Project name: HbDetector

Authors: Sutolimin Widjaja, Chelsea Miranda Lumban Gaol, Jolin Luk

Date: 7 October, 2025

1. Approach

We designed a hybrid deep learning approach through our model, HbDetector, to predict the Haemoglobin of prospective users through their facial images, specifically the color of their lips. We combined handmade skin-based color and texture features derived from color consistency and HSV masking as well as the Convolutional Neural Network (CNN) features from a pretrained ResNet-18. We decided to proceed with this approach because HB levels are deeply connected with a person's skin color and texture, which is considerably hard to predict using only a CNN. By combining handcrafted skin color statistics and CNN, we can train our model with accuracy for the HB detection. So, we aim to focus on leveraging the physiological cues and visual representations to improve robustness and the differences in skin tone and lighting.

Our model was implemented by Pytorch that not only is easy to read, but one of the best choices due to its flexibility, ease of integrating pretrained CNNs with custom features, and GPU acceleration, making it suitable for hybrid deep learning models like HbDetector. Pytorch also offers image preprocessing feature and feature extraction handled using OpenCV, NumPy, and Pillow (PIL), providing an efficient and reproducible pipeline for Hb prediction.

2. Data Usage and Augmentation

We used different facial images that centered toward the lips provided by the committee as our dataset in a folder that is in the labels.csv with the list of the Hb levels and some of them have their ethnicities. We resized our images before inputting them into our code to 224 x 224 to match the requirement of ResNet-18 and scaled the pixel value to be trained by the model with more efficiency. To prevent overfitting the images where the model can memorize the training images instead of learning the patterns of the images, we augmented our data so that it can recognise different patterns through ColorJitters to change brightness, contrast, saturation, and hue of the image so that our model still able to identity lip color under various circumstances. Other than that, we implemented Horizontal Flip to indicate facial symmetry.

We split our dataset into training and validation sets to evaluate model performance fairly. Approximately 89% of images were used for training, making the model to learn patterns, while 11%

were reserved for validation to assess generalization on unseen images. This split was implemented using train_test_split (test_size = 0.11).

3. Model

Our model is based on the ResNet-18 backbone that already got pretrained on ImageNet, where it can extract features that are high in level from the images. ResNet-18 extracts 512 deep features from each image and these features are linked together with 28 handcrafted skin features and passed through fully connected layers to produce the predicted Hb. The task is a regression task that perpetually predicts the numerical value of the Hb levels in (g/dL).

4. Loss

We identify the difference between our predicted Hb values and the real values using Mean Squared Error (MSE) and the Adam optimizer to minimize the gap between the actual values and the predicted Hb values. The smaller the MSE, the closer the prediction to the actual Hb values.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

yi = true Hb value of image i

y^i = predicted Hb value of image i

n = number of images

Our optimizer, Adam, updates the weight of the model for it to learn from errors, so that it will minimize the difference between predicted and true hemoglobin values. By adjusting the weights that determine how input features affect the output through Adam and backpropagation, the model learns the relationship between image and Hb levels and it will improve the prediction over time.

5. Calibration or Confidence

We have to ensure the accuracy of our predictions and the reliability of our model. Every Hb estimation level is accompanied by a confidence level indicated in a form of percentage. The confidence level considered and skin-tone alignment, by comparing the average lab color of the lips or face to reference values and the lighting alignment through comparing the image's color temperature to ideal daylight (6500K). The higher confidence in percentage corresponds to more reliable

predictions, while lower confidence highlights images that may be affected by poor lighting or unusual skin tones.

6. Error Analysis

There are several conditions where our model might be less reliable to produce predictions. The conditions are when lighting becomes an issue where it affects the lip color such as the underexposure and the overexposure of given images, images with occlusion where the subjects wear mask or glasses, and the bad representations of light or highly dark skin tones on the images. However, the functions analyze_lighting() helps to identify images with bad lighting and the get_skin_tone() function also contributes to identifying over or under represented skin color.

Another error occurs when the input images are unclear or not framed properly, such as when the lips are positioned too close to the top of the image, resulting in the neck occupying most of the frame and when the li;p images are surrounded by a black screen from the dataset. Therefore, a standardized image capture format should be established to ensure the lips are clearly visible and properly centered for accurate detection.

7. Fairness

Since haemoglobin levels correlate with both skin tone and reflectance properties, our model includes a skin feature normalization step using an offset-based method.

We first define a reference skin tone—a standard Lab color profile representing neutral, medium-tone skin under ideal daylight (≈6500 K). Each image's average Lab values from the detected lip region are then adjusted relative to this reference, producing a color deviation that quantifies how far the observed tone differs from the standard. These offsets are incorporated into the handcrafted feature vector, allowing the model to focus on relative color patterns.

8. Ablations

The baseline used only ResNet-18's convolutional features to predict hemoglobin levels from facial images, while the ablated version included additional handcrafted skin-based features (statistical descriptors derived from HSV-masked regions and normalized via a gray-world algorithm). The inclusion of these skin features consistently lowered the validation mean squared error (MSE) compared to the baseline, indicating that explicit color and texture cues contribute meaningful information beyond deep visual embeddings. This confirms that combining handcrafted physiological features with deep representations enhances both robustness and interpretability in hemoglobin prediction.

9. Compute Cost

Our implementation emphasizes computational efficiency and reproducibility. The model was trained using moderate resources under typical deep learning environments, demonstrating that reliable haemoglobin prediction can be achieved without specialized hardware.

By using the compact ResNet-18 architecture together with pre-computed skin color features, our model keeps the training process lightweight and efficient. The handcrafted features are calculated once and reused, preventing unnecessary recomputation during training. Image resizing, flipping and color adjustments are handled through optimized PyTorch and OpenCV functions.

10. Failure Modes

While HgbDetector demonstrates strong overall performance, its accuracy decreases under certain visual conditions. The most significant failure mode involves extreme lighting, where excessive brightness or contrast distorts color gradients and saturation levels. These distortions reduce the model's ability to capture subtle visual cues in the lip region, leading to unreliable feature extraction and weaker gradient responses during training.

A second failure arises from the interaction between lighting and skin tone. Because lighting intensity affects surface reflectance differently across pigmentation levels, the combined effect becomes multiplicative—amplifying errors in both the handcrafted skin features and the CNN output. Under very bright or dark conditions, this interaction causes the model to misinterpret tone variations as changes in haemoglobin concentration.

Overall, these failure modes show that lighting normalization and adaptive tone calibration remain critical for improving model robustness across diverse users.

11. Next Steps

To further improve HgbDetector, several steps can be taken in the future. Firstly, the dataset should be expanded and balanced to include a wider range of skin tones, lighting conditions and demographics. This would help the model generalize better across different users. Secondly, additional data labeling and verification should be done to correct any inconsistencies in filenames and haemoglobin values.