Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Chenlu Shan

Date

Approval Sheet

Evaluation of color space transformations in separating anemia vs control subjects

By

Chenlu Shan

MSPH

Biostatistics and Bioinformatics

Amita Manatunga

Committee Chair

Jeong Hoon Jang

Committee Member

Abstract Cover Page

Evaluation of color space transformations in separating anemia vs control subjects

By

Chenlu Shan

B.S.

Zhejiang Normal University

2017

Thesis Committee Chair:Amita Manatunga

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

Biostatistics and Bioinformatics Department

2021

Abstract

Evaluation of color space transformations in separating anemia vs control subjects

By

Chenlu Shan

Anemia detection requires invasive blood testing with the gold standard hemoglobin level by a complete blood count (CBC). This procedure involves phlebotomy, which could lead to certain adverse effects. We consider a recently proposed non-invasive method for monitoring anemia with a fingernail image taken from patients [1]. Non-invasive anemia detection involves procedure of extracting color data from patients' fingernail bed smartphone images. Usually, color data are extracted in RGB (Red,Green,Blue) color space. Since it is unknown that which color space transformation is most useful in discriminating anemia population vs healthy population, our work focused on determining which transformation of RGB (Red,Green,Blue) data would produce the best separation among anemic and control subjects. Several metrics are used to access the separation between anemia and healthy groups based on within sum of squares and between sum of squares Metric 1:Ratio of within to between variation). It is shown that HSI (Hue,Saturation,Intensity) color space perform the best in comparison to all other color spaces (Metric 1: 1.02 (0.82-1.35) for HSI, Metric 1: 0.76 (0.56-1.05) for RGB). The HSI color space also produced the least mean squared errors in predicting anemia status using Lasso Regression method - (MSE for HSI: 0.085).

We conclude the HSI color space has the best discriminative power in separating anemia vs control and the future prediction models or clustering models should focus on the analysis with HSI color space.

Keywords – non-invasive anemia detection, color space transformation, cluster method, better separation

Acknowledgement

Two-year study at Rollins School of Public health is an unforgettable and important experience in my life, which led me to the field of public health. I am grateful to my parents for supporting my study in the US.

I am extremely grateful to my supervisor, Prof. Amita Manatunga for her invaluable guidance, continuous support, and patience during my thesis work. She is very knowledgeable and has always given me useful suggestions in the past six months. Working with her was a joy and I have learnt a lot from her.

I would like to thank Ying Cui for her help with my thesis. When I asked her questions, she always answered me patiently and gave me a lot of inspiration.

I am grateful to my reader Jeong Hoon Jang for being my reader and providing me with useful comments.

I am grateful to my husband for supporting me during the hard time because of COVID-19. We met at Rollins and studied together.

Lastly, I would like to thank all the faculty members and staff of Rollins School of Public Health. Thanks for your support.

Contents

1 Introduction

Anemia is caused by the body's inability to produce enough hemoglobin, which is a protein that transports oxygen to read blood cells and various tissues of the body. The common symptoms of anemia include fatigue exhaustion, moodiness, and irritability. Other symptoms include dizziness, fainting, apathy, irritability, decreased concentration and unbearable cold feeling [2]. There are about 3 billion people in the world with varying degrees of anemia, and tens of millions of people die every year due to various diseases caused by anemia according to statistics form the World Health Organization (WHO,2008). In the United States, iron deficiency, thalassemia, and anemia of chronic disease are three major causes of anemia [3]. Therefore, detection and timely treatment for anemia is crucial as it affects people in all populations. In particular, anemia among children and older adults may lead to more severe problems like infant mortality and many serious short- and long- term complications such as heart failure, gout attacks, cancer etc. [4].

According to the Word Health Organization definition of anemia, the threshold for detecting anemia vary for different population of subjects. For example, hemoglobin less than 11g/dL for children aged 0 to 4 years and pregnant women, hemoglobin less 12g/dL for children aged 5 to 12 years and non-pregnant women, and hemoglobin less than 13/dL for men [5]. Diagnosing anemia requires invasive blood testing with the gold standard hemoglobin level by a complete blood count (CBC). This procedure involves phlebotomy, which could lead to certain adverse effects. Adam C. Salisbury's work [6] showed that blood loss from greater use of phlebotomy is independently related to Hospital-Acquired anemia. As a result, the procedure of invasive detection of anemia may contribute to worsening anemia. Furthermore, phlebotomy is painful, especially for infants, which makes difficult to complete the procedure. In addition, the procedure of taking blood and testing may take a considerable amount of time. As such, there is a strong need for non-invasive detection and monitoring of anemia, particularly for infant populations.

The Integrated Management of Childhood Illness strategy developed by the World Health

Organization recommends the use of clinical pallor in various anatomical sites as the initial screening tool for anemia in children under 5 years of age for whom laboratory-based hemoglobin is not readily available (WHO, 1997). Recent technology advancements such as smartphones have allowed convenient detection of anemia of subjects. M D Anggraeni [7] introduced a method that allowed non-invasive early detection of anemia among pregnant women. MD Anggraeni's study [7] utilized a smartphone camera to capture digital images of the inferior palpebral conjucttiva of pregnant women and then extracted color intensity of Red, Green, Blue from images of the inferior palpebral conjucttiva. Recently, Manninio et al. [1] introduced a smartphone app which holds promise for detecting anemia non-invasively based on images of patients' fingers. They extracted color data (Red,Green,Blue) on a 51×51 pixel domain from fingernail bed regions and then utilized an algorithm based on robust multi-linear regression with a bisquare weighting algorithm to transfer color data to CBC hemoglobin levels. The smartphone app utilized data of pallor in patient-sourced photos to do the non-invasive diagnosis of anemia. This procedure has many advantages. First, it does not require phlebotomy, which could avoid blood loss of patients and reduce the risk of aggravating anemia of patients. Secondly, this device only needs data of patients' fingernail images and does not make patients uncomfortable during diagnosis. Thirdly, this smartphone app makes the detection of anemia convenient and efficient, not requiring a long time to get the results. The fundamental idea behind of Manninio et al. [1] is that skin detection is useful in diagnosing anemia. In Manninio et al.'s work [1], color data of pallor in patient-sourced photos are collected as RGB(Red,Green,Blue) data. RGB(Red,Green,Blue) is the color space representing the three channels of red, green, and blue. This standard includes almost all the colors that human vision can perceive, which is one of the most widely used color systems.

Development of prediction rules or cluster classification algorithm depends on the structure and types of the data in color space. For example, often, certain model based algorithms requires the assumption of normal distribution to obtain valid predictions. It is also known that certain transformations of data may facilitate the separation between populations. In computer vision community, there are many other color space transformations done

in expressing data in addition to the RGB form [8]. These include NRGB (normalized RGB), HSI(Hue,Saturation,Intensity), YUV(Luma,Blue projection,Red projection),YIQ (NTSC color TV system), CIEXYZ [9], CIELAB. HSI space represents color with three components: hue(H), saturation(S), intensity(I) and can be viewed as transforming the points in the RGB(Red,Green,Blue) color to a cylindrical coordinate system. YUV space represents color with one luma component(Y) and two chrominance components which is U(blue projection) and V(red projection) respectively. YIQ color space is used by the NTSC color TV system. It has Y component which represents the luma information and I and Q component which represents the chrominance information. CIEXYZ color space is the first color space based on the perception of human color vision and the basis of all other color spaces [9].

In the diagnosis of anemia based on clinical pallor, it is unknown that which color space transformation based on patient-sourced photos is most useful in discriminating anemia population vs control population. The objective of this thesis is to determine which transformation of RGB data would produce the best separation among anemic and control subjects. We use the data collected in Mannino et al. [1] to address this objective. After performing 7 color space transformations including RGB, NRGB, HSI, YUV, YIQ CIEXYZ and CIELAB [10], we measure the distance between anemia cluster and healthy cluster under different color spaces. We use two indices to quantify the distance between cases and controls for each color transformation based on the comparison of within sum of squares and between sum of squares. Furthermore, we apply three clustering algorithm including K-means algorithm, discriminative functional mixture model (funFEM) and the K-mean alignment for curve clustering (fdakma) to the whole mixed data set, and evaluate the performance of each clustering algorithm under 7 different color spaces. Finally, we apply Lasso Regression model to the training data set(155 observations) and drive the estimated parameters. The we test the Lasso Regression model using testing data and present the mean squared error.

2 Methodology

2.1 Data Sources

Our study is based on a data set collected in Mannino et al. [1]. During Mannio et al.'s study, data were obtained from a clinical assessment which was conducted at Children's Healthcare of Atlanta, Emory University of School of Medicine, and Georgia Institute of Technology [1]. In the clinical assessment, 265 patients who were diagnosed with the Homozygous Sickle Cell Anemia (HbSS) were recruited and have their hemoglobin levels measured via a complete blood count (CBC). Another 72 healthy subjects from Emory University and The Georgia Institute of Technology were also recruited to the study and have their hemoglobin levels measured via a CBC. Then images of both patients and healthy subjects' fingernail beds were taken. An Apple iPhone 5s (Apple, Cupertino, CA) with default settings were used to take fingernail pictures from four fingers (excluding thumb) of those subjects. Before taking images, brightness adjustment and auto-focus of the smartphone's camera were set in order to make camera lens focus on fingernail bed. Distance between smartphone and subjects' fingernail were set as 0.5m. Images were taken at invariant lighting conditions and room illuminations. In our study, the data set consists 221 subjects. Among 221 subjects, 152 subjects were diagnosed with the Homozygous Sickle Cell Anemia (HbSS), 69 subjects were at a healthy status. Fingernail images of total 221 subjects were obtained. After manually selecting regions of interest, color data were extracted in RGB color space with 51×51 pixel matrix. Figure 2.1 shows the fingernail image and the extracted color data. Finally there were RGB color data with 51×51 pixel matrix of 221 subjects including 152 anemia subjects and 69 healthy subjects.

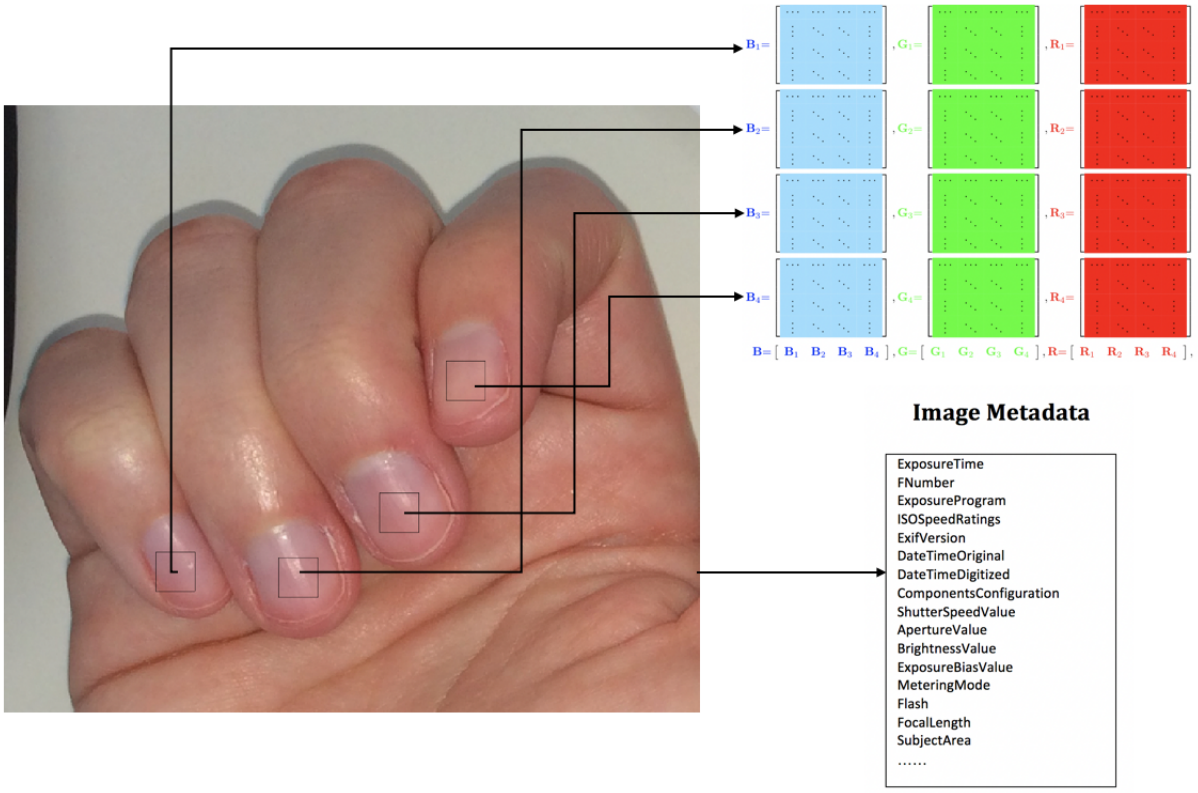

Figure 2.1: Fingernail images collected in Mannio et al. [1]'s study. Color data were extracted in RGB color space with 51×51 pixel matrix.

2.2 Methods

The fundamental concept behind clustering is that given a set of objects, objects within a cluster are more similar to each other than to those in other clusters. This concept lead to different measurements of similarity and dissimilarity within and between clusters. Our study focuses on discriminating anemia cluster and control cluster, which requires quantifying class separability in the respective feature space. In our study, we utilize three methods to evaluate class separability in different color spaces. The first is metrics based on the comparison between between-cluster sum of squares and within-cluster sum of squares. [11]. The second is applying clustering methods to mixed data set and comparing performance of these clustering methods in different color spaces. The third is using Lasso regression model and calculating the minimum square error (MSE) of the model in each color space.

2.2.1 Univariate Case

Suppose that we observe a random univariate variable $x_{i1}, i = 1, ..., m_1$ of sample size m_1 and $x_{i2}, i = 1, ..., m_2$ of sample size m_2. The similarity within a cluster of size m_1 and m_2 can be summarized as within-cluster sum of squares. That is,

$$d_{within} = \sum_{i=1}^{2} \sum_{j=1}^{m_i} (x_{ij} - \overline{x_i})^2$$

Within-cluster sum of squares examines the error variation or variation of individual scores around each group mean. The deviation between two clusters can be summarized as between-cluster sum of squares, that is

$$d_{between} = \sum_{i=1}^{2} m_i (\overline{x_i} - \overline{x})^2$$

Between-cluster sum of squares implies how center of one group is away from the overall mean. The ratio within-cluster sum of squares to between-cluster sum of squares provides the extent to which two clusters are dissimilar while accounting for similarity within the cluster.

$$Ratio(withintobetween) = \frac{d_{within}}{d_{between}}$$

When the subjects within a group are close to each other and far from subjects of the other group, within-cluster sum of squares becomes lower and between-cluster sum of squares becomes larger. Then the ratio of within-cluster sum of squares to between-cluster sum of squares becomes lower. Therefore, lower values of this ratio indicates that clusters are very well separated to each other.

2.2.2 Multivariate Case

Suppose that a random vector $X = [X_1, ..., X_p]'$ of $p \times 1$ is observed. Define $E(X) = \mu_{p\times 1}$,$var(X) = \Sigma_{p\times p}$. Then the variance-covariance matrix is defined as the following

$p \times p$ matrix:

$$\Sigma = cov(X) = \begin{bmatrix} var(X_1) & cov(X_1, X_2) & \cdots & cov(X_1, X_p) \\ cov(X_2, X_1) & var(X_2) & \cdots & cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_p, X_1) & cov(X_p, X_2) & \cdots & var(X_p) \end{bmatrix}$$

Suppose $X_1, ..., X_m$ observations are considered. The sample mean is

$$\overline{X} = \sum_{k=1}^{m} X_k$$

and sample variance-covariance matrix is

$$S_1 = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \overline{X})(X_i - \overline{X})'$$

Consider two sample case. Let $X_{i1}, i = 1, ..., m_1$ be a vector of $p \times 1$,observation from a sample of size m_1 and $X_{i2}, i = 1, ..., m_2$ be a vector of $p \times 1$, observation from a sample of size m_2. Let $P_k = \frac{m_k}{m_1+m_2}, k = 1, 2$, sample mean and estimator of variance-covriance matrix is given by

$$\overline{X_k} = \sum_{i=1}^{m_k} X_{ik}, k = 1, 2$$

and

$$S_k = \frac{1}{m_k - 1} \sum_{i=1}^{m_k} (X_{ik} - \overline{X_k})(X_{ik} - \overline{X_k})'$$

Then within-clsuter sum of squares is defined as

$$S_{within} = \sum_{k=1}^{2} P_k S_k$$

where S_k is the variance-covariance matrix for class $k, k = 1, 2$ here we can use $trace[S_{within}]$ as a measure of the total within variance of the features.

The between-cluster sum of squares is defined as

$$S_{between} = \sum_{k=1}^{2} P_k(\overline{X_k} - \overline{X})(\overline{X_k} - \overline{X})'$$

where $\overline{X}$ is the global mean vector, that is

$$\overline{X} = \sum_{k=1}^{2} P_k \overline{X_k}$$

Usually $trace[S_{between}]$ is a measure of the average distances of the each class mean from the global value. Based on the matrix above, a number of class-separability measures are built around them. Our study used two main measures which are defined as following:

$$metric1 = \frac{trace[S_{within}]}{trace[S_{between}]}$$

Metric1 takes small values when samples of each class are well clustered around their class mean and the clusters of the different classes are well separated.

$$metric2 = trace[{S_{within}}^{-1} S_{between}]$$

Metric2 takes large values when samples of each class are well clustered around their class mean and the clusters of the different classes are well separated. Since these metrics are symmetric positive definite, their eigenvalues are positive. The trace is equal to the sum of eigenvalues, while the determinant is equal to their product, this leads to

$$metric2 = trace[{S_{within}}^{-1} S_{between}] = \sum_{i=1}^{d} \lambda_i$$

where $\lambda_1, ..., \lambda_d$ are the eigenvalues of $trace[{S_{within}}^{-1} S_{between}]$. In our study, smaller metric1 and larger metric2 are desirable for better separation between anemia group and control group.

2.2.3 Color Space Transformation

In order to determine which transformation of RGB (Red, Green, Blue) data would produce the best separation among anemic and control group, We make color space transformation to transform RGB (Red, Green, Blue) data to 6 other color spaces including NRGB (Normalize Red, Green, Blie) color space, HSI (Hue, Saturation, Intensity) color space, YUV (Luma, Blue projection, Red projection) color space, YIQ (Luma, Chrominance information, Chrominance information) color space, CIEXYZ (CIE Red, Green, Blue) color space and CIELAB [9] color space.

NRGB Color Space

We use following transformation formula to transform RGB data to NRGB space:

$$r = \frac{R}{R+G+B}$$

$$g = \frac{G}{R+G+B}$$

$$b = \frac{B}{R+G+B}$$

NRGB space can be viewed as normalizing the RGB (Red, Green, Blue) values. It is a simple and effective way to get rid of distortions caused by lights and shadows in an image.

HSI Color Space

To convert RGB color space to HSI space, we utilized following transformation formula:

$$H = \begin{cases} arcos(\frac{(R-G)+(R-B)}{2\sqrt{(R-G)^2+(R-B)(G-B)}}) & , \quad G > B \\ 2\pi - arcos(\frac{(R-G)+(R-B)}{2\sqrt{(R-G)^2+(R-B)(G-B)}}) & , \quad G \leq B \end{cases}$$

$$S = 1 - \frac{3}{R+G+B} min(R,G,B)$$

$$I = \frac{R+G+B}{3}$$

. HSI space represents color with three components: hue(H), saturation(S), intensity(I) and can be viewed as transforming the points in the RGB color to a cylindrical coordinate system.

YUV Color Space

We use following transformation formula to transform RGB data to YUV space:

$$Y = 0.229R + 0.587G + 0.114B$$

$$U = 0.492(B - Y)$$

$$V = 0.877(R - Y)$$

YUV space represents color with one luma component(Y) and two chrominance components which is U(blue projection) and V(red projection) respectively.

YIQ Color Space

We use following transformation formula to transform RGB data to YIQ space:

$$Y = 0.299R + 0.587G + 0.114B$$

$$I = 0.596R - 0.275G - 0.321B$$

$$Q = 0.212R - 0.523G + 0.311B$$

YIQ color space is used by the NTSC color TV system. It has Y component which represents the luma information and I and Q component which represents the chrominance information.

CIEXYZ Color Space

We use following transformation formula to transform RGB data to CIEXYZ space:

$$X = \frac{1}{0.17697}(0.49R + 0.31G + 0.2B)$$

$$Y = \frac{1}{0.17697}(0.177R + 0.812G + 0.011B)$$

$$Z = \frac{1}{0.17697}(0.01G + 0.99B)$$

CIELAB Color Space

We uitilize following transformation formula to transform CIEXYZ color space to CIELAB color space:

$$L^* = 116f(\frac{Y}{Y_n}) - 16$$

$$a^* = 500(f(\frac{X}{X_n}) - f(\frac{Y}{Y_n}))$$

$$b^* = 200(f(\frac{Y}{Y_n}) - f(\frac{Z}{Z_n}))$$

where $t = \frac{X}{X_n}, \frac{Y}{Y_n}, or \frac{Z}{Z_n}$:

$$f(t) = \begin{cases} \sqrt[3]{t} & if t > \delta^3 \\ \frac{t}{3\delta^3} + \frac{4}{29} & otherwise \end{cases}$$

$$\delta = \frac{6}{29}$$

X,Y,Z describe the color simulus considered and X_n,Y_n,Z_n describe a specified white achromatic reference illuminant [12]. For Standard illuminant D65 [13]:

$$X_n = 85.0489, Y_n = 100, Z_n = 108.8840$$

2.2.4 Statistical Analysis

We employ three methods for measuring separability of anemia and control subjects in different color spaces.

We separate data set consisting of 221 subjects to two groups. The first is case group which includes 152 subjects who are diagnosed with who were diagnosed with the Homozygous Sickle Cell Anemia (HbSS). The second is control group includes 69 healthy subjects. We vectorize each subject using 51×51 and get a 2601×1 vector. Actually we have 51×51

for three coordinates and get a 2601×3 matrix.

Method1:

Suppose $X_1 = [X_{11}, X_{21}, X_{31}]'$ is a random vector for case group and $X_2 = [X_{12}, X_{22}, X_{32}]'$ is a random vector for control group. $X_{1k}, k = 1, 2$ represents red color, $X_{2k}, k = 1, 2$ represents green color, $X_{3k}, k = 1, 2$ represents blue color. Then the mean vector of $X_k, k = 1, 2$ can be computed as

$$\overline{X_k} = \begin{bmatrix} E(X_{1k}) \\ E(X_{2k}) \\ E(X_{3k}) \end{bmatrix} k = 1(anemia), 2(control)$$

The variance-covariance matrix of each vector can be computed as

$$S_k = cov(X_k) = \begin{bmatrix} var(X_{1k}) & cov(X_{1k}, X_{2k}) & cov(X_{1k}, X_{3k}) \\ cov(X_{2k}, X_{1k}) & var(X_{2k}) & cov(X_{2k}, X_{3k}) \\ cov(X_{3k}, X_{1k}) & cov(X_{3k}, X_{2k}) & var(X_{3k}) \end{bmatrix}, k = 1, 2$$

Let $P_k = \frac{m_k}{m_1 + m_2}, k = 1, 2$, P_1 is equal to 0.688 and P_2 is equal to 0.312. Then within-cluster sum of squares can be computed as

$$S_{within} = \sum_{k=1}^{2} P_k S_k$$

where S_k is the variance-covariance matrix for class $k, k = 1, 2$

Between-cluster sum of squares can be computed as

$$S_{between} = \sum_{k=1}^{2} P_k (\overline{X_k} - \overline{X})(\overline{X_k} - \overline{X})'$$

where $\overline{X}$ is the global mean vector, that is

$$\overline{X} = \sum_{k=1}^{2} P_k \overline{X_k}$$

Then $metric1 = \frac{trace[S_{within}]}{trace[S_{between}]}$ and $metric2 = trace[{S_{within}}^{-1} S_{between}]$ can be calculated.

Bootstrap Method

In order to dirive 95% confidence interval of the metric1 and metric2, we apply Bootstrap method [14]. By resampling observations with replacement 1000 times, 95% confidence interval are presented for $metric1 = \frac{trace[S_{within}]}{trace[S_{between}]}$ and $metric2 = trace[{S_{within}}^{-1} S_{between}]$.

Method2:

Clustering

After measuring the distance between anemia cluster and healthy cluster under six color spaces, we combine anemia cluster and healthy cluster together. Now the whole data set consists of 221 subjects without anemia information. We apply three clustering methods including K-means Algorithm, discriminative functional mixture model (funFEM) and the K-mean alignment for curve clustering (fdakma) to the whole mixed data and evaluate the accuracy, sensitivity and specificity of these algorithms with respect to the two groups. In K-means clustering, given a data set with n points in l-dimensional space R^l and a set of k points in R^l, we need to minimize the mean squared distance from each point to the closest center of this point [15,16]. The discriminative functional mixture model (funFEM) is to cluster a set of observed curves $x_1, ..., x_n$ into K groups in each group the curves are homogenous [17]. The k-mean alignment for curve clustering (fdakma) performs clustering and alignment of functional data by means of k-mean alignment [18]. By comparing clustering results with its true anemia status, we can get sensitivity, specificity, positive predictive value, negative predictive value and accuracy of each clustering method under 6 color spaces.

Method3:

Lasso Regression Model

In the third method of measuring class separability in different color spaces, we apply Lasso regression model to the whole data set consisting of 221 subjects with information about anemia status. We separate the whole data set to training data set (155 observations) and

testing data set (66 observations). Then we apply Lasso regression model to the training data set and drive the estimated parameters. Finally we test the Lasso Regression model using testing data set and calculate the mean squared error (MSE).

3 Results

Class separability measurement results in RGB, NEGB, HSI, YUV, YIQ, CIEXYZ and CIELAB color spaces are summarized in Table 3.1. metric1 is the ratio of within-cluster sum of squares to between-cluster sum of squares. Since within-cluster sum of squares calculates the deviation of every subject value away from the overall mean within that particular group and between sum of squares calculate the deviation between two groups, metric1 takes small value when two groups are well separated. Similarly, metric2 takes large value when two groups are well separated.

As seen in Table 3.1, metric1 takes smallest value 7.02 (95% CI, 4.22-12.17) in HSI color space while takes largest value 12 (95% CI, 6.26-30.40) in YUV and YIQ color spaces. This means HSI color space performs pretty well in separating among anemic and control groups based on metric1 while YUV and YIQ color spaces appear to be unsatisfactory in separation among these two groups. Based on metric1, CIEXYZ,CIELAB and RGB color spaces perform pretty well in separating anemia and healthy groups since their metric1 values are relatively small. The result of NRGB appears to be unsatisfactory since its value was above 10. Metric2 takes largest value 1.02 (95% CI, 0.82-1.35) in HSI color space while takes smallest Value 0.6 (95% CI,0.47-0.84) in NRGB color space. This means HSI color space performs pretty well among anemic and control groups based on metric2 while NRGB produces bad separation among two clusters. Under metric1, NRGB is also not doing well. CIELAB color performs also pretty well since its metric2 value is 0.78 (95% CI, 5.65-12.99), which is larger than other color spaces except HSI.

Combining the results of two metric1 and metric2, HSI has good performance based on both metric1 and metric2, which means under HSI color space anemia group and control group are well separated. NRGB performs poorly in the cluster separation. Interestingly, RGB color space does relatively well. Also, CIELAB color space performs pretty well too, which means it did well in producing separation among anemia group and healthy group.

Table 3.1: Performance of different color spaces in separating anemia and control groups

	metric1		metric2	
Color Space	mean	95% CI	mean	95% CI
RGB	8.92	4.99,19.82	0.76	0.56,1.05
NRGB	10.63	6.33,17.35	0.60	0.47,0.84
HSI	7.02	4.22,12.17	1.02	0.82,1.35
YUV	12.04	6.26,30.40	0.76	0.56,1.05
YIQ	12.04	6.26,30.40	0.76	0.56,1.05
CIEXYZ	8.54	4.71,19.17	0.76	0.56,1.05
CIELAB	8.81	5.65,12.99	0.78	0.58,1.07

Smaller values of metric1 and larger values of metric2 indicate better performance

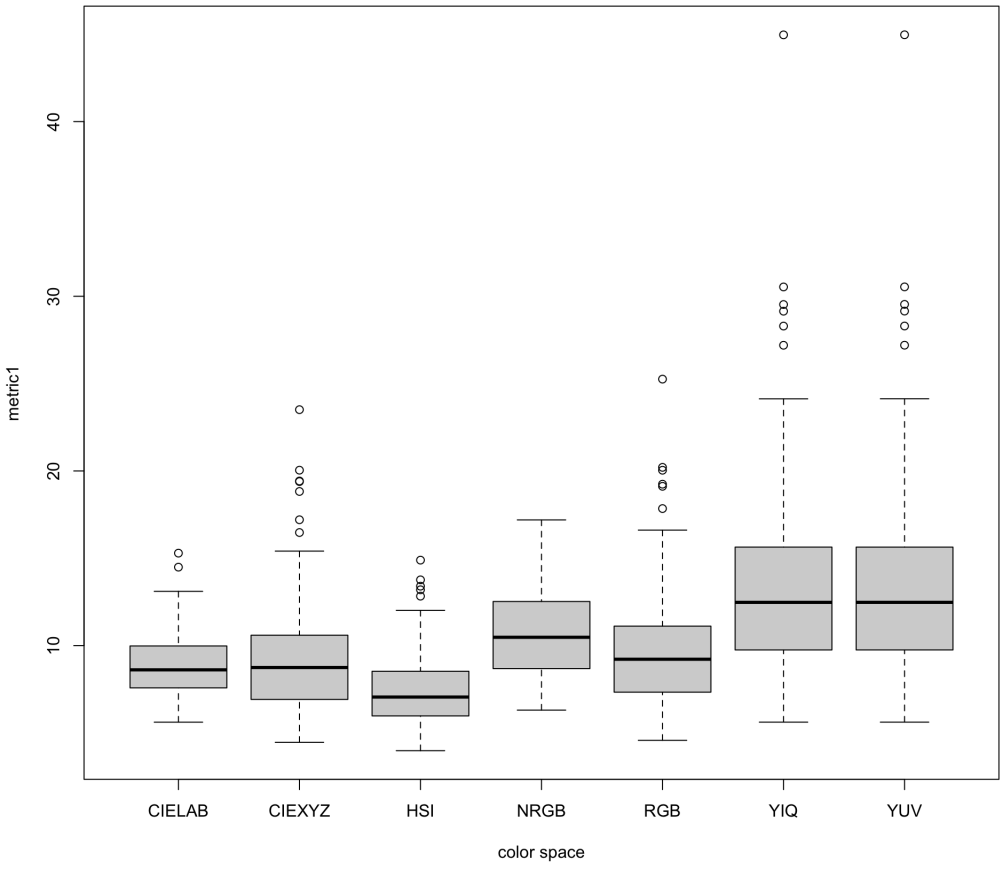

Figure 3.1: Box plots for metric1 in different color spaces

Figure 3.1 presents box plots for metric1 in RGB, NRGB, HSI, YUV, YIQ, CIEXYZ and CIELAB color spaces. The distribution of HSI color space is lower than others, however,

it overlaps with CIELAB, CIEXYZ and RGB color spaces. Also, metric1 in HSI color space is significantly smaller than NRGB, YIQ and YUV color space. Figure 3.1 gives us more intuitive understanding that HSI, RGB, CIEXYZ and CIELAB color spaces produce well separation among case and control class while YUV, YIQ and NRGB color spaces produce least separation among case and control groups.

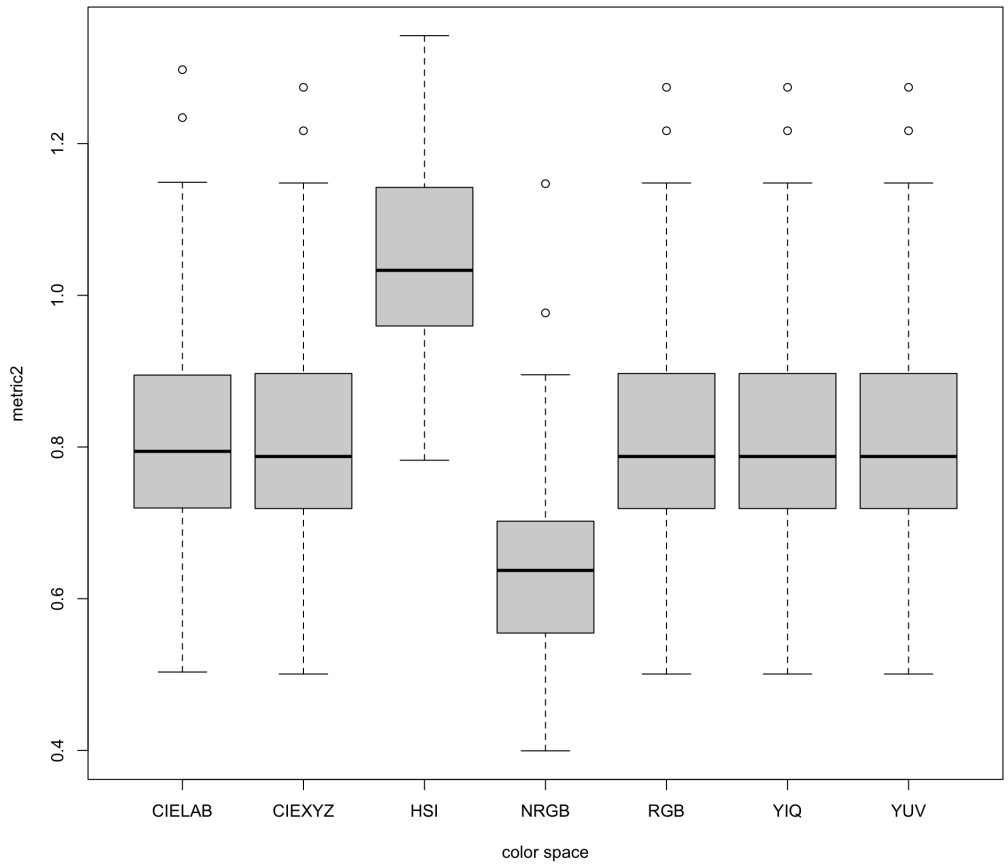

Figure 3.2: Box plots for metric2 in different color spaces

Figure 3.2 presents box plots for metric2 in RGB, NRGB, HSI, YUV, YIQ, CIEXYZ and CIELAB color spaces. The distribution of HSI color space does not overlap with other color spaces. This gives us information that HSI has the best performance in producing separation among anemia and healthy groups.

Table 3.2 shows the performance of K-means cluster model in 7 different color spaces (accuracy,sensitivity, specificity) with the whole data set. The whole data set consists of 221 subjects without information about anemia status. Then K-means algorithm is

applied to the mixed data set. With regard to the accuracy, K-means algorithm has the highest accuracy value 0.64 (95% CI,0.64-0.65) in HSI and CIEXYZ color spaces. The accuracy values of k-means algorithm in RGB is also pretty high as its value is above 0.60. With regard to the sensitivity, RGB color space has the highest value 0.68 (95% CI, 0.64-0.74). The sensitivity of HSI color space is 0.66 (95% CI,0.64-0.65) and the sensitivity of CIELAB color space is 0.67 (95% CI, 0.67-0.68), which also indicate good performance of these two color spaces. With regard to the specificity, the value of HSI color space is 0.64 (95% CI,0.64-0.65) and the value of CIEXYZ color space is 064 (95% CI, 0.64-0,65), which is the highest specificity value among all color spaces.

Table 3.2: Accuracy, sensitivity and specificity of the k-means clustering method under different color spaces

	ACC		SEN		SPE	
Color Space	mean	95% CI	mean	95% CI	mean	95% CI
RGB	0.63	0.60,0.64	0.68	0.64,0.74	0.60	0.54,0.64
NRGB	0.60	0.60,0.60	0.63	0.62,0.63	0.48	0.47,0.48
HSI	0.64	0.64,0.65	0.66	0.65,0.67	0.64	0.64,0.65
YUV	0.60	0.60,0.60	0.64	0.63,0.64	0.59	0.59,0.60
YIQ	0.60	0.60,0.60	0.64	0.64,0.65	0.59	0.59,0.60
CIEXYZ	0.64	0.64,0.65	0.65	0.64,0.65	0.64	0.64,0.64
CIELAB	0.60	0.60,0.61	0.67	0.67,0.68	0.57	0.57,0.57

ACC: Accuracy; SEN: Sensitivity; SPE: Specificity

Table 3.3 shows the performance of the discriminative functional mixture model (funFEM) in 7 different color spaces (accuracy, sensitivity, specificity) with the whole data set. With regard to the accuracy, funFEM algorithm has the highest accuracy value 0.67 (95% CI,0.66-0.67) in HSI color space. The accuracy values of k-means algorithm in RGB and CIEXYZ are also pretty high as their values are above 0.65. With regard to the sensitivity, YUV color space has the highest value 0.68 (95% CI, 0.68-0.68). The sensitivity of HSI color space is 0.66 (95% CI,0.65-0.66), the sensitivity of CIELAB color space is 0.65 (95% CI, 0.65-0.66), the sensitivity of CIEXYZ color space is 0.65 (95% CI, 0.65-0.65), the sensitivity of RGB color space is 0.64 (95% CI,0.63-0.64), which also indicate good performance of these color spaces. With regard to the specificity, HSI color space has the

best performance with value 0.68 (95% CI,0.67-0.68). RGB, CIEXYZ and CIELAB color spaces also perform well since their specificity is above 0.65.

Table 3.3: Accuracy, sensitivity and specificity of the discriminative functional mixture model (funFEM) under differnt color spaces

	ACC		SEN		SPE	
Color Space	mean	95% CI	mean	95% CI	mean	95% CI
RGB	0.66	0.66,0.67	0.64	0.63,0.64	0.66	0.66,0.66
NRGB	0.53	0.52,0.53	0.51	0.51,0.51	0.55	0.54,0.55
HSI	0.67	0.66,0.67	0.66	0.65,0.66	0.68	0.67,0.68
YUV	0.65	0.65,0.66	0.68	0.68,0.68	0.64	0.64,0.65
YIQ	0.63	0.62,0.63	0.61	0.61,0.61	0.64	0.64,0.65
CIEXYZ	0.66	0.65,0.67	0.65	0.65,0.65	0.66	0.66,0.67
CIELAB	0.65	0.65,0.65	0.65	0.65,0.66	0.67	0.66,0.67

ACC: Accuracy; SEN: Sensitivity; SPE: Specificity

Table 3.4 shows the performance of the k-mean alignment for curve clustering (fdakma) in 7 different color spaces (accuracy, sensitivity, specificity) with the whole mixed data set. With regard to the accuracy, fdakma algorithm has the highest accuracy value 0.65 (95% CI,0.64-0.65) in HSI color space, which indicates good performance of HSI color space. With regard to the sensitivity, HSI color space has the highest value 0.75 (95% CI, 0.73-0.76). The performance of RGB color space is also pretty good since its sensitivity is 0.72 (95% CI,0.70-0.73). With regard to the specificity, the value of HSI color space is 0.52 (95% CI,0.51-0.52), which is the highest specificity value among all color spaces..

Table 3.4: Accuracy, sensitivity and specificity of the k-mean alignment for curve clustering(fdakma) under different color spaces

	ACC		SEN		SPE	
Color Space	mean	95% CI	mean	95% CI	mean	95% CI
RGB	0.52	0.51,0.52	0.72	0.70,0.73	0.42	0.41,0.45
NRGB	0.49	0.48,0.51	0.52	0.52,0.52	0.48	0.47,0.48
HSI	0.65	0.64,0.65	0.75	0.73,0.76	0.52	0.51,0.52
YUV	0.58	0.58,0.59	0.65	0.65,0.65	0.46	0.45,0.46
YIQ	0.58	0.57,0.59	0.64	0.63,0.64	0.46	0.46,0.46
CIEXYZ	0.57	0.56,0.57	0.64	0.63,0.64	0.45	0.45,0.45
CIELAB	0.58	0.57,0.58	0.68	0.68,0.68	0.46	0.45,0.47

ACC: Accuracy; SEN: Sensitivity; SPE: Specificity

Table 3.5: Mean squared error (MSE) of Lasso regression model (n=221)

Color Space	MSE
RGB	0.096
NRGB	0.123
HSI	0.085
YUV	0.091
YIQ	0.109
CIEXYZ	0.100
CIELAB	0.094

Table 3.5 shows the performance of the Lasso regression model in 7 different color spaces. It shows that in HSI color space, Lasso regression model has the smallest MSE 0.085, which means the Lasso regression model performs best in HSI color space. In YUV,CIELAB and RGBcolor space, the Lasso regression model also performs well with MSE 0.091, 0.094 and 0.096. Result shows that NRGB's performance is unsatisfactory.

The over mean by pixels for each color space for anemia and control groups are ploteed in Figure A0.1, Figure A0.2, Figure A0.3, Figure A0.4, Figure A0.5, Figure A0.6, Figure A0.7. HSI color space shows clear separation between cases and controls compared to other color spaces.

4 Discussion

In the study, we evaluate the separability of anemia and healthy groups in RGB, NRGB, HSI, YUV, YIQ, CIEXYZ and CIELAB color spaces by three methods. The first method is based on comparison between within-cluster sum of squares and between-cluster sum of squares. The first method indicates that HSI has the best separation performance based on both metric1 and metric. The second method is applying three clustering methods (k-means,funFEM,fdakma) to the mixed data set without information about anemia status. Results show that all the three clustering algorithm perform well in HSI color space. Furthermore, the performance of RGB, CIEXYZ and CIELAB color spaces appear to be reasonable. The third method involves Lasso regression model. Results shows that in HSI color space Lasso regression model has minimum MSE, which indicates better performance of HSI color space. Meanwile, RGB and CIELAB color spaces also show good results. All the three methods combine together indicates that HSI color space produce better sepatation among anemia and healthy groups than any other color space.

There are some limitations of our study. Our study assumes that color data are capable of separating anemia and healthy groups. This might be a reasonable assumption given that WHO recommends color can produce separation between anemia and healthy subjects. In our study, anemia patients may have other chronic diseases. This may confound the results.

Given separation information of these color spaces, we can further improve non-invasive detection device using HSI color data. Also, for classification problems which involve image analyses, HSI color space transformation could be used for better classification.

References

[1] Robert G Mannino, David R Myers, Erika A Tyburski, Christina Caruso, Jeanne Boudreaux, Traci Leong, GD Clifford, and Wilbur A Lam. Smartphone app for non-invasive detection of anemia using only patient-sourced photos. *Nature communications*, 9(1):1–10, 2018.

[2] Dominique Bron, Natalie Meuleman, and Cecile Mascaux. Biological basis of anemia. In *Seminars in oncology*, volume 28, pages 1–6. Elsevier, 2001.

[3] Ernest Beutler. The common anemias. *JAMA*, 259(16):2433–2437, 1988.

[4] Ravi M Patel, Andrea Knezevic, Neeta Shenvi, Michael Hinkes, Sarah Keene, John D Roback, Kirk A Easley, and Cassandra D Josephson. Association of red blood cell transfusion, anemia, and necrotizing enterocolitis in very low-birth-weight infants. *Jama*, 315(9):889–897, 2016.

[5] Kushang V Patel. Epidemiology of anemia in older adults. In *Seminars in hematology*, volume 45, pages 210–217. Elsevier, 2008.

[6] Adam C Salisbury, Kimberly J Reid, Karen P Alexander, Frederick A Masoudi, Sue-Min Lai, Paul S Chan, Richard G Bach, Tracy Y Wang, John A Spertus, and Mikhail Kosiborod. Diagnostic blood loss from phlebotomy and hospital-acquired anemia during acute myocardial infarction. *Archives of internal medicine*, 171(18):1646–1653, 2011.

[7] MD Anggraeni and A Fatoni. Non-invasive self-care anemia detection during pregnancy using a smartphone camera. In *IOP Conference Series: Materials Science and Engineering*, volume 172, page 012030. IOP Publishing, 2017.

[8] Rolf G Kuehni. Color space and its divisions. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color,*

The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur, 26(3):209–222, 2001.

[9] Thomas Smith and John Guild. The cie colorimetric standards and their use. *Transactions of the optical society*, 33(3):73, 1931.

[10] Ian L Weatherall and Bernard D Coombs. Skin color measurements in terms of cielab color space values. *Journal of investigative dermatology*, 99(4):468–473, 1992.

[11] Brian Everitt and Anders Skrondal. *The Cambridge dictionary of statistics*, volume 106. Cambridge University Press Cambridge, 2002.

[12] K McLaren. Xiii—the development of the cie 1976 (l* a* b*) uniform colour space and colour-difference formula. *Journal of the Society of Dyers and Colourists*, 92(9):338–341, 1976.

[13] YONG-KEUN LEE, Huan Lu, and John M Powers. Fluorescence of layered resin composites. *Journal of esthetic and restorative dentistry*, 17(2):93–100, 2005.

[14] Christopher Z Mooney, Christopher F Mooney, Christopher L Mooney, Robert D Duval, and Robert Duvall. *Bootstrapping: A nonparametric approach to statistical inference*. Number 95. sage, 1993.

[15] Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[16] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[17] Charles Bouveyron, Etienne Côme, Julien Jacques, et al. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *Annals of Applied Statistics*, 9(4):1726–1760, 2015.

[18] Laura M Sangalli, Piercesare Secchi, Simone Vantini, and Valeria Vitelli. K-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5):1219–1233, 2010.

Appendix

Table A0.1: Positive predictive value and negative predictive value of the k-means clustering method under different color spaces

	PPV		NPV	
Color Space	mean	95% CI	mean	95% CI
RGB	0.44	0.42,0.45	0.81	0.80,0.82
NRGB	0.43	0.43,0.44	0.89	0.89,0.89
HSI	0.45	0.45,0.46	0.81	0.80,0.81
YUV	0.41	0.41,0.41	0.78	0.78,0.79
YIQ	0.41	0.41,0.42	0.78	0.78,0.79
CIEXYZ	0.45	0.45,0.46	0.80	0.80,0.80
CIELAB	0.41	0.41,0.42	0.79	0.79,0.80

PPV: Positive predictive value; NPV: Negative predictive value

Table A0.2: Positive predictive value and negative predictive value of the discriminative functional mixture model (funFEM) under differnt color spaces

	PPV		NPV	
Color Space	mean	95% CI	mean	95% CI
RGB	0.46	0.46,0.47	0.80	0.80,0.80
NRGB	0.34	0.34,0.34	0.71	0.71,0.72
HSI	0.52	0.50,0.53	0.85	0.84,0.85
YUV	0.46	0.46,0.47	0.82	0.82,0.82
YIQ	0.43	0.43,0.43	0.78	0.78,0.79
CIEXYZ	0.47	0.47,0.47	0.81	0.81,0.82
CIELAB	0.46	0.46,0.46	0.80	0.80,0.81

PPV: Positive predictive value; NPV: Negative predictive value

Table A0.3: Positive predictive value and negative predictive value of the k-mean alignment for curve clustering(fdakma) with under different color spaces

	PPV		NPV	
Color Space	mean	95% CI	mean	95% CI
RGB	0.48	0.46,0.49	0.76	0.75,0.76
NRGB	0.32	0.32,0.33	0.70	0.68,0.71
HSI	0.49	0.47,0.50	0.83	0.82,0.83
YUV	0.46	0.45,0.46	0.77	0.76,0.78
YIQ	0.48	0.48,0.49	0.78	0.78,0.78
CIEXYZ	0.49	0.47,0.50	0.78	077,0.78
CIELAB	0.49	0.49,0.49	0.79	0.78,0.79

PPV: Positive predictive value; NPV: Negative predictive value

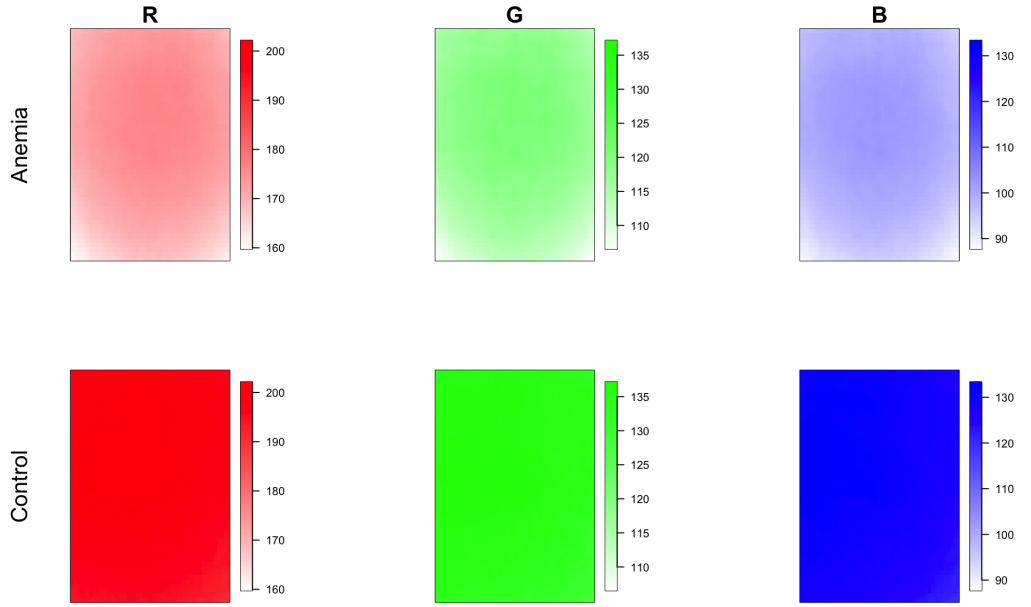

Figure A0.1: Heatmaps for the mean color data at each pixels for anemia and control groups in RGB color space.

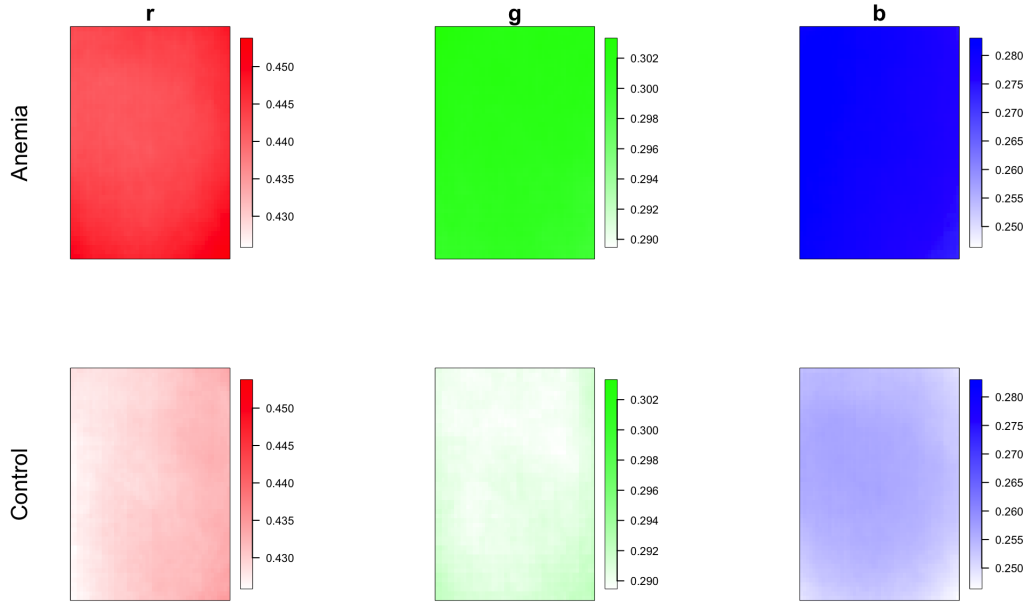

Figure A0.2: Heatmaps for the mean color data at each pixels for anemia and control groups in NRGB color space.

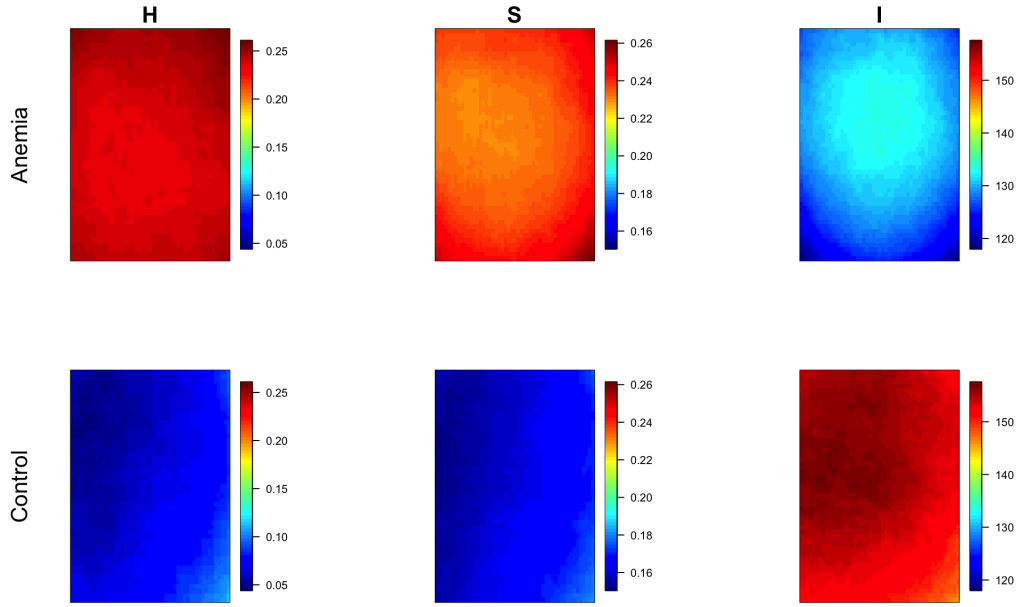

Figure A0.3: Heatmaps for the mean color data at each pixels for anemia and control groups in HSI color space.

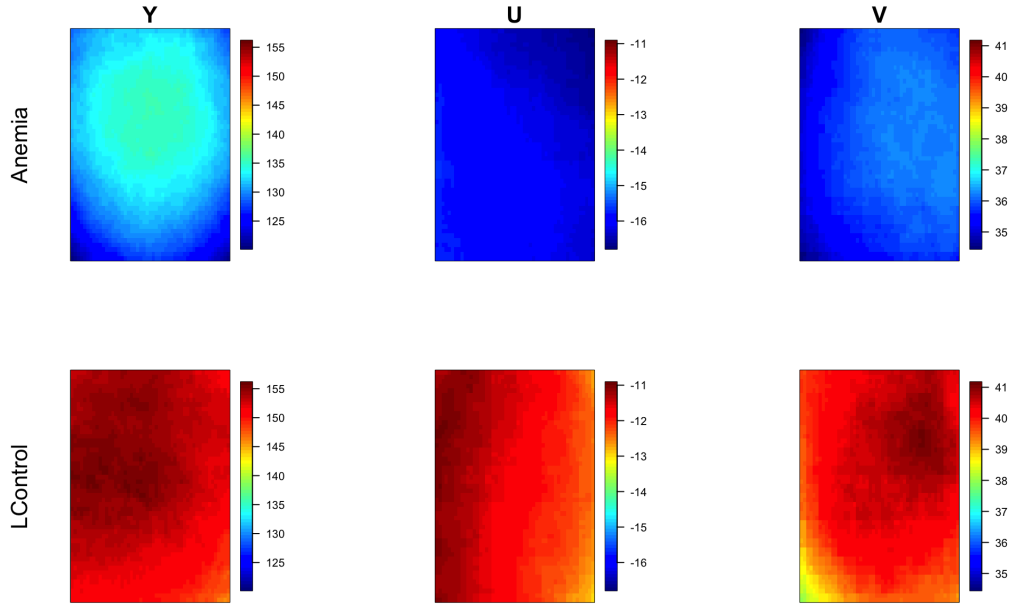

Figure A0.4: Heatmaps for the mean color data at each pixels for anemia and control groups in YUV color space.

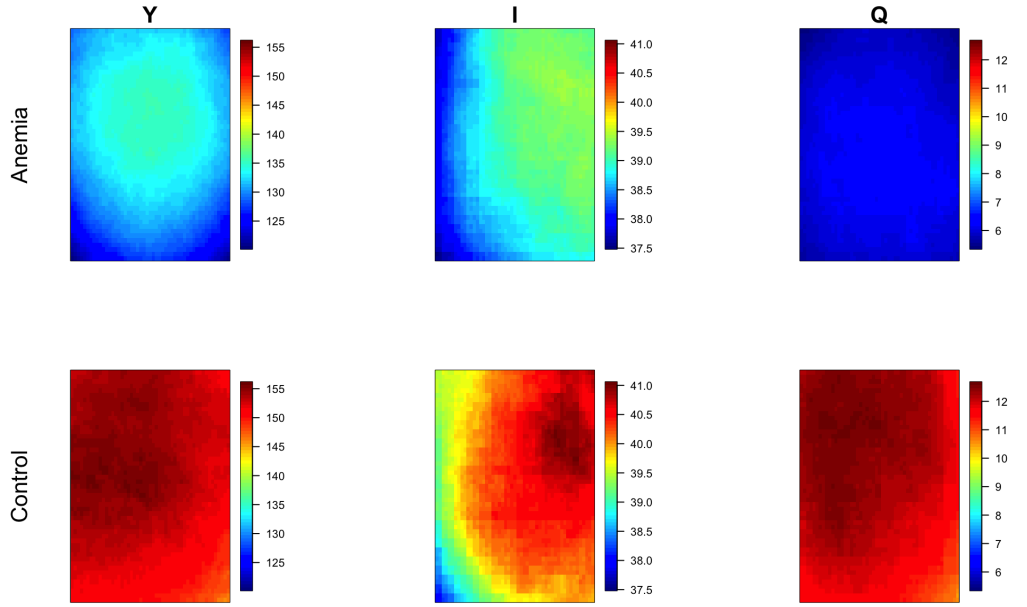

Figure A0.5: Heatmaps for the mean color data at each pixels for anemia and control groups in YIQ color space.

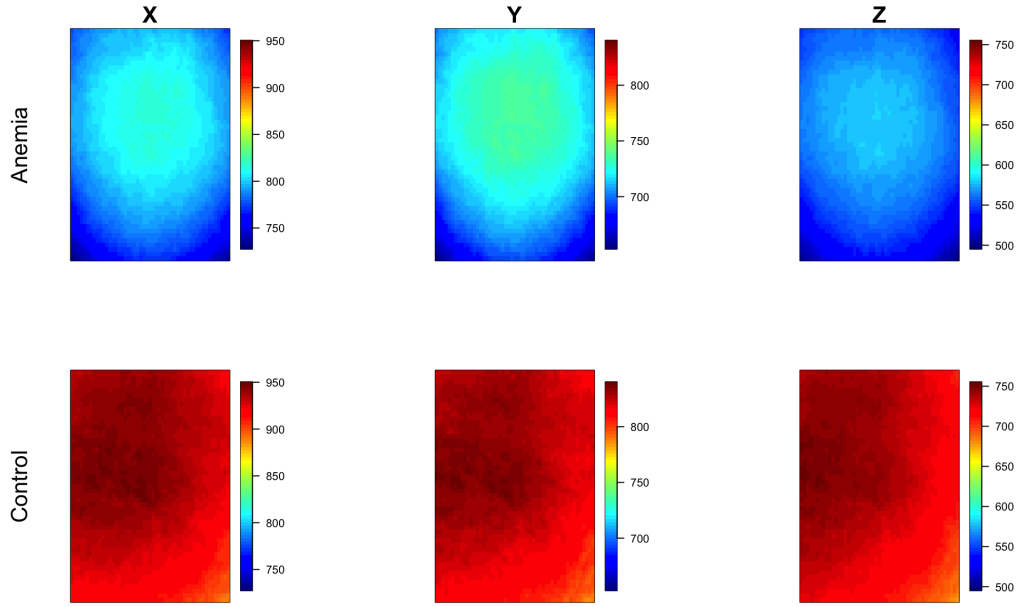

Figure A0.6: Heatmaps for the mean color data at each pixels for anemia and control groups in CIEXYZ color space.

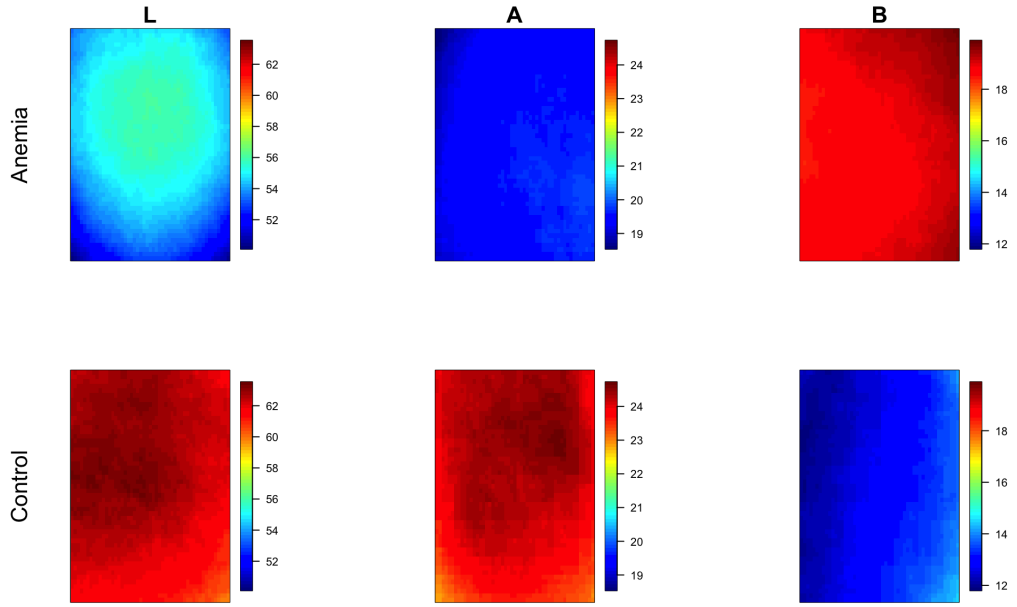

Figure A0.7: Heatmaps for the mean color data at each pixels for anemia and control groups in CIELAB color space.