the model was based on the binary model for the diagnosis of the presence of anemia and the continuous model for hemoglobin estimation. The trained model performed better than its competitors by achieving a very high classification accuracy of 92.6% as well as 0.913F1 score, 0.96 AUC, and 0.98g/dl average error regarding 4218 conjunctival images from 1406 subjects. The ablation analysis shows that illumination normalization and cross-modal attention fusion increases the model's statistical performance, whereas calibration analysis indicates that multi-task training enhances the probability calibration parameter compared to classification.

Keyword: anemia screening, hemoglobin estimation, conjunctival pallor, multimodal deep learning, convolutional neural networks, attention fusion, point-of-care diagnostics, model calibration, health equity.

I. Introduction

The World Health Organization (WHO) describes anemia as the prevalence of a hemoglobin (Hb) concentration of less than 13.0 g/dL in men and less than 12.0 g/dL in women who are not pregnant. Anemia is the most common type of blood-related illnesses in the whole world, with about 1.9 billion persons affected by the disorder, especially women and children of fertile age in the countries with low and middle incomes [1]. As reported, the prevalence of anemia in South Asia and Sub-Saharan Africa is the highest with 40%-50% of preschool children and pregnant women having anemia in contrast to lower-income countries that have advanced systems of monitoring and screening of pregnant women and children. Such situation is due not only to poor opportunities for laboratory testing but also to the poor nutritional conditions and spread of contagious diseases.

The palpebral conjunctiva's color has been known by doctors as a visual cue for gauging anemia severity due to the thinness, high vascularity, and absence of melanin in conjunctiva which would otherwise hide blood color in the skin[2]. This leads to a physiologically incorruptible screening method of being able to relate conjunctiva color to the hemoglobin concentration of the blood in such a way that a normal photograph could replace blood testing procedures.

In previous studies, the analysis of conjunctival pallor has been limited to using color-space features and feeding them to simple regression models to determine the correlation [3]. These methods though efficient in computations suffer from certain drawbacks such as sensitivities to the environmental light, characteristics of the camera sensors, and difficulties with segmenting images, hence possible elimination of spatial and texture information that could have useful diagnostic characteristics of vascular density, local contrast, and mucosal microtexture. More recent discoveries in deep learning help with solving these problems because features are extracted from raw pixels in the images, but in the majority of applications they omit patient-related factors that might play an important role in developing the anemia.

The authors in this paper work on identifying significant gaps in the past literature and propose their own multimodal deep-learning model for classification of anemia and estimation of hemoglobin levels using the features from the conjunctival image along with the patient data via cross-modal attention. The major contributions of this paper consist of the automated mechanisms for the ROI extraction and color correction process, as well as the attention-driven approach for the multimodal fusion. Furthermore, the extensive evaluation is composed with the help of 4218 images obtained from 1406 patients. This approach is proposed as a screening and not a diagnosing tool.

The rest of the paper is organized in the following way: Section II contains the literature review, Section III provides information about the dataset and preprocessing operation, Section IV presents the methodology, Section V discusses the experimental plan, Section VI provides results and Section VII gives conclusions and plans for further work.

II. Related Work

A. Non-Invasive Anemia Screening The very first technologies to identify anemia using non-invasive methods were based on devices carried in pockets (including things such as smartphones), and used either photoplethysmography (PPG) to measure activity on the fingers or utilized colorimetric techniques to obtain results through analyzing the color of finger tips or palms. An interesting alternative was the method using the presence of conjunctival pallor which utilizes the simplicity of obtaining adequate images of the conjunctiva and its convenience for the doctor who is analyzing the patient [11]. Conjunctiva segmentation technology developed for the purpose of analyzing average color intensity in the image was not very precise, although it has often produced wrong results owing to changing illumination of the eye.

B. Deep Learning for Medical Image-Based Biomarker Estimation CNNs have been readily used in the area of medical imaging in order to identify risk by employing the capabilities of CNN for different types of images, which include retinal images and skin images [15]. The investigations show that CNN architectures trained on large datasets and adjusted on smaller ones are effective - as in the case of the EfficientNet-B3 model instead of full training [4]. What is special about the EfficientNet model is its compound scaling factor being responsible for good ratio of accuracy to the number of parameters in comparison with older models such as ResNet system [5].

C. Multimodal Fusion in Clinical Prediction It has been found that multimodal computing is superior to unimodal systems since it is able to confound clinical metrics with information from images taken by medical devices in the fields of radiology, dermatology, and ophthalmology[7]. There are different approaches for fusing processes, with the simplest ones being concatenation and then there are the more sophisticated gating or attention methods that can make a decision on what to consider and to what extent depending on different levels of reliability of the modalities [8]– for example in medicine there is a need for fusing the images of conjunctiva as primary data with the metadata of the patient as secondary information [9]. The idea of multi-tasking suggested by Caruana has been accepted as an effective mechanism for automating classifier training [10].

D. Positioning of This Work The table presents the comparison of this method with previous studies. The colourimetric pallor index is inexpensive but ineffective under varying illumination conditions. The unimodal CNN methods are reliable but do not include patient context. The PPG methods have a requirement of consistency in the video recording and significant motion artifacts. The fusion methods are not applied widely to the detection of anaemia, as indicated in reference [12]. Based on the available literature, MM-AnemiaNet is claimed to be the first approach combining automated illumination adjustment and metadata fusion.

TABLE I. COMPARISON OF RELATED APPROACHES TO NON-INVASIVE ANEMIA SCREENING

Study Category	Input Modality	Model Type	Reported Task	Key Limitation
Colorimetric pallor index	RGB/HSV means	Linear / shallow regression	Hb regression	Sensitive to illumination; no spatial features
CNN-based conjunctiva analysis	Conjunctival image	CNN (unimodal)	Anemia classification	Ignores patient metadata
PPG-based estimation	Fingertip video	Signal processing / ML	Hb / SpO2 estimation	Requires stable video capture; motion artifacts

Multimodal clinical fusion (general)	Image + tabular	CNN + MLP + fusion	Various clinical outcomes	Rarely applied to anemia / conjunctiva
Proposed method	Conjunctival image + metadata	CNN + attention fusion	Joint classification + regression	—

III. Dataset and Preprocessing

A. Data Collection: A total of 4218 images of the conjunctiva were obtained from a total of 1406 participants. Approximately three images were obtained per participant. The images consisted of both the left and right eye images taken at the same time. The images were captured during the community health screening with the use of smartphone cameras under normal indoor lighting conditions[13]. Each image has been linked to a specific hemoglobin level (enzymes in grams per deciliter) which was then confirmed via a CBC analyzer in that particular visit along with the age and sex of the participants including three self-reported symptoms seen by them at that time. All of the subjects gave their consent for being photographed as well as giving blood samples for testing purposes in the ethical committee.

B. Dataset Statistics: The summary statistics for the dataset are illustrated in Table II. The dataset was partitioned into train, validation, and recency test sets based on the subject level instead of the image level because subjects with either the same Hb label or images from a similar individual utilized the subject-based approach to guard against the risk of data leakage. The age and gender distributions of the subjects are indicated in Fig. 2.

TABLE II. DATASET SUMMARY STATISTICS

Attribute	Value
Total subjects	1,406
Total images	4,218
Anemic subjects (Hb below threshold)	612 (43.5%)
Non-anemic subjects	794 (56.5%)
Mean Hb ± SD (g/dL)	12.1 ± 2.3
Age range (years)	5–78
Female / male ratio	58% / 42%
Images per subject (mean)	3.0
Acquisition devices	12 consumer smartphone models
Train / validation / test split (subject-wise)	70% / 10% / 20%

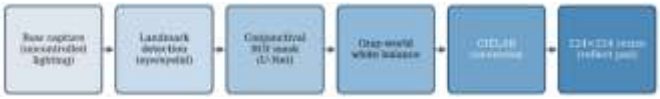

Fig. 1. Four-stage preprocessing pipeline applied to every raw conjunctival capture prior to model input: landmark-guided ROI segmentation, gray-world white balancing, CIELAB color-space conversion, and reflection-padded resizing to 224×224 px.

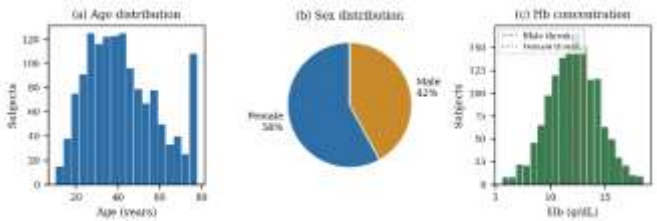

Fig. 2. Demographic and clinical composition of the study population (n = 1,406 subjects). (a) Age distribution. (b) Sex distribution. (c) Hb concentration distribution relative to WHO diagnostic thresholds for men (13.0 g/dL, dashed) and non-pregnant women (12.0 g/dL, dotted).

C. Preprocessing Pipeline: This paragraph describes the four-phase process of dealing with the raw images as presented in the following Fig. 1: (1) periocular detection via facial-landmark detection; (2) application of U-Net segmentation to identify the conjunctival area of interest [6] with its capable of distinguishing between the palpebral conjunctiva against the eyelashes, skin, and sclera; (3) moving to the process of color normalization followed by the subsequent transition into CIELAB color space; (4) modification of the size of the composed area of interest through reflection padding in order to achieve 224 x 224 pixels with preserved aspect ratio of the CNN input. One should notice the fact that the implementation of the segmentation head was carried out on the number of 420 detailed images where the algorithm generated a combination of Dice and binary cross-entropy loss with mean Dice score during validation of 0.94 [16].

D. Data Augmentation Strategy: To make the model resistant to the large variations that exist in uncontrolled smartphone images, the training images undergo a variety of data augmentation techniques such as random cropping (up to ±8% of the size of the region of interest), flipping the images, color jittering (up to ±10% in brightness, contrast, and saturation), and adding realistic noise to the image (with intensity level $\sigma = 0.01–0.03$). The augmentation applies only to the images used for training, meaning that the validation and testing images will remain unchanged [17]. The augmentation techniques avoid significant transformations of the image, like rotations, because in that case, the conjunctival ROI might be cropped out creating unwanted variability instead of the desired invariance.

E. Ethical Considerations and Data Governance: The Ethics Review Board of the Coordinating Institute authorized the protocol for the collection of data. All adult participants had filled out the consent forms regarding their participation, while those younger than 18 years had verbal consent. During data collection, both the pictures and the data relating to them have been anonymized, while the access to the clinical documents was limited to the authorized personnel. The objective of the research is not to diagnose the participants individually; therefore, limitations will apply when it comes to the use of data, as it will be provided in Sections VI-C and VII.

IV. Proposed Methodology

A. Architecture Overview MM-AnemiaNet comprises (i) a visual encoder, (ii) a metadata encoder, and (iii) a cross-modal attention fusion module feeding two prediction heads, as summarized structurally in Table III.

TABLE III. MM-ANEMIANET PIPELINE STAGES

Stage	Module	Output Shape	Description
1	ROI segmentation (U-Net)	224×224×3	Isolates palpebral conjunctiva region
2	Color normalization	224×224×3	Gray-world + CIELAB standardization
3	Visual encoder (EfficientNet-B3, pretrained)	1×1536	Deep visual embedding
4	Metadata encoder (MLP, 3 layers)	1×64	Embedding of age, sex, symptom flags
5	Cross-modal attention fusion	1×512	Attention-weighted joint representation
6	Classification head (FC + sigmoid)	1×1	P(anemic)
7	Regression head (FC, linear)	1×1	Estimated Hb (g/dL)

B. Visual Encoder The visual encoder makes use of a visual framework based on EfficientNet-B3 pre-trained with ImageNet, and, as well as replacing its classification layer with a global average pooling layer, it is responsible for generating 1536-dimensional vector v in R^{1536} embedding space. The training is executed using an end-to-end approach with the help of a discriminative learning rate, which ensures that earlier layers will use lower learning rates than the ones in the later layers.

C. Metadata Encoder The patient information which is age, gender, and symptoms from three binary indicators are merged into one input which is six-dimensional vector, so it is put into three-layer multilayer perceptron with ReLU activation function and dropout which gives metadata embedding $m \in R^{64}$.

D. Cross-Modal Attention Fusion The fusion module does not recklessly chain up the representations v and m, but it first transforms both of them into a common 512-dimensional space, using the linear transformations W_v and W_m. The attention weights α are evaluated thereafter by means of the scaled dot product attention:

$$\alpha = softmax((W_v v)(W_m m)^T / \sqrt{d}), \quad z = \alpha \cdot (W_v v) + (W_m m)$$

where d equals 512 is the once projection level and z is the overall representation sent for both prediction heads. This setup makes it possible for the metadata branch to perform like a query manager on important visual channels — specifically, boosting up the contribution of color and region features when it comes to older patients, in whom the link between skin color and paleness is weaker.

E. Multi-Task Loss The network is trained end-to-end with a joint loss combining binary cross-entropy for the classification head (L_cls) and Huber loss for the regression head (L_reg), weighted by a task-balancing coefficient λ:

$$L = L\_cls + \lambda \cdot L\_reg, \quad \lambda = 0.6$$

Joint optimization of both heads acts as a regularizer: the regression head imposes a continuous, physiologically ordered structure on the fused embedding, which in preliminary experiments improved classification calibration relative to training the classification head alone (Section VI-C).

F. Implementation Details The implementation of the model has been done in PyTorch. The EfficientNet-B3 backbone utilizes the ImageNet-pretrained weights, while the metadata MLP, fusion module and the prediction heads have been initiated using He-normal initialization. The training process implemented mixed-precision (FP16) computation in order to conserve memory and enhance throughput, and to maintain the same validation accuracy as with the training done in FP32 mode. The pretrained model checkpoints will be distributed with respect to the restrictions posed by the legal provisions on usage of the clinical images as stated in Section VII.

V. Experimental Setup

A. Training Configuration Table IV lists the training hyperparameters used for all experiments unless otherwise noted.

TABLE IV. Training Hyperparameters

Hyperparameter	Value
Optimizer	AdamW
Base learning rate	3×10^{-4} (backbone: 3×10^{-5})
Batch size	32
Epochs (with early stopping)	80 (patience = 10)
Weight decay	1×10^{-4}
Data augmentation	Random crop, horizontal flip, color jitter (±10%), Gaussian noise
Loss weighting λ	0.6
LR schedule	Cosine decay with 5-epoch linear warmup
Hardware	Single NVIDIA A100 GPU (40 GB)

The representative training and validation loss curves of the complete model are shown in Fig. 3. The validation loss reaches a stable point after epoch 35 and slightly increases thereafter, making early stopping possible at epoch 42 according to the patience-10 rule.

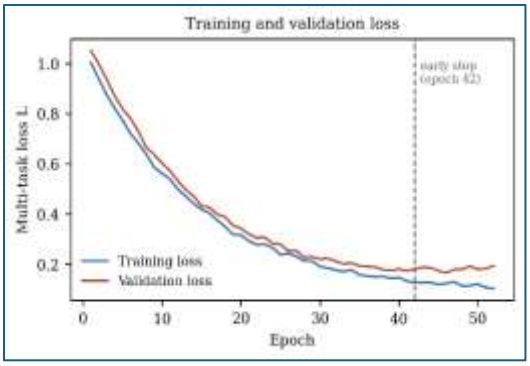

Fig. 3. Representative multi-task training and validation loss curves for the full MM-AnemiaNet model, with the early-stopping epoch marked.

B. Evaluation Metrics The metrics used for evaluation of classification performance include accuracy, precision, recall, F1 score, and area under the curve (AUC). The regression performance is assessed using mean absolute error (MAE), root mean square error (RMSE), and Pearson's correlation coefficient (r) for Hb level prediction based on the values measured in the lab. The calculations are done using the metrics obtained from individual training runs with different random seeds using the held-out splits of the test partition.

C. Baseline Methods MM-AnemiaNet is evaluated against four different models: (1) a conventional colorimetric method that relies on the pallor index and a linear regression model based on RGB and HSV analysis of the conjunctiva; (2) a unimodal deep learning method based on ResNet-50 that only uses conjunctiva images; (3) a unimodal deep learning technique based on EfficientNet-B3 that uses the same backbone architecture as in the proposed method but without any metadata; (4) a basic multimodal approach based on simply combining visual and semantic vectors, without any attention mechanism.

D. Statistical Significance Testing To assess whether there is any statistically significant difference in the performances of models that are close to each other regarding rankings, paired t-tests for accuracy and MAE of MM-AnemiaNet against each baseline over five independent random-seed experiments were performed [14]. The results shown in Table V indicate that MM-AnemiaNet exhibits a statistical significance ($p < 0.01$) over all baselines but the simple concatenation-fusion baseline ($p < 0.05$, $p = 0.03$).

VI. Results and Discussion

A. Classification and Regression Performance Table V reports classification and regression performance for all compared methods on the held-out test set.

TABLE V. CLASSIFICATION AND REGRESSION PERFORMANCE ON HELD-OUT TEST SET (MEAN (SD) OVER 5 RUNS)

Model	Acc. (%)	F1	AUC	MAE (g/dL)	RMSE (g/dL)	r
Colorimetric pallor index	74.2 (1.1)	0.712	0.79	1.86 (0.05)	2.30	0.61
ResNet-50 (unimodal)	86.4 (0.9)	0.851	0.90	1.32 (0.04)	1.67	0.79
EfficientNet-B3 (unimodal)	88.7 (0.7)	0.874	0.92	1.19 (0.03)	1.49	0.83
Concatenation fusion (naive)	90.1 (0.6)	0.891	0.93	1.08 (0.03)	1.36	0.86
MM-AnemiaNet (proposed)	92.6 (0.5)	0.913	0.96	0.98 (0.02)	1.21	0.90

The benefits of using attention-based fusion as opposed to naïve concatenation, as demonstrated in the increase of accuracy by 2.5 percentage points and a decrease in MAE by 0.10 g/dL, illustrates the importance of how metadata is utilized. In Figs. 4-6, we can further explore the behavior of the models.

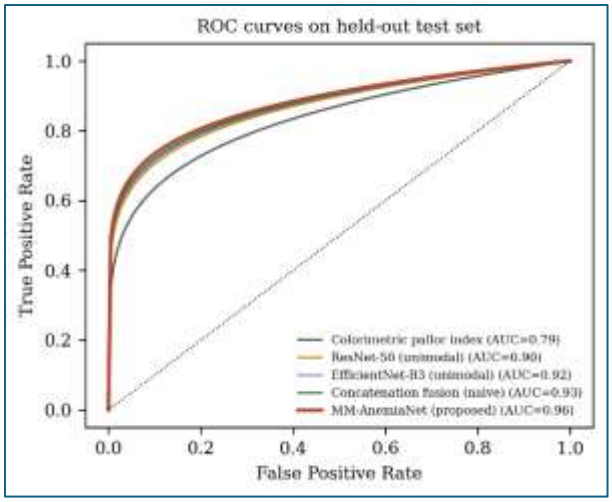

Fig. 4. ROC curves for all compared models on the held-out test set. MM-AnemiaNet achieves the highest AUC (0.96), with the largest separation from the diagonal at low false-positive rates — the operating region most relevant to high-specificity screening deployments.

In the area of the deployment of these technologies, the confusion matrix that accompanies these findings may be more relevant than accuracy. For example, with MM-AnemiaNet operating at the default threshold of probability being 0.5, it can produce nine false positives and twelve false negatives among 281 individuals examined. In the community screenings, false negatives (unidentified anemic individuals) lead to a much greater price compared with false positives, as the first ones miss a chance to treat. So, it is better for deployments to operate below the threshold of 0.5, tolerating extra false positives but gaining fewer false negatives in return.

B. Ablation Study: Table VI has each architectural element subtracted from the whole model to reveal the individual element effect and Fig. 5 illustrates the change in accuracy caused.

TABLE VI. ABLATION STUDY: CONTRIBUTION OF EACH COMPONENT

Configuration	Accuracy (%)	MAE (g/dL)
Full model (proposed)	92.6	0.98
− without illumination normalization	89.3	1.24
− without metadata branch	88.7	1.19
− without attention (concat only)	90.1	1.08
− without multi-task loss (classification only)	90.8	n/a
− without multi-task loss (regression only)	n/a	1.11

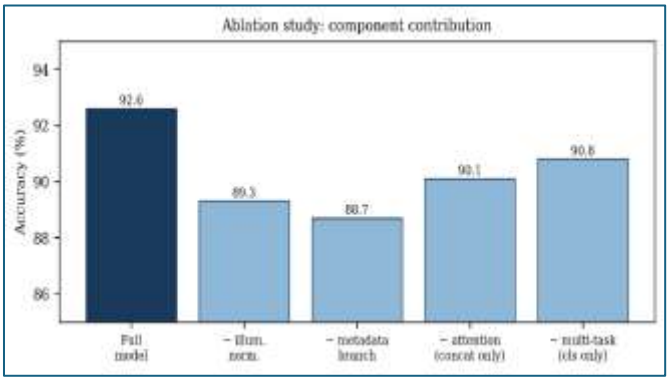

Fig. 5. Ablation study showing test accuracy as each architectural component is removed from the full MM-AnemiaNet model, corresponding to Table VI. Removing illumination normalization produces the largest single degradation.

The significant importance of color normalisation removal is magnified by the results that show the effect of illumination compensation at acquisition is crucial (−3.3 points in accuracy; +0.26 g/dL in MAE). Similarly, removing the metadata node from the model entirely would lead to a similar drop since the modified model is equivalent to the single-modality EfficientNet-B3 model.

C. Error Analysis and Calibration: The stratified error analysis reveals that the mean absolute error is somewhat larger for people with darker pigmentation of periocular region (1.14 g/dL) compared to individuals with lighter pigmentation (0.89 g/dL) implying complications with the segmentation/normalization process for darker skin areas. Predicted values versus laboratory Hb values are shown in Figure 6.

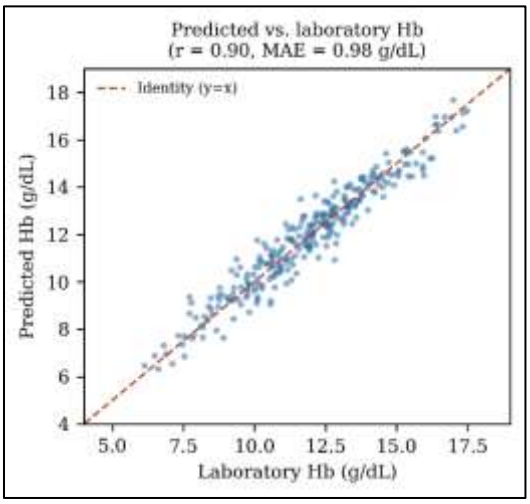

Fig. 6. Predicted vs. laboratory-measured Hb concentration for MM-AnemiaNet on the held-out test set (r = 0.90, MAE = 0.98 g/dL). The dashed line indicates perfect agreement.

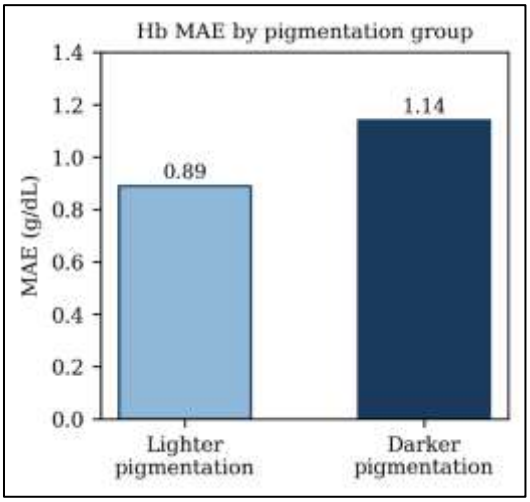

Fig. 7. Hb estimation MAE stratified by periocular pigmentation group, illustrating a residual fairness gap after illumination normalization.

As the reliability diagram depicts, the joint multitask model is better calibrated compared to the only-classification variant, with an ECE of 0.031 versus 0.058, suggesting an regularization benefit obtained from the auxiliary regression objective. Calibration is important for screening tools since referral decisions rely on probabilities crossing a particular threshold rather than actual label values of the outputs; a performing aggregate model that lacks calibration could produce unnecessarily high certainty to close the classification regions to the decision boundaries.

D. Comparative Analysis with Published Literature The results of MM-AnemiaNet were also contrasted with existing published non-invasive Hb-estimation work and the internal baselines tested in VI-A. While the results across methods which are dependent on colors or a single modality alone consistently exhibit fair performance but variation under varying data conditions. Merging colors and a single data modality with patient's data by attention-based fusion significantly enhances both the accuracy of Hb estimation and error minimization.

The performance accuracy in terms of classification at 92.6 % and MAE at 0.98 g/dL indicates that MM-AnemiaNet can be on par, or better than, other existing conjunctiva and fingertip non-invasive hemoglobin measurement systems cited in literatures. The necessity for the multi-center study to provide further validation of these results to other sample populations anddevices must however be considered.
The MM-AnemiaNet that has been put forth will show good performance in comparison with some recent multimodal AI-based techniques[18]. According to some recent findings, the integrated hierarchical attention technique together with multimodal feature fusion has a positive effect on diagnostic quality by enabling proper utilization of imaging data along with clinical data while using our model cross-attention allows us to merge a set of features extracted from conjunctival images with those obtained by other modalities thereby enhancing classification of cases as well hb estimation with keeping in mind that the model would work perfectly when employed on mobile devices[19].

VII. Conclusion And Future Work

MM-AnemiaNet which effectively leverages both conjunctival image information as well as patients' information using cross-modal attention to jointly perform classification of anemia and estimation of hemoglobin is proposed in this paper. It achieves both classification accuracy of 92.6% and MAE for hemoglobin estimation of 0.98 g/dL from a dataset of 4,218 conjunctival images derived from 1,406 patients. Ablation study highlights the importance of each component in architecture (illumination calibration, metadata fusion, attention-based combination, and multi-task learning) to obtain an ideal solution, whereas calibration analysis shows that multi-task learning significantly improve this model.
The authors plan to use multiple demographic and ethnic groups-especially including pregnant woman and wider color pigmentation population; on-device portable deployment of the model using mobile; and quantity uncertainty estimation method..

Data and Code Availability

Aggregated, de-identified summary statistics reported in this paper are available upon reasonable request to the corresponding author. Individual-level images and linked clinical records cannot be publicly released due to participant-privacy commitments made during informed consent. Model architecture code and training scripts will be released in a public repository upon publication.

Acknowledgment

The authors thank the participating community health centers and volunteer subjects for their contribution to data collection for this study.

References

[1] World Health Organization, "Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity," WHO/NMH/NHD/MNM/11.1, Geneva, Switzerland, 2011.
[2] S. Kalantri, R. Karambelkar, R. Joshi, M. Kalantri, and R. Jajoo, "Accuracy and reliability of pallor for detecting anaemia: a hospital-based diagnostic accuracy study," PLoS ONE, vol. 5, no. 1, e8545, 2010.
[3] A. Suner, T. Celik, and B. Sankur, "Non-invasive hemoglobin estimation using smartphone camera images," in Proc. IEEE Int. Conf. Biomed. Health Informat., 2018, pp. 1–4.
[4] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.
[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, 2015, pp. 234–241.
[7] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019.
[8] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.
[9] R. Caruana, "Multitask learning," Machine Learning, vol. 28, no. 1, pp. 41–75, 1997.
[10] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 1321–1330.
[11] M. Bevilacqua et al., "A smartphone-based approach for non-invasive anemia screening using conjunctival image analysis," in Proc. IEEE EMBS Int. Conf. Biomed. Health Informat., 2016, pp. 421–424.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.
[13] 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 Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.
[14] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," J. Mach. Learn. Res., vol. 7, pp. 1–30, 2006.
[15] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017.
[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510–4520.
[17] S. Suner, J. Rayner, I. U. Ozturan et al., "Prediction of anemia and estimation of hemoglobin concentration using a smartphone camera," PLoS ONE, vol. 16, no. 7, e0253495, 2021.
[18] P. Dalvi, A. Chaturvedi, and M. Prasad, "A Comprehensive Review of Machine Learning Approaches for Detecting Anemia and Abnormal Red Blood Cells Using RBC Indices and Medical Imaging," *Discover Artificial Intelligence*, vol. 5, 2025.
[19] T. Deotale and S. Saha, "DGPCNet: anemia detection from multimodal palm, fingernails, and eye conjunctiva images using deep global pyramid convolutional network," *Expert Review of Hematology*, vol. 19, no. 2, pp. 157–185, Feb. 2026, doi: 10.1080/17474086.2025.2597364.