Universiteit
Leiden
The Netherlands

Teaching Machines to See Seasonal Color: An Explainable Approach for Automated Personal Color Analysis

Natascha Dinh (s3990834)

Supervisors:
Hazel Doughty & Kaiting Liu

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 9, 2026

Abstract

Personal color analysis is the practice of finding color palettes that harmonize with an individual's natural look. It is traditionally implemented by color consultants, which requires significant time and effort. The best performing models for automated personal color analysis are convolutional neural networks (CNNs), however, this CNN-based approach suffers from limited interpretability and requires extensive annotation effort while only achieving moderate performance.

This thesis investigates whether a distance-based approach can provide greater explainability while remaining competitive with existing CNN-based models. Multiple design choices are systematically evaluated, including different distance-based algorithms (K-Means versus Nearest Centroid), structures (flat versus hierarchical), assignment strategies (hard versus soft), color spaces (RGB, LAB, HSV), feature engineering approaches, training data size, and labeled data requirements.

Experimental results show that Nearest Centroid consistently outperforms K-Means across most tasks. LAB and HSV representations achieve substantially higher performance than RGB, suggesting that personal color analysis relies primarily on the clear separation between chromatic information and luminance information. Furthermore, hierarchical classification yields better results than flat classification, supporting the hypothesis that decomposing the task into perceptually meaningful stages aligns well with the underlying structure of personal color analysis. The results also demonstrate that stable centroid estimation can be achieved using a relatively small amount of labeled data, reducing the need for extensive annotation. Increasing the amount of unlabeled training data, however, provides only limited performance improvements.

Overall, the proposed framework offers a computationally lightweight and interpretable alternative to CNN-based approaches. While achieving slightly lower accuracy than the best-performing deep learning models, it provides transparent intermediate reasoning steps and requires significantly less labeled data, making it a promising direction for future research in automated personal color analysis.

Contents

1 Introduction **1**

2 Research Question and Work Outline **2**

3 Preliminaries **3**
3.1 Background and Personal Color Theory . . . 3
3.1.1 Seasonal System - The 4 Basic Seasons . . . 3
3.1.2 Seasonal System - The 12 Color Wheel . . . 5
3.2 Distance-Based Models: K-Means and Nearest Centroid . . . 6
3.3 Color Spaces . . . 7
3.3.1 RGB . . . 7
3.3.2 HSV . . . 8
3.3.3 LAB . . . 9

4 Related Work **10**
4.1 Traditional Machine Learning Methods . . . 10
4.1.1 Rule-based Methods . . . 10
4.1.2 Machine learning classifiers . . . 11
4.2 Deep Learning Methods . . . 12
4.3 Research Gaps and Motivation . . . 13

5 Methodology **14**
5.1 Problem Definition . . . 14
5.2 Feature Extraction . . . 15
5.2.1 Facial region segmentation . . . 15
5.2.2 Color values extraction . . . 15
5.3 Feature Selection and Engineering . . . 17
5.4 Experimental Setup . . . 19
5.4.1 Color spaces . . . 19
5.4.2 Hierarchical versus Flat . . . 19
5.4.3 Soft versus Hard . . . 21
5.4.4 Low-annotation setting . . . 21
5.4.5 Data Size . . . 22
5.5 Baselines . . . 22

6 Dataset **23**
6.1 Data Collection and Processing . . . 23
6.2 Data Inspection . . . 24
6.2.1 Test set . . . 24
6.2.2 First training set: Personal Color Dataset . . . 25
6.2.3 Second training set: Extended Dataset . . . 26

7 Experiments and Results **27**
7.1 K-Means vs Nearest Centroid . . . 27
7.2 Color Spaces . . . 28
7.3 Hierarchical versus Flat . . . 31
7.4 Low-annotation setting . . . 33
7.5 Data Size . . . 35
7.6 Final Distance-Based Model versus CNN . . . 36
7.6.1 Accuracy . . . 36
7.6.2 Interpretability . . . 38
7.6.3 Data efficiency . . . 40
7.6.4 Feature Space Analysis . . . 41

8 Discussion **44**
8.1 Practical Implications . . . 44
8.2 Limitations . . . 45
8.3 Future Work . . . 46

1 Introduction

Personal color analysis is the practice of determining the color palette that complements and harmonizes well with an individual's natural skin tone, hair and eye color to be perceived as soothing and pleasing. It plays a crucial role in maximizing their natural beauty by coordinating clothing, makeup, and hairstyle. Besides serving aesthetic purposes, this practice also carries social impact, as clothing style and color can affect perceptions of competence and professionalism. This idea is supported by Neil Hester and Eric Hehman in their paper "Dress is a Fundamental Component of Person Perception"[4].

Personal color analysis is widely used in the fashion and entertainment industries. It is mostly conducted by professional stylists to enhance celebrities' public images and stage presence. Traditionally, it is done by directly draping a series of fabrics next to the individual's face to determine which color palette best complements them. These consultations often cost between 100 and 250 euros per session and last between 1 and 3 hours. Furthermore, although it has gained popularity in East Asia, it remains less publicly accessible in other regions. As a result, there has been growing interest in developing automated methods for quick, cheap and accessible personal color analysis using computer vision and machine learning.

Although personal color might seem straightforward in theory, as will be discussed in Section 3.1.1 and 3.1.2, in practice, it is much more challenging to determine a person's color palette due to the extremely subtle visual cues and subjective judgment. Factors such as lighting condition, background and even the person´s health can significantly affect how a person´s facial coloring is perceived. Furthermore, adjacent classes in personal color analysis always share one or more visual characteristics (Section 3.1.1 and 3.1.2), resulting in overlapping feature distributions. Due to the overlapping characteristics of classes, the task is not trivial, even for human experts. In fact, personal color analysis is inherently a subjective task, which relies mostly on expert perception and interpretation rather than universally accepted biological traits. Therefore, this study assumes that the seasonal classes reflect sufficiently stable expert judgements that can be modelled computationally rather than objective natural classes. The aim of this research does not seek to validate the actuality of seasonal categories, but to show if the classifications made by experts can be emulated by the machine learning algorithms.

One of the most remarkable studies in this field was carried out by Stacchio. His paper, Deep Armocromia: A Novel Dataset for Face Seasonal Color Analysis and Classification [19], targets the previously mentioned problems by developing a CNN-based model that automatically classifies facial portraits into seasonal colors. This model achieved the highest reported accuracy among existing approaches. Despite this effort, challenges remain. Firstly, the researchers and color experts had to go through a labor-intensive annotation process due to the lack of publicly available, well-labeled data. Secondly, since the backbones of this model are CNNs, it is not capable of providing details regarding its predictions, which might be valuable for users seeking deeper insight. Lastly, the accuracy achieved by this model is 31.8% for 12-subtype classification and 55.4% for 4-season classification , which is lower than expected for a model of this architectural complexity.

Among the limitations of CNN-based approaches, the lack of explainability is critical for personal color analysis. Due to the subjective nature of the task, it is usually not verifiable whether the classifications are strictly correct or incorrect. In this case, users may be particularly prone to be skeptical of the model's decisions, making explainability crucial for increasing user trust and acceptance. From a scientific point of view, explainability allows researchers to validate whether

the model is learning meaningful patterns that align with the principles of personal color analysis or just relying on spurious correlations. This is specifically useful in the presence of data bias, where the model may rely on irrelevant majority-class cues such as ethnic backgrounds, lighting conditions or makeup styles instead of relevant features such as skin tone, hair color and eye color. This enables researchers to assess the validity and generalizability of the results.

This thesis aims to address the gaps introduced above, not only by developing an explainable, distance-based model for personal color analysis, but also by investigating the impact of different design decisions, such as various architectural choices, training strategies, and data-related modifications.

2 Research Question and Work Outline

The primary goal of this thesis is to answer the following research question:

> *To what extent can explainable distance-based methods achieve competitive accuracy while improving interpretability and data efficiency?*

In order to investigate the main research question, several sub-questions are formulated:

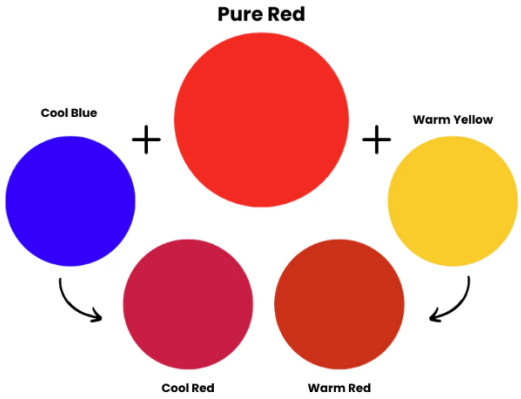
> *Which algorithm among **K-Means** and **Nearest Centroid** achieves the best accuracy and robustness for seasonal color classification?*
>
> *What is the impact of different color spaces (**RGB vs. LAB vs. HSV**) on model performance?*
>
> *How does **hierarchical clustering** compare to **flat clustering** for personal color classification?*
>
> *What is the impact of **the amount of annotated data** on Nearest Centroid performance?*
>
> *What is the impact of **training data volume** on K-Means performance?*

The remainder of this thesis is structured as follows to address the research questions mentioned above. In the Preliminaries section, fundamental concepts required to understand the rest of this study is introduced. It covers the theory of personal color analysis, distance-based techniques and color spaces. The Related Work section reviews previous studies on machine learning strategies for personal color analysis. The study design is discussed in the Methodology section. The Experiment and Result section follows with the experimental outcomes. Finally, the Discussion section offers a general interpretation of the findings, discusses the limitations of the study and suggests future research options.

3 Preliminaries

3.1 Background and Personal Color Theory

In the twentieth century, Swiss painter Johannes Itten laid the foundation for personal color by developing several theoretical principles about color harmony. He introduced the concepts of warm versus cool hues and light versus dark contrasts in colors, shown in Figure 1 [5].



(a) Warm and cool hues of the same color. Source: [20].

(b) Low-, medium-, and high-contrast colors. Source: [17].

Figure 1: Illustrations of key concepts in Johannes Itten's color theory.

His concepts paved the way for later work in personal color analysis, as it fundamentally treats a person as a work of art, creating visual harmony between them and their clothing. Subsequently, American artist Robert C. Dorr expanded Itten's work by exploring the impact of skin undertones on color harmony and developed a system that could differentiate cool and warm undertones [7]. Lastly, Carole Jackson projected Itten's and Dorr's concepts onto a seasonal color wheel system, which is used today in personal color analysis [6].

3.1.1 Seasonal System - The 4 Basic Seasons

The seasonal system, shown in Figure 2 is the most popular framework for personal color analysis. It categorizes people into 4 main seasons, determined mainly by 2 factors: the undertone (warm versus cool) and the level of contrast (light versus dark) in one's facial coloring.

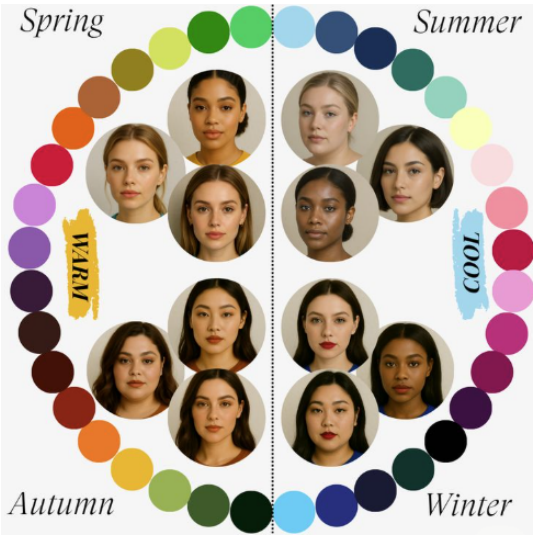


Figure 2: Illustration of the 4 primary seasonal color categories. Source: [13].

A "Spring" person has a warmer skin tone and a low contrast facial coloring. They often have lighter skin and hair with either a peach or golden undertone. Their eyes are often light blue, green or hazel. A color palette that blends well with Spring individuals includes light, clear and warm colors, such as coral, mint and turquoise.

Summer represents people with a cooler undertone and a low contrast facial coloring. Similar to the Spring type, their skin and hair are on the lighter side, however, they typically have a lack of warmth in their coloring, with ashier hair colors and rosier skin tones. Colors that enhance the Summer type are typically soft, cool, and muted, such as mauve, dusty pink and dove gray.

The Autumn type is characterized by a warm skin tone and a high contrast facial coloring. "Autumn" people often have deeper hair color with copper shades, such as dark brown and auburn. Their skin ranges from fair to deep with golden or olive undertones, and their eyes are commonly dark. Warm and earthy colors such as chocolate brown, olive green, and terracotta elevate their featural depth.

Lastly, Winter defines individuals with a cool skin tone and a high contrast facial coloring. The skin tones can range from light to dark with undertones of either pink or blue. The hair color typically ranges from dark, cool brown to black. "Winter" people often have a cool eye colors such as black, blue or gray. The Winter color palette consists of cold and bold colors, for instance red, royal blue, black and white, exhibiting maximum contrasts.

Season	**Undertone**	**Contrast Level**
Spring	Warm	Low
Summer	Cool	Low
Autumn	Warm	High
Winter	Cool	High

Table 1: General Characteristics of the 4 Seasonal Categories

3.1.2 Seasonal System - The 12 Color Wheel

According to Carole Jackson, the categorization of personal color is determined by two primary dimensions of color: contrast (low to high) and undertone (warm to cool) and each seasonal type is a combination of these two dimensions. However, this system is considered to be oversimplified as it does not take the continuous nature of these underlying dimensions into account. There exist individuals who fall close to the boundary between seasons. Under those circumstances, the color palette that belongs to the individual is a combination of the color palettes from the 2 adjacent seasons. Therefore, categorizing those individuals in a single season may not be entirely valid.

To tackle these limitations, the 4-season system is expanded into a 12-season system, shown in Figure 3. This expanded framework introduces a third classification factor to provide a more precise differentiation within each season. It is determined by the most prominent characteristic of an individual, which typically falls into one of six categories: light, deep, warm, cool, clear, or soft.

The Light subtype corresponds to people with light features, such as light hair, light skin and light-colored eyes. This subtype is only applicable for the seasons with low contrast, which are Spring and Summer. Conversely, the Deep subtype is applied for the individuals with darker features and is only adapted to the seasons with high contrast, which are Autumn and Winter. The Warm subtype is characterized by a richer warmth to the skin, hair and eyes. It is therefore only applicable for the warm seasons, Spring and Autumn. On the other hand, the Cool subtype is only applicable for the other two cool seasons and refers to individuals with a dominance of cool hues to their features. The Clear subtype describes people who have a strong, vivid coloring to their features, such as bright blue eyes or bright red hair. It can be applied to Spring and Winter individuals. Lastly, the Soft subtype is characterized by a muted coloring with more grayish tone and is applicable for Summer and Autumn.

Since both contrast and undertone exist on a continuum, the 12-subtype system can provide much greater precision in its recommendations than the traditional 4-season system. If a person falls between two seasons, it would be beneficial for them to think about their primary season as well as their secondary characteristic when selecting colors that will complement their natural features.

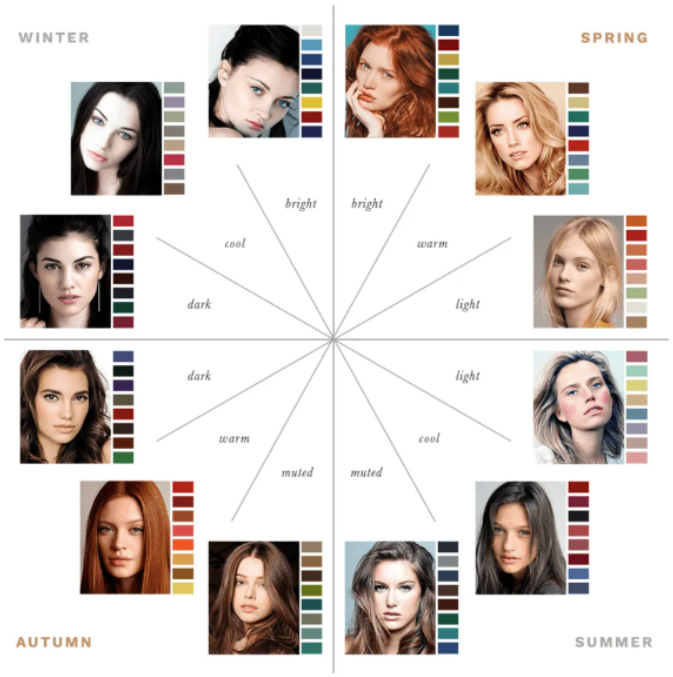


Figure 3: In the 12-Subtype Color Analysis Wheel, each season is further divided into three subtypes reflecting variations in warmth, coolness, lightness, softness/mutedness, clearness/brightness, and depth/darkness. Source: [10].

3.2 Distance-Based Models: K-Means and Nearest Centroid

Distance-based models perform classification or clustering on the data points based on the distance between them. The more similar the data points, the closer they are to each other in a feature space and the more likely they belong to the same group. Distance-based models are therefore effective in continuous feature spaces, specifically color representations. Euclidean distance is one of the commonly used distance metrics in distance-based methods, it is also used as the sole distance metric throughout this study. This study considers two distance-based models, which are K-Means and Nearest Centroid. Since these models are fundamentally defined in terms of Euclidean geometry in their default configuration, the selection of Euclidean distance as the distance metric is standard. It measures the straight-line distance between two data points in a continuous feature space, making it a simple and intuitive choice, which does not require specific assumptions about the structure of the data. The use of one consistent metric also helps in preventing additional variability in distance definitions from being introduced, allowing fair comparison between methods.

K-Means is an unsupervised, distance-based clustering model. It groups similar data points into clusters without any prior knowledge of the true labels. It is suitable for personal color analysis, since colors can be represented as vectors, and similar colors are close to each other in the color space. Since it is unsupervised, the time and cost to annotate the data can be reduced. Furthermore, the number of clusters K is already known. However, K-Means models assume the clusters to be spherical, therefore, they only work well if seasonal groups are indeed compact and geometrically separable.

Nearest Centroid is a distance-based supervised classification method. The training phase

consists of only one step, that is, to compute the positions of the class centroids based on the true labels of the training set. After training, during prediction, any given data point is assigned to its nearest class centroid. Similarly to K-Means, it can work with color vectors in different color spaces. In contrast to K-Means, which assumes that the data has a natural class structure that can be discovered, Nearest Centroid assumes that each class already has a meaningful centroid, which should be a good summary of the entire class. In other words, while K-Means is more data-driven, Nearest Centroid is more label-driven. Although Nearest Centroid does not explicitly make the assumption that the clusters are spherical, it implicitly favors compact and isotropic clusters by relying on the Euclidean distance to the class centroid. One disadvantage of this method is that it is supervised, which means that the training set should be fully labeled. Nevertheless, it is still possible to compute centroids with a smaller set of labeled samples, therefore, the effort to annotate the data can still be reduced. From this perspective, Nearest Centroid provides a way to determine how large the labeled set should be to produce the desired results, as its performance depends heavily on the representativeness and size of the labeled training set.

3.3 Color Spaces

Color spaces are mathematical, numerical representations of colors. It is crucial for this study since distance-based methods are not capable of processing raw images like CNNs. Therefore, those raw pixels need to be translated into explicit numerical vectors. Different color spaces provide different feature vectors and the distances between them. Therefore, the choice of color space directly influences the quality of the resulting clusters. This study experiments with three color spaces: RGB, HSV and LAB.

3.3.1 RGB

RGB stands for red, green and blue and is the most well known color space. It creates colors by combining light from the three colors with different intensities. The intensity of each color channel ranges from 0 (total blackness) to 255 (maximum brightness). When images are stored in RGB, each pixel is translated into a vector of RGB values.

$$\mathbf{x}_i \in R^3, \quad \mathbf{x}_i = [r_i, g_i, b_i]^\top$$

where $\mathbf{x}_i$ denotes the feature vector of the i-th pixel, and r_i, g_i, and b_i correspond to the red, green, and blue channel intensities respectively.

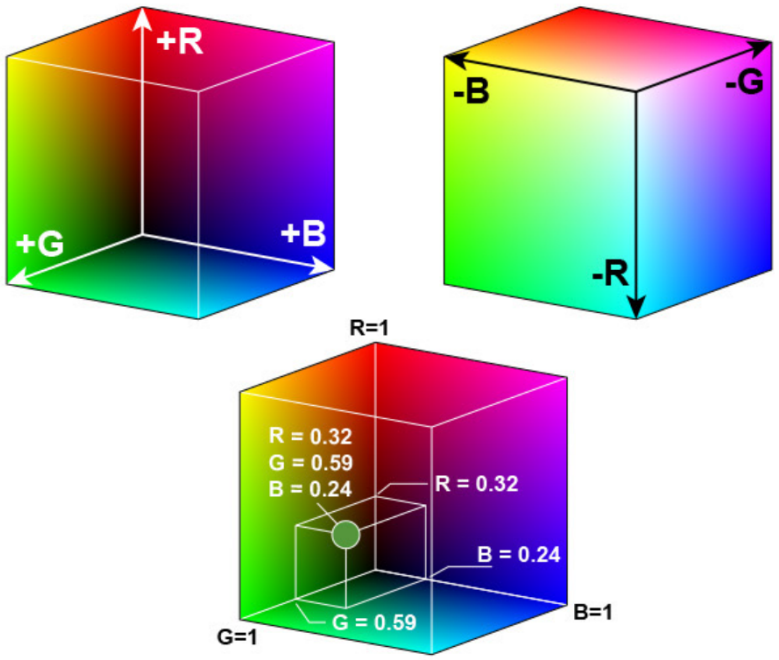


Figure 4: RGB color space as a 3D cube using normalized channel values from 0.0–1.0 instead of the standard 0–255 RGB scale. The axes represent increasing Red, Green, and Blue intensity. Source: [11].

RGB combines light colors and therefore works well in displaying images on screens, since screens emit light. Theoretically, personal color analysis relies on both brightness and chroma. Brightness helps determine contrast and chroma helps determine skin undertone and vividness of one's facial coloring. Therefore, both lighting and chromatic information need to be extracted separately in order to determine the overall season or subtype of a person. However, in RGB, brightness and chroma are mixed across channels, making seasonal structure harder to isolate. It is therefore solely device-oriented and does not align with how a human perceives colors. This makes it less suitable for personal color analysis where the annotation is done by human experts.

3.3.2 HSV

HSV stands for hue, saturation and value. Hue represents the type of the color, ranging from 0° to 360. By rotating around the wheel, different base colors are selected, for instance, red is at 0°, green is at 120° and blue is at 240°. Saturation represents the vividness or intensity of the color, ranging from 0% to 100%. At 0% saturation, colors are essentially on a black-and-white scale, and at 100%, colors appear at their highest vibrance. Value represents the brightness of the color, it also ranges from 0% (black) to 100% (white). Similarly to RGB, when images are stored in HSV, each pixel is translated into a vector of HSV values.

$$\mathbf{x}_i \in R^3, \quad \mathbf{x}_i = [h_i, s_i, v_i]^\top$$

where $\mathbf{x}_i$ denotes the feature vector of the i-th pixel, and h_i, s_i, and v_i correspond to hue, saturation and value respectively.

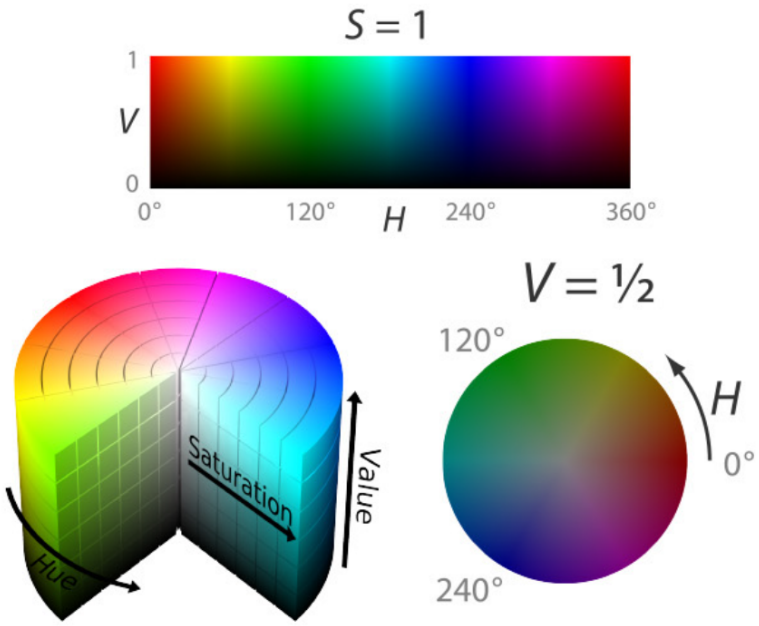


Figure 5: HSV Color Space Representation. Hue (H) determines the color type and is represented as an angle from 0°–360°, Saturation (S) controls color intensity from grayscale to fully vivid color, and Value (V) controls brightness from black to full brightness. Source: [15]

HSV isolates lighting information in the V-channel from the chromatic information in the H- and S-channel. Therefore, chromatic information extracted in HSV is less sensitive to lighting conditions than in RGB. The S-channel is possibly a good indication for vividness which is crucial for the determination of subtypes, while the H-channel may work well for the determination of skin undertones. The V-values can be extracted from different facial features to determine the variations in lightness and conclude if the individual has low or high contrast. These channels are more effectively separated and align more directly with the theoretical seasonal descriptors discussed in Section 3.1. This reduces inter-channel interference and improves the clarity of feature representations, which makes it more suitable for personal color analysis.

3.3.3 LAB

LAB is specifically designed to mimic human color perception. The L-channel represents luminance. It is comparable to the V-channel in HSV, where its value ranges from from 0% (black) to 100% (white). The A-channel represents the color gradient from green to red, its value ranges from -127 (green) to +128 (red). The B-channel represents the color scale from blue to yellow, ranging from -128 (blue) to +127 (yellow). When images are stored in LAB, each pixel is translated into a vector of LAB values.

$$\mathbf{x}_i \in R^3, \quad \mathbf{x}_i = [l_i, a_i, b_i]^\top$$

where $\mathbf{x}_i$ denotes the feature vector of the i-th pixel, and l_i, a_i, and b_i correspond to the values in L-channel, A-channel and B-channel.

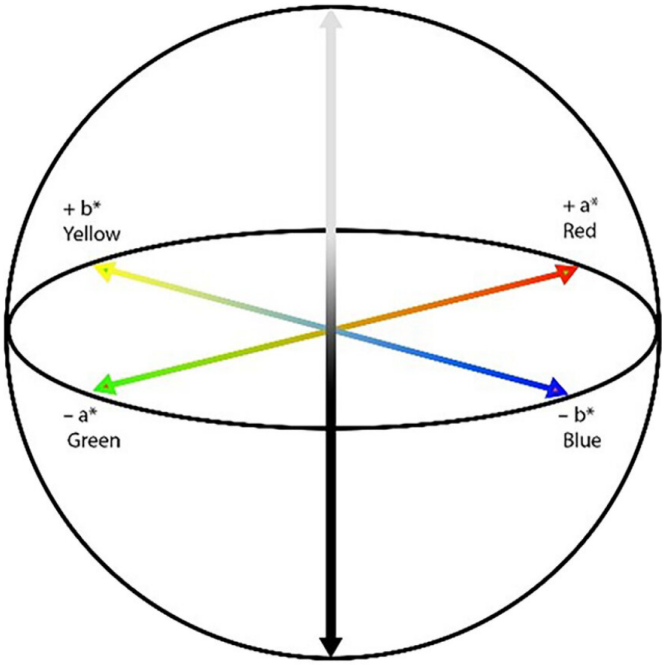


Figure 6: LAB Color Space Representation.The vertical axis (L) represents lightness from black to white, while the horizontal chromatic axes describe color opponency: +a indicates red, a green, +b yellow, and b blue. Source: [16]

Similarly to HSV, LAB also stores lighting information in one channel, separating it from the chromatic information. It is therefore also insensitive to lighting variations. Both A- and B-channel are helpful for determining the skin undertones, while the L-values can be used to determine the contrast. Again, LAB provides a more disentangled representation of color compared to RGB, enabling more effective separation of relevant characteristics. It is therefore more suitable for personal color analysis.

4 Related Work

In this section we review the existing works in automated personal color analysis, consisting of rule-based methods, traditional machine learning methods and deep learning methods. The final part focuses on identifying the research gap in these approaches and highlighting the need for a more flexible and explainable approach.

4.1 Traditional Machine Learning Methods

4.1.1 Rule-based Methods

Park et al. (2018) [12] propose a rule-based method to determine a person's skin undertone. They begin by isolating the skin region from the facial-landmark using a Gaussian Mixture Model that distinguishes skin regions from non-skin regions based on color distribution. Next, they convert all skin pixels from the device-dependent RGB space to the perceptually uniform LAB space. As discussed in Section 3.3.3, the larger the B-value, the more yellow the pixel is and the larger the

A-value, the more red the pixel is. Yellowness is strongly associated with warm skin undertone and redness with cool skin undertone. This fact is exploited in their study to develop a decision rule: if the average A-values of the skin pixels is larger than the average B-values, the person is classified as having cool undertone and vice versa. Overall brightness is represented by the L-channel and is not necessary for identifying the color temperature. While the approach is conceptually sound, due to a lack of labeled data, the authors have to rely on visual inspection and subjective ratings to evaluate the performance of the model instead of with a labeled test set. Evaluation metrics such as accuracy, precision, recall and F1 are therefore not computed and reported.

Another work by Alyoubi et al. in 2025 [1] also relies on the rule-based approach to classify skin undertones using wrist images, which aligns with how some human experts would carry out the task. They begin with by transforming the wrist images from RGB to LAB to enhance color differentiation. Thresholds in LAB are then applied to distinguish vein regions from skin regions. Since green veins are associated with warm undertones and blue veins are associated with cool undertones, the author define two LAB references, one for warm undertone which represents a dark green color ([70, 20, 40]) and one for cool undertone, which represents a dark blue color ([60, -20, -30]). If the computed average LAB values for the vein pixels is closer to the warm reference, then the person is classified as having a warm undertone. An accuracy of 70% for cool undertone detection and 80% for warm undertone detection is reported.

Although these studies may not have covered the contrast- and vividness layer in personal color analysis specifically, they still provided a foundation for skin undertone classification, which is an important component of the overall task.

In 2023, the first available implementation of 4-season personal color classification was published on the GitHub repository Colorinsight [14]. A dataset of Korean celebrity photos was used and partially labeled. The FaRL(Facial Representation Learning) is used for facial segmentation and isolation of the skin regions. They classify a person's season based on the Euclidean distance between RGB values from a randomly sampled skin pixel and predefined reference RGB values extracted from prior work. The image is assigned to the label corresponding to the closest reference color. This method results in low accuracy rates ranging from 20% to 30%, which is around the random baseline.

4.1.2 Machine learning classifiers

Besides rule-based approaches which rely on manually defined thresholds to assign a specific seasonal category to a person, there are also machine learning classifiers that can learn decision boundaries directly from the data, allowing them to adapt to the natural variations present in it. While rule-based personal color analysis enforces a rigid structure on a highly variable and subjective problem, machine-learning methods relax this by enabling the data to speak for itself. Patterns that naturally emerge from the data can be captured, while variability in images is automatically handled. Colorinsight [14] also trains a few other supervised classifiers using the RGB values of pixels .Table 2 presents the recorded accuracy values for each of the evaluated methods, which are slightly better than the accuracy of the rule-based method.

Method	Accuracy
Logistic Regression	35%
LDA (Linear Discriminant Analysis)	33.4%
KNN (K-Nearest Neighbors)	33.6%
Naive Bayes	31.3%
SVM (Support Vector Machine)	34.1%

Table 2: Classification accuracy of evaluated methods for personal color analysis

4.2 Deep Learning Methods

In addition to rule-based methods and classical classifiers, the authors of Colorinsight [14] also experimented with deep learning methods on the skin masks only. These methods include MobileNet, ResNet and Efficient Net. Among them all, ResNet with Adam optimizer performs the best with 56% accuracy, which is substantially higher than the accuracies of traditional machine learning and rule-based methods. The accuracy values of other evaluated deep learning models are presented below in Table 3.

Method	Accuracy
ResNet18	56%
MobileNetV3	36%
EfficientNet	27%

Table 3: Classification accuracy of evaluated deep learning methods for personal color analysis

Building on Colorinsight's implementation, Stacchio et al. (2025) [18] conduct a formal study with a well-defined methodology, providing detailed documentation of the experimental procedures and results. For training, the study use 3000 facial images sampled from CelebA dataset [9], which provides greater diversity in terms of facial attributes and ethnicity compared to a dataset limited to only Korean celebrities used in the work of Colorinsight. For testing, around 1900 facial images were collected from social-media. Both the training set and test set are annotated by trained analysts, following a strict protocol. Relevant facial features such as the skin, the hair and the eyes are extracted using the Facer deep-learning segmentation toolbox. The relevant generated masks were passed into three pre-trained visual backbones, which are FaRL-16, FaRL-64 and ResNeXt-50 to be transformed into numerical representations. These vectors are then passed into two trainable fully connected networks with ReLU activation, one maps them to the right season and the other one maps them to the right subtype. This study reported an accuracy of 55.4% for 4-season classification and 31.8% for 12-season classification. It is the only study that makes the dataset publicly available for research purposes.

Although explainability techniques for CNNs such as Grad-CAM, saliency maps, or SHAP do exist, the paper did not employ any of these techniques. The addition of explainability methods is also not mentioned in the Future Work section of this paper. It is understandable that explainability is not typically a central consideration in CNN-based approaches, since the existing post-hoc explainability methods offer limited practical benefit for end users. These techniques typically identify input regions that contribute most to the prediction, in this case, they can determine which pixels in the image contribute the most to a decision of the model. With this

pixel-level information, there is little room for the users to interpret or act upon. These techniques fail to provide relevant meaningful explanations related to personal color analysis, such as undertone, contrast, and vividness, and therefore have limited utility for user understanding and trust.

4.3 Research Gaps and Motivation

Despite the progress made in personal color analysis, several limitations remain in existing approaches.

Rule-based methods primarily rely on handcrafted thresholds, which is intuitive, interpretable and easy to implement. In some cases, these methods perform quite well, achieving up to 80% accuracy in binary classification for skin undertone. Assuming that contrast and skin undertone are independent components, combining both classifications could potentially yield an overall accuracy of approximately 64% for 4-season classification. Although this indicates that rule-based approaches has further potential when extended to additional feature layers, we need to take into consideration that rules are only effective when the problem is clearly defined and well separated. For instance, an accuracy of 80% is reached when the task is to detect the vein color, which is relatively obvious. Detecting the undertone from the facial skin is significant more challenging and subjective, which can reduce the effectiveness of these rules. Although Park et al. has made an attempt to apply a rule-based method on facial images, little to no research evidence of its effectiveness was provided. In theory, the approach is overly rigid and simplistic to work with variations in ethnicity, skin tone and lighting condition. Nevertheless, the findings demonstrate that LAB may be more suitable than RGB for capturing the chromatic information of the skin and predicting skin undertone.

Classical machine learning classifiers are more flexible than rule-based methods, since they can automatically learn decision boundaries and adapt to the real-world variations. However, when combined with low-level color space such as RGB, these methods only reached an accuracy ranging from 20% to 30% for 4-season classification. This limited performance can be attributed to using a non-perceptual color space that does not adequately represent features and hinders proper identification of subtle differences.

Deep learning techniques achieve significantly higher accuracies, since they are able to capture complex patterns in facial images, representing information beyond the 3-element structure of standard color spaces. Although these methods are rich in representation power, they are not without limitations. Firstly, they are computationally expensive. Deep learning models typically need a lot of data and training time, in addition to significant computational resources. Even though these models are architecturally complex, there is no direct correlation to high performance, as they only achieve a moderate accuracy ranging from 27% to 56% for 4-season classification. This raised questions about the efficiency–performance trade-off of deep learning approaches. Secondly, they are black boxes, which means that their explainability is nearly nonexistent. On one hand, this makes it difficult to improve accuracy, on the other hand, while these systems can predict the seasonal color of an individual, they are not capable of providing meaningful details regarding this prediction. Thirdly, the researchers and color experts had to go through a labor-intensive annotation process due to the lack of publicly available, well-labeled data.

Among the three types of approaches, traditional machine learning methods offer the most promising balance of performance, interpretability, and flexibility. Although previous related works have reported relatively modest accuracy for these traditional methods, this may be attributable to limitations in the chosen color representation - RGB - rather than the underlying algorithms

themselves. They have considerable potential when combined with the appropriate, perceptually meaningful color spaces. In this study, our focus is on traditional, distance-based machine learning methods. These methods retain the interpretability of rule-based approaches through explicit similarity comparisons between samples, while still benefiting from flexible data-driven learning. Compared to deep learning methods, distance-based approaches can reduce the need for fully labeled datasets. This results in a more transparent and cost-efficient framework. In order to compensate for their lower representational power of traditional classifiers in comparison to deep learning models, we incorporate perceptually meaningful color spaces such as HSV and LAB. Inspired by the rule-based methods, the proposed pipeline is structured into modular stages, where different aspects of the problem such as predicting the undertone and contrast, are handled in separate layers.

It is important to note that while each approach has its own unique benefits and suffers from its own limitations, they all face several common challenges. Since human perception is a major variable in the development of seasonal labels, the annotation of all data sets used is subjective. This leads to overlapping and non-strictly separable seasonal categories. These inherent ambiguities can never be perfectly solved by any computational approach, as the underlying class structure itself is not clearly defined. In addition, all approaches are sensitive to the conditions under which the image is obtained. These include camera settings, image quality, lighting condition and the presence of make-up. Although variations in lighting can be mitigated with the use of illumination-invariant color spaces such as HSV and LAB, their influence on the model is not fully eliminated. All factors mentioned above can significantly alter perceived color characteristics.

5 Methodology

At the global level, this study focuses on the comparison between CNNs and the proposed interpretable distance-based models. At the experimental level, internal ablation studies are conducted to evaluate the impact of different feature representations, distance-based algorithms and dataset configurations, on the proposed models. The aim is to identify which components contribute most to model performance to enhance it and close the performance gap with CNN-based models.

5.1 Problem Definition

The purpose of personal color analysis is to classify people by facial images to previously mentioned seasonal groups. Let us define the following notation:

$$X = \{facial\ images\}$$

$$Y = \left\{\ Spring, Summer, Autumn, Winter\ \right\}$$

$$Z = \left\{ \begin{array}{l} Bright\_Spring, Light\_Spring, Warm\_Spring, \\ Cool\_Summer, Light\_Summer, Soft\_Summer, \\ Deep\_Autumn, Soft\_Autumn, Warm\_Autumn, \\ Bright\_Winter, Cool\_Winter, Deep\_Winter \end{array} \right\}$$

The goal of personal color analysis is to produce the following mapping functions:

- $f : X \to Y$ predicts the most suitable season based on a given input facial image
- $g : X \to Z$ predicts the most suitable subtype based on a given input facial image

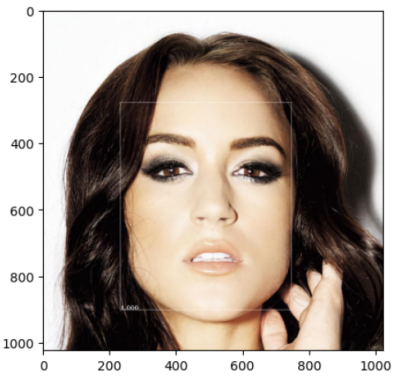
This study aims to approximate both mappings using various distance-based methods with various setups. It is important to note that the labels used for training are inherently subjective and ambiguous, which can introduce a degree of label noise.

5.2 Feature Extraction

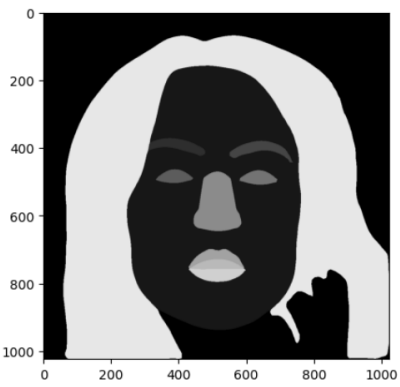
Feature Extraction is divided into 2 main stages: facial region segmentation and color values extraction, where facial region segmentation follows the same procedure as the previous study [19].

5.2.1 Facial region segmentation

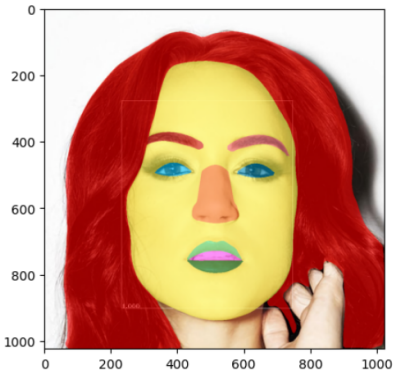
Facial region segmentation is employed with the Facer toolbox [2], which is a Deep Learning-based Module used for face detection and segmentation. We exploited this toolbox to segment the face and extract relevant features for personal color analysis, which are the skin region, hair region, and eye region.



(a) Face detection with the detected facial boundary highlighted by a bounding box,



(b) segmentation mask identifying the main facial regions, and



(c) extracted semantic regions including skin, hair, and eyes for personal color analysis.

Figure 7: Facial region segmentation using the Facer toolbox.

5.2.2 Color values extraction

After retrieving the relevant masks, pixel values of the masks are converted from RGB into LAB and HSV.In order to do this, RGB values are first normalized to the range $[0, 1]$:

$$I_{norm} = \frac{I}{255}$$

where:

- I is the original RGB image with pixel values in $[0, 255]$

- I_{norm} is the normalized image in $[0, 1]$

Each pixel in a region r is transformed from the normalized RGB space into LAB using the function (rgb2lab) from the skimage.color module:

$$X_r^{LAB} = RGB2Lab(X_r^{RGB})$$

where:

- X_r^{RGB} represents the set of normalized RGB pixels belonging to region r
- X_r^{Lab} represents the corresponding pixels in LAB space

Pixels are also mapped into the HSV color space using the function (rgb2hsv) from the skimage.color module:

$$X_r^{HSV} = RGB2HSV(X_r^{RGB})$$

where:

- X_r^{HSV} denotes the HSV representation of region r

For each region $r \in \{skin, hair, eyes\}$, a feature vector of 3 elements is computed by averaging pixel-wise values within the region.

$$\mu_r^{RGB} = \frac{1}{N_r} \sum_{i=1}^{N_r} X_{r,i}^{RGB}, \quad \mu_r^{Lab} = \frac{1}{N_r} \sum_{i=1}^{N_r} X_{r,i}^{Lab}, \quad \mu_r^{HSV} = \frac{1}{N_r} \sum_{i=1}^{N_r} X_{r,i}^{HSV}$$

where:

- r denotes the region (skin, hair, or eyes)
- N_r is the number of pixels belonging to region r
- $X_{r,i}^{RGB}$, $X_{r,i}^{Lab}$, and $X_{r,i}^{HSV}$ denote the i-th pixel in region r in the respective color spaces
- μ_r^{RGB}, μ_r^{Lab}, and μ_r^{HSV} are the resulting mean feature vectors for region r

Each raw RGB image is thus represented by 27 features: 9 mean values (RGB, LAB, HSV) per region multiplied by 3 regions (skin, hair, and eyes). The specific choice of mean over standard deviation or range is motivated by the need for a robust representation of facial color features. Given that facial segmentation may not be perfect, where a leakage between skin, hair, and background regions can be present, the use of standard deviation or range may introduce these segmentation errors since they are highly sensitive to outliers. Furthermore, in personal color analysis, the key signal is the overall facial coloring features and not the intra-facial variation. The mean is suitable for this task because it can capture the global chromatic tendency, while the standard deviation and range often capture texture, make-up and shadows, which are unwanted in personal color analysis.

Feature Group	Variables	Description
Skin RGB	skin_R, skin_G, skin_B	RGB channels of skin color
Hair RGB	hair_R, hair_G, hair_B	RGB channels of hair color
Eye RGB	eye_R, eye_G, eye_B	RGB channels of eye color
Skin HSV	skin_H, skin_S, skin_V	HSV channels of skin color
Hair HSV	hair_H, hair_S, hair_V	HSV channels of hair color
Eye HSV	eye_H, eye_S, eye_V	HSV channels of eye color
Skin LAB	skin_L, skin_a, skin_b	LAB channels of skin color
Hair LAB	hair_L, hair_a, hair_b	LAB channels of hair color
Eye LAB	eye_L, eye_a, eye_b	LAB channels of eye color

Table 4: Feature dictionary for color channels

Among these features, it is expected that skin-related features will be a significant factor in determining an individual's seasonal color, as they directly reflect the undertone and how vivid that undertone is. This already covers two out of the three steps required for personal color analysis. Specifically, the A- and B- channels of the skin are expected to have high predictive power for the undertone since they can capture the yellowness and redness of the skin. The S-channel of the skin is directly related to how saturated or vivid one's facial coloring is, and therefore may perform well in predicting vividness. In addition to the skin, hair and eye features are considered secondary cues to determine the contrast of an individual, which is the last step required in personal color analysis. Here, the L- and V- channels can play a crucial role in determining the difference in brightness between different facial features. The RGB features are expected to have the lowest predictive power due to their entangled representation of colors, which does not appropriately decouple perceptual dimensions that are relevant for personal color analysis.

5.3 Feature Selection and Engineering

In order to answer the sub-question concerning the predictive ability of the individual color spaces, feature selection is limited to the extracted color values in Section 5.2.2. As the feature set of each color space is relatively small, consisting of only nine elements, an exhaustive search over that specific color space to find the best feature combinations is computationally feasible. For the remaining sub-questions, instead of relying on one single color space, the optimal feature set is selected from the full pool of available features, ensuring full exploitation of the set. The fact that increasing feature dimensionality generally improves performance was supported by an additional experiment. We specifically conducted exhaustive search on combined color spaces: RGB and LAB, RGB and HSV, and LAB and HSV. The feature sets discovered from these pairwise color space combinations achieve significantly higher accuracy than those obtained from individual color spaces. However, the computational expense of performing exhaustive searches over combined color spaces becomes exponentially greater as the dimensionality increases. This makes it impossible to conduct and exhaustive search through the whole feature space of 27 features. Therefore, a greedy search strategy is introduced, where features are incrementally added based on their contribution to cross-validation performance.

Although greedy search is not optimal, it has the advantage of being significantly faster than exhaustive search, allowing it to work with high-dimensional feature spaces. To exploit this

advantage, a set of engineered features are added, resulting in a richer feature space.

Feature	Variable(s)	Formula	Definition / Why it can be useful
Hair warmth	`hair_warmth_lab`	$b_{hair} - a_{hair}$	Approximates hair warmth in LAB space. Positive values correspond to warmer red/yellow tones.
Skin warmth	`skin_warmth_lab`	$b_{skin} - a_{skin}$	Measures skin warmth and may help distinguish warm and cool undertones.
Eye warmth	`eye_warmth_lab`	$b_{eye} - a_{eye}$	Captures warm versus cool eye coloration.
Overall warmth	`overall_warmth`	$\frac{W_{hair}+W_{skin}+W_{eye}}{3}$	Provides a global warmth score across all facial regions.
Hair–skin lightness difference	`hair_skin_L_diff`	$L_{skin} - L_{hair}$	Measures brightness contrast between skin and hair regions.
Eye–hair lightness difference	`eye_hair_L_diff`	$L_{eye} - L_{hair}$	Captures brightness differences between eyes and hair.
Skin–eye lightness difference	`skin_eye_L_diff`	$L_{skin} - L_{eye}$	Represents contrast between skin and eye regions.
Overall lightness range	`overall_L_range`	$\max(L) - \min(L)$	Measures total brightness variation across all regions. Higher values indicate stronger contrast.
Hair chroma	`hair_chroma`	$\sqrt{a_{hair}^2 + b_{hair}^2}$	Represents hair color saturation in LAB space.
Skin chroma	`skin_chroma`	$\sqrt{a_{skin}^2 + b_{skin}^2}$	Measures saturation of skin coloration.
Eye chroma	`eye_chroma`	$\sqrt{a_{eye}^2 + b_{eye}^2}$	Represents eye color vividness.
Chroma ratio	`chroma_ratio`	$\frac{C_{hair}}{C_{skin}}$	Compares saturation levels between hair and skin. Relative differences may be more informative than absolute values.
Hair redness	`hair_redness`	$\frac{R_{hair}}{G_{hair}+B_{hair}}$	Measures the relative dominance of red intensity in hair color.
Skin redness	`skin_redness`	$\frac{R_{skin}}{G_{skin}+B_{skin}}$	Captures relative redness in skin tones.
Hair–skin hue difference	`hair_skin_hue_diff`	$H_{hair} - H_{skin}$	Measures hue contrast between hair and skin regions.
Normalized LAB features	`hair_a_over_L`, `hair_b_over_L`, `skin_a_over_L`, `skin_b_over_L`, `eye_a_over_L`, `eye_b_over_L`	$\frac{a}{L}, \frac{b}{L}$	Normalizes chromatic information by brightness to reduce sensitivity to illumination variations.
Normalized RGB features	`hair_G_over_L`, `eye_R_over_L`	$\frac{R}{L}, \frac{G}{L}$	Represents color channels relative to brightness rather than absolute intensity.

Table 5: Additional engineered features derived from color attributes.

The addition of these 23 engineered features results in samples with 50 features in total. The set of engineered features bridges the gap between raw color values and the perceptual dimensions that are relevant to personal color analysis. It incorporates the theoretical foundations by explicitly encoding higher-level descriptors of warmth, saturation/vividness and inter-regional contrast. Therefore, the resulting feature set aligns more closely with the criteria used by human experts when performing personal color analysis.

Based on principles of personal color analysis, among these features, it is expected that warmth-, chroma- and contrast-related features will be the most informative for classification. Specifically, skin warmth and skin redness directly measure how warm or cool toned a person is. Inter-region lightness differences, such as hair–skin lightness difference, skin–eye lightness difference and overall lightness range are expected to play a significant role in predicting the contrast level of a person. Chroma-related features such as hair chroma and skin chroma quantify color intensity and therefore are important for predicting the vividness of a person's facial coloring. Overall, it is hypothesized

that features that encode perceptual dimensions directly related to personal color analysis such as warmth, contrast and vividness will have the most predictive power.

5.4 Experimental Setup

The experiments are implemented with two distance-based methods: K-Means and Nearest Centroid. They are chosen for their interpretability, low complexity, the flexibility in handling different feature representations and the compatibility with the hierarchical pipeline. Both methods are implemented using the scikit-learn library. To assess which factors are most likely to reduce the performance gap with the CNN-based baselines, which will be discussed in Section 5.5, we define several experimental settings.

5.4.1 Color spaces

In order to evaluate the influence of color representation on model performance, we conduct an exhaustive search for each color space to find the best combination of features in that color space. A stratified 5-fold cross-validation was used to evaluate the feature sets within each color space. For each color space, the feature set that yields the maximum average cross-validation score is selected to represent that color space. This maximum average cross-validation score itself is used to evaluate the effectiveness of the different color spaces directly instead of testing them on the test set. This approach is chosen since cross-validation is a more robust metric to evaluate the generalization capability of the model with multiple train-test splits. Moreover, the goal at this stage is not to optimize the final model, but instead to isolate and analyze the individual effect of each color space to identify the most promising one. Since a final comparison with CNN-based models is not made during this stage, evaluating the models on the final test set was considered unnecessary. This progress is performed independently for both K-Means and Nearest Centroid. This allows for a fair comparison of the impact of color spaces on different algorithms.

5.4.2 Hierarchical versus Flat

This study will explore both hierarchical and flat approaches. These two approaches differ in how structure is imposed on the data.

A flat approach refers to methods that separate data into groups in one single step. For personal color analysis, this means that data is immediately grouped into 4 seasons, or 12 subtypes, depending on the goal. This approach is simple and efficient, however, it assumes that there are no relationships between the groups, which does not hold for personal color analysis. Seasons, or subtypes are not entirely independent, they share gradual transitions along different perceptual dimensions, such as skin undertone, contrast and vividness. For instance, the Spring and Summer groups both share the low contrast attribute, while Spring and Autumn share the warm undertone attribute. This suggests that similarities between data points exist at multiple layers. Therefore, a hierarchical approach might align better with this task.

In this paper, the term *hierarchical* does not refer to hierarchical clustering methods such as agglomerative clustering. Instead, it refers to the task-specific approach that follows a multi-layer pipeline, enabling it to be progressively refined with each layer. For personal color analysis task, in the first layer, instances are separated into warm-toned and cool-toned groups. In the second

layer, these two groups are further divided into low- and high contrast clusters. The two layers, therefore, form four seasonal groups: Spring, Summer, Autumn and Winter. In the last layer, each season is further refined into three subtypes based on vividness. This approach preserves the relationships between seasonal groups instead of flattening them. Moreover, this decomposition of the task aligns well with human expert decision processes, which may lead to improved accuracy on the human-annotated test set. This is motivated by the fact that the annotations in the test set themselves are produced based on similar layers of perceptual cues such as undertone, contrast and vividness. Therefore, models that are explicitly built upon these perceptual decompositions align better with the underlying annotation strategy, and might score better on the human-annotated test set.

Compared to a flat approach, a hierarchical result can provide explainable intermediate grouping results. While hierarchical classification could also be implemented with CNN-based models , it is not considered in this study. The main goal of this study is not to further enhance CNN architectures, but rather to determine if domain knowledge derived from personal color theory can be explicitly incorporated within an interpretable classification system. Our goal is to understand how perceptual attributes, based on theory, contribute to prediction by focusing on a feature-based and explainable approach. A CNN would shift the focus away from the interpretability and domain-driven nature of the proposed method.

The efficacy of the hierarchical approach is assessed by comparing it with the flat approach for 4-season and 12-subtype classification tasks.

K-Means

For hierarchical K-Means, classification is performed in stages. First, a K-Means model is trained to group skin undertone (warm vs cool). Second, a separate K-Means model is trained to group contrast (high vs low). Each model operates on its own independently selected feature set. The true class labels of the samples allocated to each cluster are collected to determine the final labels of the cluster through majority voting. The outputs of these two layers are combined to produce the final seasonal clusters. During inference, a test sample is assigned to the nearest cluster centroid, and it inherits the majority label associated with that cluster. This approach is compared to a flat clustering baseline, where a single K-Means model directly group the data into four seasonal clusters.

For the 12-season classification task, a third layer that predicts vividness is implemented and two hierarchical pipelines are evaluated. The first consists of a 2-layer hierarchy, where classification is performed sequentially from flat season category to subtype. The second employs a 3-layer hierarchy, where classification proceeds from undertone to contrast to hierarchical season and finally to subtype. To ensure valid subtype predictions, the final layer is applied within each previously assigned seasonal cluster. Specifically, 4 different K-Means models are trained on 4 different seasonal groups to further divide each of them into 3 subtypes. The final subtype classification is obtained by combining the outputs of all stages. This hierarchical approach is compared against a flat clustering model that directly predicts all 12 subtypes using a single feature set optimized via cross-validation. The same majority voting, feature selection, and inference stage as 4-season classification are used.

Nearest Centroid

The hierarchical Nearest Centroid model follows the same structure as the K-Means approach, including the layers, the feature selection and how it ensures valid subtype classification. One difference from the K-Means model is that Nearest Centroid does not require a majority voting step, as class centroids are computed directly from the true labels during training. During inference,

each test sample is assigned to the nearest explicitly defined class centroid.

5.4.3 Soft versus Hard

Within the hierarchical framework, during the inference phase, we evaluate and compare the robustness of hard assignment versus soft assignment.

Hard prediction refers to deterministic assignment of the sample at each level of the hierarchical structure. For both hard K-Means and hard Nearest Centroid, during the inference phase of each layer, samples are assigned to the cluster with the closest centroid. This creates a clear binary decision. However, with the hard prediction, in cases where a sample lies close to the decision boundary, an incorrect assignment occurring early in the hierarchy can result in irreversible propagation of classification error to subsequent stages, reducing the model performance for borderline cases.

Soft prediction refers to a probabilistic assignment of samples to each class based on the distance between them and the class centroid. It is computed by converting distances into probabilities using an inverse-distance function, followed by normalization:

$$p_{i,j} = \frac{\frac{1}{d_{i,j}+\epsilon}}{\sum_{k=1}^{K} \frac{1}{d_{i,k}+\epsilon}} \tag{1}$$

where $d_{i,j}$ represents the distance between sample i and centroid j, and $p_{i,j}$ denotes the probability that sample i belongs to cluster j and ϵ is a small constant added to avoid division by zero. The inverse-distance mapping ensures that samples that are relatively closer to the centroid will receive higher probabilities compared to samples that are farther away from the centroid. The normalization ensures that the sum of probabilities over all clusters equals 1. These probabilities are propagated through the hierarchical structure, and at the final stage, class probabilities are determined by multiplying the contributions from each level. The predicted season or subtype is the one with the highest overall probability.

Not only does a soft approach better capture the continuous nature of skin undertone, contrast and vividness, it also provides a more informative representation by preserving uncertainty throughout the decision process, which can improve interpretability for users.

Although there are other alternative formulas for probability calibration such as softmax-normalized negative distance and Gaussian kernel functions, they require more parameters and thus further assumptions about personal color analysis. Since this study's primary goal is to develop an interpretable and lightweight distance-based framework, the inverse-distance function is chosen for its simplicity and compatibility with geometric interpretation of the feature space.

5.4.4 Low-annotation setting

This experiment is specifically conducted for the soft hierarchical Nearest Centroid model, as this model produced the best overall performance in our earlier experiments. Our aim is to identify the minimal amount of labeled data needed for centroid computation to achieve performance plateau for both 4-season and 12-subtype classification, where further increases in the number of labeled samples result in only marginal improvements in classification accuracy.

For this experiment, different subsets of the labeled dataset are used to compute the class centroids. For both 4-season classification and 12-subtype classification we used $n \in$

$\{10, 100, 400, 700, 1000, 1600, |trainset|\}$. In order to ensure balanced representation across classes, a stratified strategy by season or by subtype is used to sample n: for each season or subtype, a fixed number of samples is randomly selected without replacement. This stratified sampling is repeated across 10 seeds to ensure robustness.

After calculating the centroids using the labeled subsets, the same hierarchical soft assignment described in Section 5.4.2 and Section 5.4.3 is used to assess model performance on the test set across 10 seeds for each n.

5.4.5 Data Size

This experiment is conducted exclusively for the K-Means model, using the Extended Dataset, to be introduced in Section 6. Since this dataset is mostly unlabeled, it is not suitable to experiment with Nearest Centroid method, which relies on labeled samples to compute class centroids.

The labeled overlap between this dataset with the previous training set contains 331 instances, which guide the K-Means clustering process in the following two ways:

- **Method 1 - Majority Vote**: 331 labeled samples are used to assign labels to the clusters post-hoc using a majority voting scheme.
- **Method 2 - Centroid initialization**: 331 labeled samples are used to initialize centroid.

Hierarchical and Flat K-Means methods are employed on both datasets, the larger Extended Dataset and the smaller Personal Color Dataset (Section 6) for both 4-season and 12-subtype classification. All evaluations are performed using the same test set.

5.5 Baselines

The work Deep Armocromia: A Novel Dataset for Face Seasonal Color Analysis and Classification by Stacchio et al. [19] provides two CNN-based models, one is for 4-season classification and the other is for 12-subtype classification. With an accuracy of 55.4% for 4-season classification and 31.8% for 12-subtype classification, this method is currently recognized as the most accurate automated method in personal color analysis. A more detailed description of the model architecture and training procedure can be found in Section 4.2. Both of these models are used as the benchmark models for this study. We directly use the reported results for the comparison with our proposed methods. This enables us to avoid retraining the CNN-based models from scratch, especially when there is insufficient implementation details mentioned in the paper about the use of seed value, loss function, normalization or the presence of early stopping. Furthermore, the previous paper utilized an ASUS ROG Strix G814JI with an Intel Core i9-13980HX CPU and an NVIDIA GeForce RTX 4070 (8 GB VRAM), which was not accessible in this study.

To ensure a valid comparison between the baselines and the proposed models, an identical annotated test set was used to measure the performance of all models. Except for the evaluation of the impact of training data size, all other experiments are conducted using the same training set as the previous study. The evaluation metrics are also kept the same as in the previous work, including accuracy, precision, recall, and F1-score. This guaranties that any difference in performance is due to the modeling approach, rather than variations in the evaluation method and data splits. Facial region segmentation is kept consistent with the previous work, while color value

extraction is adapted to downstream distance-based methods. This is motivated by the fact that the high-dimensional feature vectors produced by the CNN in the previous work is not compatible with classical distance-based methods, as the curse of dimensionality reduces the meaning of distances.

The comparison in this study is restricted to the above-mentioned models from prior work. The reason why other large-scale vision-language models such as Gemini are not considered is because they are trained on proprietary datasets and operate with undisclosed training pipelines. Due to this lack of transparency, they cannot be evaluated under controlled and fully reproducible conditions, making direct comparison with the proposed approach methodologically inconsistent. On the other hand, the CNN baselines in prior work are the only methods that are well-documented, making them reliable reference points for comparison. Furthermore, the addition of comparisons to other large-scale vision-language models would require significant computational resources, which fall outside the scope of this study.

6 Dataset

6.1 Data Collection and Processing

Two datasets of different sizes are collected. The first data set is retrieved directly from the previous study [19]. This dataset is used to evaluate the impact of the algorithm, color space, structure imposed on the data (flat vs hierarchical), and the amount of available annotated data. It consists of 4920 labeled images, annotated by annotators under the supervision of personal color experts. According to the author, this is the first and only publicly available labeled dataset for personal color analysis. This dataset is already split into a training set of 4000 images and a test set of 920 images. Throughout this study, the training set from this data set will be called the **Personal Color Dataset**. The test set will serve as a fixed test set for all experiments in this study to keep the evaluation fair and comparable.

Since we also need to investigate the impact of increased dataset size on K-Means accuracy and generalization capability, we have to obtain a second, different data set of greater volume. The second data set is a subset of CelebA-Mask-HQ [8]. Due to the time and computational constraints associated with the facial segmentation procedure, in which it took 20 seconds to segment one image, we could only manage to process 20,000 images. Although a larger dataset is generally preferable as it creates a bigger gap between limited-data and higher-data scenarios, which facilitates the assessment of the impact of dataset scale on model performance, 20,000 samples are considered sufficient to serve this purpose. Average perceptual hashing (aHash) is applied to detect visually similar images with the test set. These images are removed from this data set to prevent data leakage. In addition, to support a semi-supervised learning framework, overlap between this data set and the previously introduced training set is detected. This allows the retrieval of existing labels, which are then integrated into the large data set to guide the K-Means clustering process. Throughout this study, this dataset will be called the **Extended Dataset**

6.2 Data Inspection

6.2.1 Test set

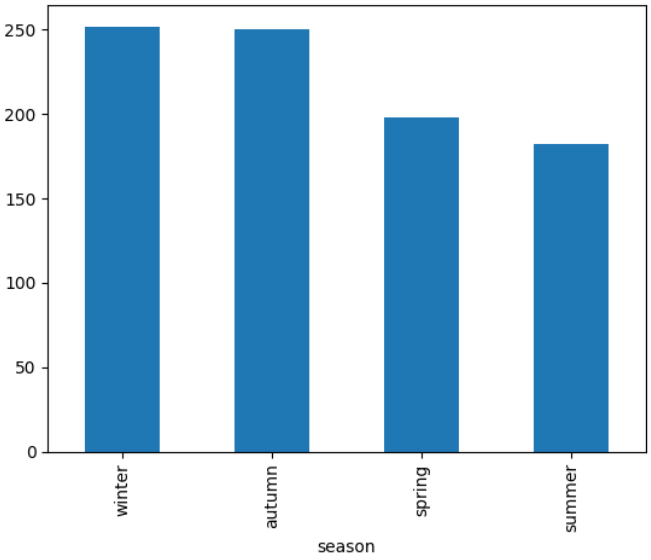


Figure 8: Distribution of seasonal classes in the test set.

An imbalance ratio of the test set is found to be 1.38, which means that the largest class (Winter) contains 1.38 times more samples than the smallest class (Summer). This is considered a mild imbalance.

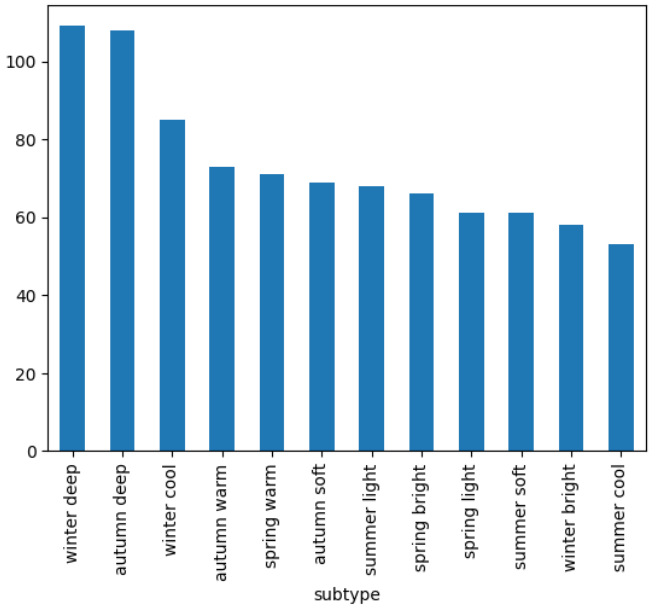
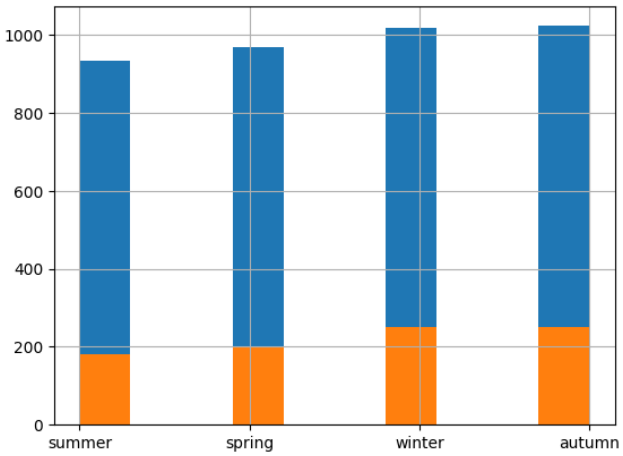


Figure 9: Distribution of subtype classes in the test set.

For the subtypes, an imbalance ratio of the test set is found to be 2.06, suggesting a moderate

class imbalance.

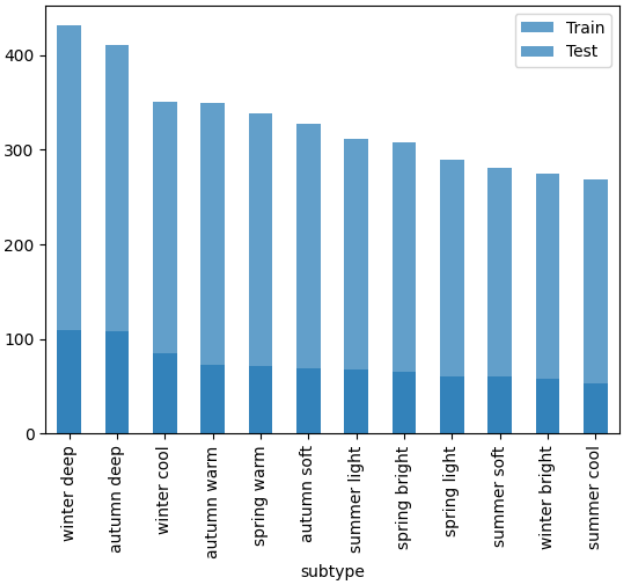
6.2.2 First training set: Personal Color Dataset



Season	Training Set	Test Set
Autumn	25.94%	28.34%
Spring	24.58%	22.45%
Summer	23.66%	20.63%
Winter	25.82%	28.57%

Figure 10: Relative class distributions of the seasonal categories in the training and test sets.

An imbalance ratio of the Personal Color Dataset for the 4 seasons is found to be 1.10, with the largest class being Autumn and the smallest class being Summer. This is also considered a mild imbalance. This mild imbalance prevents the evaluation metrics from being dominated by a single majority class. The relative class distribution in this set is broadly similar to that of the test set. Since there is no substantial distribution shift between training and test data, the evaluation is expected to be relatively fair across the four seasonal classes.



Season	Training Set	Test Set
Autumn Deep	10.93%	12.24%
Autumn Soft	7.88%	7.82%
Autumn Warm	7.13%	8.28%
Spring Bright	7.35%	7.48%
Spring Light	8.32%	6.92%
Spring Warm	8.90%	8.05%
Summer Cool	6.97%	6.01%
Summer Light	7.81%	7.71%
Summer Soft	8.89%	6.92%
Winter Bright	6.82%	6.58%
Winter Cool	8.57%	9.64%
Winter Deep	10.42%	12.36%

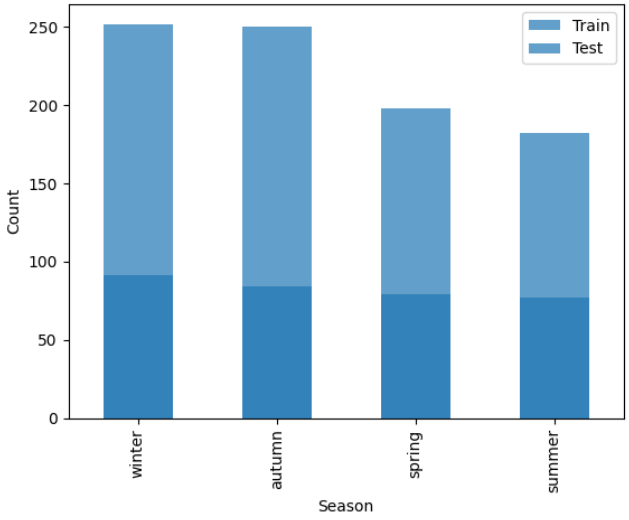
Figure 11: Relative class distributions of the subtype categories in the training and test sets.

An imbalance ratio of the Personal Color Dataset for the 12 subtypes is found to be 1.60, with the 2 largest classes being Winter Deep and Autumn Deep and the smallest class being Summer Cool. This indicates a moderate class imbalance across the subtype classification task. Majority classes such as Winter Deep and Autumn Deep are expected to achieve higher classification performance compared to minority classes such as Summer Cool.

6.2.3 Second training set: Extended Dataset

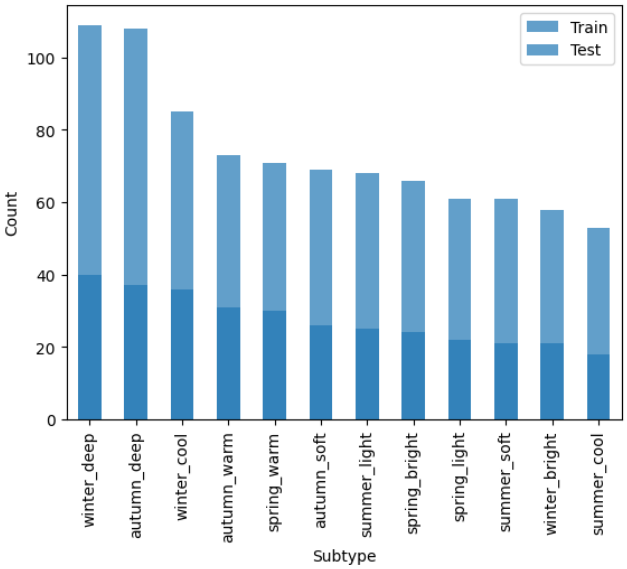
The Extended Dataset has not gone through the annotation process and therefore the majority of the labels is unknown. To help support the semi-supervised clustering framework, an overlap with the Personal Color Dataset was found, allowing us to retrieve 331 labeled samples to be used to guide the clustering process on this Extended Dataset. The performance of K-Means therefore depends heavily on the distribution of these samples.

Figures 12 and 13 represent the class distributions of only the retrieved labeled subset. As shown, the seasonal categories remain relatively balanced, although mild subtype imbalance is present. In particular, Winter Deep and Autumn Deep are the most represented subtype classes, while Summer Cool and Winter Bright occur less frequently.



Season	Training Set	Test Set
Autumn	23.26%	28.34%
Spring	27.49%	22.45%
Summer	25.38%	20.63%
Winter	23.87%	28.57%

Figure 12: Relative class distributions of the seasonal categories in the training and test sets.



Season	Training Set	Test Set
Autumn Deep	7.55%	12.24%
Autumn Soft	9.37%	7.82%
Autumn Warm	6.34%	8.28%
Spring Bright	11.18%	7.48%
Spring Light	7.25%	6.92%
Spring Warm	9.06%	8.05%
Summer Cool	7.86%	6.01%
Summer Light	6.65%	7.71%
Summer Soft	10.88%	6.92%
Winter Bright	5.44%	6.58%
Winter Cool	6.34%	9.64%
Winter Deep	12.08%	12.36%

Figure 13: Relative class distributions of the subtype categories in the training and test sets.

7 Experiments and Results

7.1 K-Means vs Nearest Centroid

This experiment compares the performance of Flat K-Means and Flat Nearest Centroid under the impact of feature engineering using test and cross-validation accuracy for both 4-season and 12-subtype classification.

	Test Accuracy		Cross Validation Accuracy	
	Without Feature Engineering	With Feature Engineering	Without Feature Engineering	With Feature Engineering
K-Means	48.87	49.21	47.37	48.19
Nearest Centroid	51.81	52.49	51.74	51.69

Table 6: Test and cross-validation accuracies (%) for the 4-season classification using K-Means and Nearest Centroid classifiers, evaluated with and without feature engineering.

	Test Accuracy		Cross Validation Accuracy	
	Without Feature Engineering	With Feature Engineering	Without Feature Engineering	With Feature Engineering
K-Means	24.03	24.60	23.15	23.56
Nearest Centroid	30.73	28.68	29.22	28.40

Table 7: Test and cross-validation accuracies (%) for the 12-season classification using K-Means and Nearest Centroid classifiers, evaluated with and without feature engineering.

The results for the flat 4-season classification are shown in Table 6. For this task, the Nearest Centroid classifier consistently outperforms K-Means on both the test set and cross-validation sets, achieving a maximum test accuracy of 52.49% when feature engineering is applied. Table 7 shows the results for the flat 12-season classification task, which is considerably more difficult. Again, Nearest Centroid outperforms K-Means, achieving a maximum test accuracy of 30.73%, without feature engineering.

Nearest Centroid consistently produced better results than K-Means across all experimental settings. A possible reason for this is that while K-Means relies entirely on unsupervised clustering, Nearest Centroid benefits from the use of annotated data during centroid computation. This observation reveals that the target classes in personal color analysis are predefined categories rather than naturally occurring clusters. Since K-Means assumes convex cluster shapes, the resulting clusters are not necessarily aligned with the targeted seasonal classes. Nearest Centroid, while using class labels to define centroids, may create more meaningful and stable decision boundaries.

The fundamentally different ways in which K-Means and Nearest Centroid utilize the feature space also changes the impact of feature engineering on them. As demonstrated in Table 6, feature engineering generally seems to have a positive effect on model performance, especially for K-Means where both test and cross-validation accuracy increased. The impact of feature engineering is less significant for Nearest Centroid since there is a small improvement in test accuracy and a negligible decrease in cross-validation accuracy. This small decrease from 51.74% to 51.69% is more likely to be caused by statistical variation, rather than a reduction in model quality due to feature engineering. It appears more prominently in Table 7 that feature engineering works better on K-Means than on Nearest Centroid. Since K-Means discovers clusters purely from geometry, it is sensitive to feature space structure. Feature engineering directly reshapes the clustering landscape and therefore significantly alters cluster assignments. In contrast, since Nearest Centroid assumes that class structure is already defined by labels, and adding new features only shifts where each centroid sits slightly, feature engineering only leads to limited improvement. In other words, for K-means, feature engineering changes the global structure, while for Nearest Centroid, it only affects the local centroid positions.

Both tables show that the test accuracies are constantly slightly higher than the cross-validation accuracies, suggesting that the test set is slightly more optimistic than the cross-validation sets. Therefore, this confirms that training the model on cross-validation sets helps the models to generalize better and prevents them from overfitting.

7.2 Color Spaces

This experiment investigates the impact of different color spaces (RGB, HSV, and LAB) on the performance of both K-Means and Nearest Centroid for the individual layers of personal color analysis using cross-validation accuracy.

K-Means	Undertone	Contrast	Flat Season	Vividness	Flat Subtype
RGB	56.28	73.55	42.53	37.64	21.33
HSV	64.34	74.64	46.94	37.56	24.04
LAB	64.77	72.63	46.44	40.20	23.97

Table 8: Cross-validation accuracies (%) obtained with the K-Means classifier for the classification of individual color analysis layers (Undertone, Contrast, Flat Season, Vividness, and Flat Subtype) across different color spaces.

Nearest Centroid	Undertone	Contrast	Flat Season	Vividness	Flat Subtype
RGB	64.17	73.57	45.83	39.46	23.59
HSV	65.31	76.16	49.25	42.63	26.76
LAB	65.69	75.48	50.06	43.04	28.25

Table 9: Cross-validation accuracies (%) obtained with the Nearest Centroid classifier for the classification of individual color analysis layers (Undertone, Contrast, Flat Season, Vividness, and Flat Subtype) across different color spaces.

Table 8 and Table 9 show clear differences in performance between different color spaces. RGB generally produces the lowest accuracies across both classifiers. Although it performed relatively well for certain tasks in K-Means, such as Contrast and Vividness, it did not achieve the highest performance for any category. This suggests that the raw RGB representation is less suitable for separating perceptual color attributes in personal color analysis. An additional observation is that the performance of RGB in Nearest Centroid was better than in K-Means in almost all perceptual layers. This finding suggests that RGB data do not have a well-defined, separable structure for K-Means clustering. This poor performance of RGB is likely the result of brightness being embedded within the color channels, making it difficult to isolate and extract only the chromatic information or luminance information. The results suggest that personal color analysis relies primarily on the clear separation of chromatic information and luminance information. It relies on a layered structure of perceptual dimensions, such as undertone, contrast and vividness. These analytical layers are conceptually treated as independent layers, where contrast is determined by the overall difference in brightness, and skin undertone and vividness are determined by chroma. Therefore, it is necessary to isolate luminance from chromatic information to properly model these layers. Since RGB does not explicitly separate luminance from chromatic components, it may be less suitable for capturing the most appropriate representation for this task.

For both K-Means and Nearest Centroid, the performance of LAB is consistently strong for both the classification of Undertone and Vividness. LAB achieves the highest accuracies for Undertone (64.77% with K-Means and 65.69% with Nearest Centroid) and Vividness (40.20% and 43.04%). A possible explanation for the strong performance of the LAB color space is that LAB is able to separate luminance from chromatic information via its A- and B-channels, representing warm versus cool colored variations, respectively. The selected feature combinations in Table 10 shows

that for Undertone classification, the LAB-based models consistently rely heavily on these A- and B- channels, especially for K-Means. This implies that for unsupervised clustering, the A- and B-channels provide most dominant variance directions in the feature space for Undertone classification. When supervised information is incorporated in Nearest Centroid, it is shown in Table 11 that the model also benefits from the inclusion of L-channel. This indicates that brightness information also carries useful discriminative signals for this task.

K-Means	Undertone	Contrast	Season	Vividness	Subtype
RGB	skin_B, hair_B, eye_B	hair_R, eye_B	hair_R	hair_R, eye_B	skin_B, hair_R, hair_B, eye_G
HSV	skin_S, hair_H, hair_S	hair_H, hair_V, eye_S	hair_S, hair_V, eye_S	skin_H, hair_S, hair_V, eye_S, eye_V	skin_S, skin_V, hair_S, hair_V, eye_S
LAB	skin_a, skin_b, hair_a, hair_b	skin_L, skin_a, hair_L, hair_b, eye_L, eye_a	hair_L, hair_a, hair_b	hair_L, hair_a, hair_b, eye_L, eye_a	skin_a, hair_L, hair_a, hair_b

Table 10: Selected feature combinations that produced the highest cross-validation accuracies for each color analysis layer and color space using the K-Means classifier.

Nearest Centroid	Undertone	Contrast	Season	Vividness	Subtype
RGB	hair_R, hair_G, hair_B	hair_R, eye_B	skin_B, hair_R, hair_B, eye_B	hair_R, hair_B, eye_B	skin_B, hair_R, hair_G, hair_B, eye_R, eye_B
HSV	skin_S, hair_H, hair_S, hair_V	hair_H, hair_S, hair_V, eye_S	skin_S, hair_H, hair_S, hair_V, eye_S	skin_H, skin_S, hair_H, hair_S, hair_V, eye_S, eye_V	skin_S, hair_H, hair_S, hair_V, eye_S, eye_V
LAB	skin_L, skin_b, hair_L, hair_a, eye_L	skin_b, hair_L, hair_b, eye_L, eye_a, eye_b	skin_a, skin_b, hair_L, hair_a, hair_b, eye_L, eye_a	hair_L, hair_a, hair_b, eye_L, eye_b	skin_b, hair_L, hair_a, hair_b, eye_L, eye_a, eye_b

Table 11: Selected feature combinations that produced the highest cross-validation accuracies for each color analysis layer and color space using the Nearest Centroid classifier.

Furthermore, for Vividness classification, K-Means and Nearest Centroid Algorithms consistently

selected hair and eye features among the many channels available. This aligns with the fact that vividness is primarily related to chromatic richness and saturation rather than just skin color alone. It suggests that hair and eye features provide significant discriminative power in terms of color intensity detection since they indeed have more robust and stable chromatic properties than the skin.

HSV performs the best in Contrast classification for both K-Means and Nearest Centroid, with 74.64% in K-Means and 76.16% for Nearest Centroid. This indicates that the V-channel makes it easier to detect variations in brightness, which are the main contrast-related attributes. However, because the V-channel in HSV is comparable to the L-channel in LAB, given the strong performance of HSV in this task, it is possible that the V-channel may provide more accurate representations of contrast than the L-channel. Since contrast focuses primarily on differences in brightness between facial regions, the separation of brightness and chroma in HSV may help it better capture the relevant information. As shown in Table 10 and Table 11, both K-Means and Nearest Centroid selected hair features as part of their selection process. This suggests that hair provides substantial discriminative information for contrast classification, which is expected since human experts typically determine contrast by visually assessing how strongly hair color differs from the skin color. The repeated selections of hair V-value support this idea, as the V-channel captures lightness of the hair. The selection of hair Hue may initially appear less intuitive, however, it can suggest that different hues can indirectly contribute to perceived contrast through their interaction with surrounding facial features. Another feature that is also consistently considered when determining contrast is the saturation of the eyes. This also aligns with how human experts carry out the task, as saturated eyes can create greater contrast with surrounding skin regions.

Similarly to LAB in Undertone or Vividness classification, with HSV in Contrast classification, Nearest Centroid selects a broader combination of features. This again indicates that supervised information allows models to exploit more subtle distinctions in color properties that are not dominant enough to emerge naturally in unsupervised clustering.

Similarly to Table 6 and 7, Table 8 and 9 suggest that Nearest Centroid is more suitable for personal color analysis than K-Means, in all color spaces and perceptual dimensions. The tables also suggest that different perceptual layers require different color spaces and that the hierarchical approach may provide a more accurate prediction of the season and subtype than simply predicting them directly.

7.3 Hierarchical versus Flat

This experiment compares the impact of the hierarchical approach on K-Means and Nearest Centroid for both 4-season and 12-subtype classification under soft and hard assignment settings. The results of the hierarchical approach are compared against those of the flat approach.

	K-Means		Nearest Centroid	
	Soft	Hard	Soft	Hard
Flat	-	49.21	-	52.49
Hierarchical	51.47	51.59	53.97	53.97

Table 12: Comparison of flat and hierarchical classification accuracies (%) for the 4-season classification task using K-Means and Nearest Centroid classifiers under soft and hard assignment settings.

	K-Means		Nearest Centroid	
	Soft	Hard	Soft	Hard
Flat	-	24.60	-	30.73
2- layer hierarchical	23.58	25.74	29.67	29.97
3- layer hierarchical	27.10	27.22	30.50	30.84

Table 13: Comparison of flat and 2-layer hierarchical classification accuracies (%) for the 12-season classification task using K-Means and Nearest Centroid classifiers under soft and hard assignment settings.

The results in Tables 12 and 13 shows that the hierarchical approach slightly outperforms the flat classification approach. Table 12 shows the results for 4-season classification. For this task, both K-Means and Nearest Centroid achieve higher accuracies with the hierarchical approach. The improvement is particularly visible for K-Means, where the hierarchical model increases the hard accuracy from 49.21% to 51.59%. Nearest Centroid also benefits from the hierarchical structure, reaching 53.97% accuracy. For the 12-season classification task, shown in Table 13, the same pattern is visible, especially for K-Means. The 2-layer hierarchical model increases hard accuracy to 25.74%, and the 3-layer hierarchical model further improves it to 27.22%. Soft prediction shows a similar pattern, with the 3-layer hierarchy achieving the best performance at 27.10%. These results suggest that K-Means benefit more from the decomposition of the problem, since it provides guidance for creating more separable intermediate groups. In contrast, Nearest Centroid already performs relatively well in the flat setting with 30.73% hard accuracy, leaving less room for improvement through hierarchical decomposition. Even though the 2-layer hierarchical model performs slightly worse at 29.97%, the 3-layer hierarchical model reaches the highest hard accuracy of 30.84% and the highest soft accuracy of 30.50%. This indicates that a deeper and more fine-grained hierarchical structure aligns best with personal color analysis, even when the overall gain is modest. These results support our hypothesis that personal color analysis is inherently a hierarchical process.

In general, the hierarchical framework performs better than the flat framework. This implies that breaking the task into smaller consecutive decisions better reflects the structure of personal color analysis. Not only does the hierarchical framework allow the model to focus on more relevant features and reduce ambiguity and errors during classification, it also aligns with how human experts progressively identify characteristics such as warmth, contrast, and vividness of a person before reaching the final conclusion. The strong alignment between the proposed hierarchical pipeline and the steps used by human experts is likely a key factor contributing to its higher performance on the human-annotated test set. By effectively mirroring the decision-making process of human experts, the framework reached improved agreement with the labels.

The results further show that soft probabilistic assignment consistently performs slightly worse than hard assignment, particularly in the 12-season classification task. Several factors may explain this behavior. Even though soft predictions can theoretically preserve richer information, it can introduce ambiguity or noise in earlier stages. Soft prediction allows lower-confidence alternatives to influence downstream predictions in the hierarchical pipeline, which can influence it negatively, especially when confidence values may not accurately reflect the true likelihood of class membership. In contrast, hard prediction enforces a single decision at each stage, selecting only the strongest candidate, potentially reducing the impact of weak competing classes and the accumulation of uncertainty. For a complex task such as the 12-season classification task, where greater overlaps between visually similar subtypes are present, the negative effect of soft prediction is amplified due to more diffuse probability distributions.

7.4 Low-annotation setting

This experiment evaluates how the number of labeled samples used to initialize class centroids affects the performance of the soft hierarchical Nearest Centroid classifier for both 4-season and 12-subtype classification.

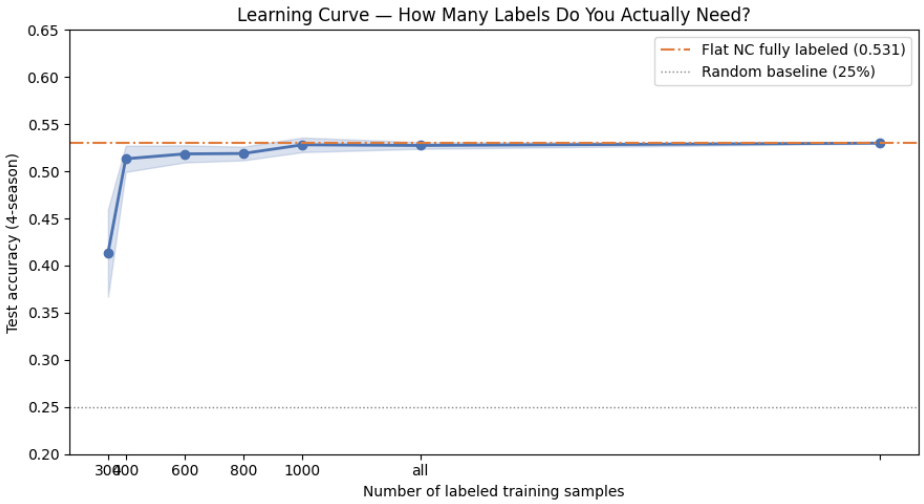


Figure 14: Effect of Label Availability on 4-Season Centroid Initialization of Nearest Centroid

As shown in Figure 14, the performance of Nearest Centroid are highly consistent, achieving accuracies of approximately 52%–53% across most numbers of labels evaluated, closely matching the fully labeled reference performance of 53.97%. Despite having a relatively small number of labeled samples, the model is able to converge to near-optimal performance quickly, indicating that approximately 100 labeled samples, which are 25 samples per class, are already sufficient to represent the underlying population distribution reasonably well. This suggests that the 4-season classification task has an easily identifiable underlying structure, and requires only a small amount of labeled data for accurate estimation of centroids. This model is therefore highly label-efficient for this task. The fact that Nearest Centroid performs significantly better than the random baseline of 25% shows that it does capture the inherent seasonal structure of the data set quite effectively.

Overall, the results show that Nearest Centroid can maintain a relatively high level of performance regardless of how much labeled initialization data is available for the 4-season classification task.

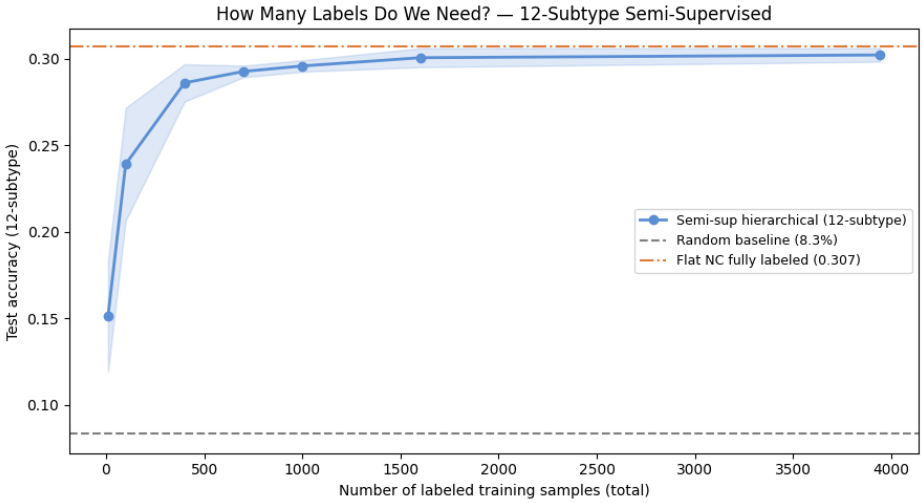


Figure 15: Effect of Label Availability on 12-Subtype Centroid Initialization of Nearest Centroid

The effect of Label Availability on 12-Subtype Centroid is shown in Figure 15. In comparison to the 4-season classification task, classification performance converges more gradually toward the fully labeled reference performance of 30.7%. This indicates that the degree of sensitivity of Nearest Centroid to the amount of labeled samples is much greater for the 12-subtype classification task than for the 4-seasonal classification task. This difference is expected since subtype classification involves a larger number of classes than season classification. Since the subtypes are more fine-grained, their feature distributions likely overlap more, making accurate centroid estimation more difficult when limited labeled data are available. As a result, the classifier requires a larger number of labeled samples to be able to create representative class centroids and establish reliable decision boundaries. After approximately 500 labeled samples, or 40 samples per class, performance tends to stabilize. Even though the results demonstrate that increasing label availability significantly improves the effectiveness of the Nearest Centroid method for fine-grained subtype classification, the method remains relatively label-efficient, as near-optimal performance is achieved without requiring the full labeled dataset.

The observed plateaus in Figure 14 and 15 can be explained by the fact that Nearest Centroid classification estimates the classes with the mean of the labeled samples. According to the Law of Large Numbers, the sample mean converges to the true mean as sample size increases. In this case, the larger the number of labeled samples available for computing the centroids, the more closely the empirical centroid reflects the true underlying class centroid. Therefore, after a sufficient number of annotated samples, additional labeled samples have a diminishing impact, as they contribute only marginal updates to the already well-estimated means. For 4-season classification, 25 annotated samples per class are sufficient to accurately represent the whole class, while for 12-subtype classification , this number is 40. Although the proposed framework significantly reduces annotation effort compared to fully supervised deep learning approaches, a minimum number of labeled samples is still required to sufficiently support the classification process.

7.5 Data Size

This experiment assesses the impact of an increase in data volume (the Extended dataset) on K-Means with the two proposed methods in Section 5.4.5.

Baseline K-Means	Personal Color Data	Extended Method	Extended Set
4-season Flat	49.21	4-season Flat method 1	48.75
		4-season Flat method 2	49.66
4-season Hard Hierarchical	51.59	4-season Hard Hierarchical method 1	52.49
		4-season Hard Hierarchical method 2	52.38
12-subtype Flat	24.60	12-subtype Flat method 1	21.09
		12-subtype Flat method 2	20.07
12-subtype 2-layer Hard Hierarchical	25.74	12-subtype 2-layer Hard Hierarchical method 1	22.00
		12-subtype 2-layer Hard Hierarchical method 2	20.29

Table 14: Performance comparison of K-means-based methods.

As shown in Table 14, classification performance does not improve consistently as the dataset is extended. For flat 4-season classification with a larger data set, method 1 decreases the accuracy from 49.21% to 48.75%, while method 2 only shows a small increase to 49.66%. It is visible that simply adding more unlabeled data will not result in a significant improvement in clustering quality, especially when the class boundaries are inherently difficult to separate.

For hierarchical 4-season classification, both methods show an overall increase in accuracy from 51.59% to approximately 52.55%. This indicates that the hierarchical structure can better exploit the increased amount of data than the flat approach. The improvement over baseline is modest but consistent.

In contrast, the 12-subtype classification task shows a decrease in performance with the larger dataset. For both flat and hierarchical 12-subtype classification, the decrease in accuracies is significant. This drop in performance likely reflects the complexity associated with the subtype classification task, as the addition of data may also introduce greater intra-class variation and overlap between neighboring subtypes.

In general, increasing the amount of training data did not lead to a significant improvement for K-Means. For 12-subtype classification, the results were even found to be worse than the baseline K-Means models. A possible explanation is the limited amount of only 331 labeled overlap samples available. Previous observations from the Nearest Centroid approach have provided a useful guideline that 500 labeled samples were required to reach a stable 12-subtype classification. While this threshold cannot be directly transferred to K-means due to differences between the algorithms, our results appears consistent with this observation. With 4-season classification, the 331 samples did not have any negative effect on the model performance since only around 100

labeled samples are required for this task. In this case, increasing the amount of training data does benefit the 4-class setting.

Another explanation for the observed results is the expansion of the dataset from 3,000 to 15,000 samples. It is likely that 3,000 samples are already sufficient to capture the underlying structure needed for the desired performance. When new unlabeled instances are added, they may provide a small amount of new information, which is not enough to significantly improve the models. In the case of 12-subtype classification, the additional information may not be sufficient to compensate for the limited amount of labeled data available, leading to diminishing performance. The results suggest that for personal color analysis, where boundaries are potentially overlapping, simply adding more data will not help with cluster separation. It may be more beneficial to add informative features and increase the quality and representativeness of the labeled data.

Table 14 also provides insight into the role of supervision in clustering. Method 1 assigns cluster labels after the clustering process, using majority voting based on 331 labeled samples. Generally, this method produces the best and most stable performance, which may indicate that even limited labeled supervision can substantially improve the interpretation of clusters. Because the labels are assigned after clustering, the clustering process itself remains unsupervised while still benefiting from external guidance. On the other hand, Method 2 begins with using labeled samples for initialization purposes. Although this can help the clustering process get off to a better start by providing more meaningful starting points, the results indicate that Method 2 is generally not as stable and tends to be more affected by the original distribution of the data. In some cases, such as for the hierarchical 4-season classification, Method 2 performs approximately as well as Method 1; however, for the much more complex 12-subtype problem, Method 2 performs considerably worse than Method 1. This indicates that the use of centroid initialization alone is not sufficiently effective in maintaining generally well-separated clusters when the classes are highly overlapping. Overall, the results suggest that limited supervised guidance through majority voting is more effective than initialization-based guidance, and that hierarchical classification benefits more from increased data than flat classification, particularly for simpler classification structures such as the 4-season system.

7.6 Final Distance-Based Model versus CNN

This section compares the performance of the best proposed classical classification approach against the CNN baselines. The primary objective is to evaluate whether interpretable classical methods can achieve competitive performance relative to deep learning approaches while requiring substantially lower computational complexity and training effort.

7.6.1 Accuracy

For the 4-season classification task, the best-performing classical approach was the soft hierarchical Nearest Centroid method, with the selected features shown in Table 16.

	Accuracy	Precision	Recall	F1-score
CNN-based	55.40	55.30	55.40	54.80
Hierarchical Soft Nearest Centroid	53.97	52.13	51.90	51.79

Table 15: Performance comparison between CNN-based and Hierarchical Soft Nearest Centroid models for 4-season classification

	Features
Undertone layer	hair_b_over_L, hair_a, skin_B, skin_a_over_L, eye_hair_L_diff
Contrast layer	hair_R, eye_S, hair_S, hair_b_over_L, skin_b, skin_eye_L_diff, skin_H

Table 16: Feature Configuration for the Nearest Centroid Hierarchical Soft Framework for 4-season classification

This feature configuration reached an accuracy of 70.18% on the first-stage undertone classification task (warm vs. cool) and 75.62% on the second-stage contrast classification task (high vs. low). When combining both stages into the final four-season prediction, the overall classification accuracy reached 53.97%. This performance is close to the CNN baseline accuracy of 55.4%, with a relatively small performance gap of 1.4 percentage point.

The results indicates that carefully engineered perceptual features based upon color theory already provide ample discriminative information for seasonal color analysis. In contrast to CNN-based models which rely on deep feature extraction, this approach utilizes domain-informed representations related to warmth, contrast and vividness. Despite its simplicity and low complexity, the method demonstrates competitive performance.

For the 12-subtype classification task, the best-performing classical approach was the hard hierarchical Nearest Centroid method, with the selected features for each layer shown in Table 18.

	Accuracy	Precision	Recall	F1-score
CNN-based	31.80	31.70	31.80	31.00
Hierarchical Hard Nearest Centroid	30.84	28.43	29.25	27.59

Table 17: Performance comparison between CNN-based and Hierarchical Hard Nearest Centroid models for 12-subtype classification

	Features
Undertone layer	hair_b_over_L, hair_a, skin_B, skin_a_over_L, eye_hair_L_diff
Contrast layer	hair_R, eye_S, hair_S, hair_b_over_L, skin_b, skin_eye_L_diff, skin_H
Vividness layer	hair_V, hair_S, eye_S, skin_B, hair_b, hair_B

Table 18: Feature Configuration for the Nearest Centroid Hierarchical Hard Framework for 12-subtype classification

For the more fine-grained 12-subtype classification task, the accuracy for the CNN-based model was found to be 31.80% and for the Nearest Centroid Hierarchical Hard framework, 30.84%. Similar

trends can be observed for precision, recall, and F1-score, where the proposed method consistently having comparable performance to the deep learning baseline.

As previously observed with 4-season classification, it is shown that the quality of engineered features and problem decomposition is sufficiently able to offset any significant disadvantage of a lower model complexity.

7.6.2 Interpretability

The interpretability of Soft Hierarchical Nearest Centroid is illustrated in Figures 16-18. They present an example output generated for one specific prediction. As shown in Figure 16, the model explains every step of the decision-making process instead of only reporting the predicted season. In the first step, it decides whether the sample is warm-toned or cool-toned by showing which individual feature contributes most to the overall decision, as well as its relative importance. The process is repeated to identify the contrast of that sample. Finally, the overall prediction of the sample is achieved by multiplying the two intermediate predictions to produce a final seasonal prediction. The model also provides the user with the probabilities assigned to each of the four seasons, allowing the user to see which season is the runner-up prediction, as well as to determine how strongly the predicted class is favored. If the two most probable classes are close together, the user can use this information together with the explanations to make an informed decision on what to wear rather than relying solely on a single class label.

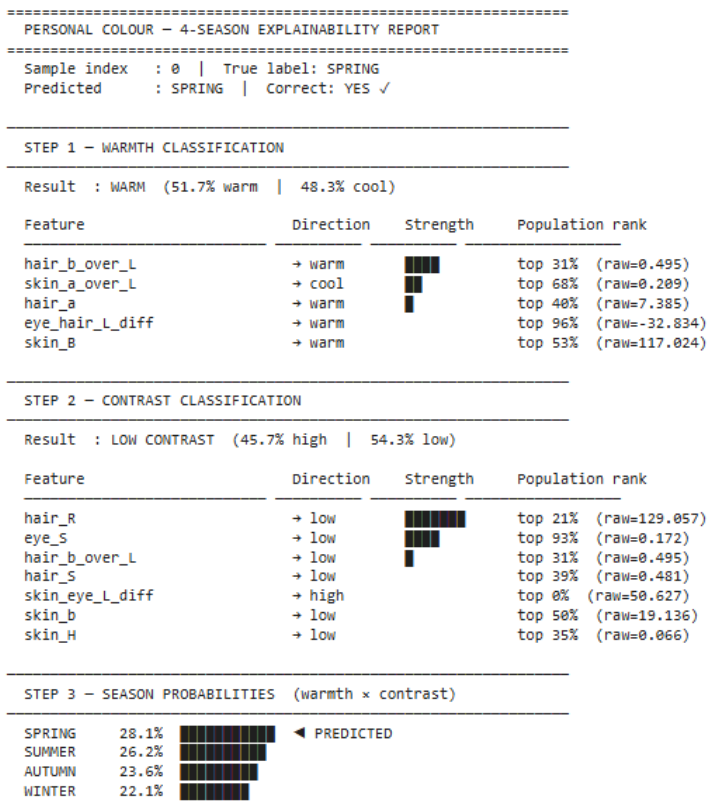


Figure 16: Example output of the proposed explainability framework, showing the step-by-step reasoning behind the seasonal prediction.

```
──────────────────────────────────────────────────────────
PLAIN-LANGUAGE SUMMARY
──────────────────────────────────────────────────────────

You are a SPRING.

Your colouring is 52% warm-toned. The key signal was
'hair_b_over_L', which pulled strongly toward the warm centroid.

Your lightness contrast is 54% low. The key signal was
'hair_R', which defined the contrast boundary.

Combined: warm tone × low contrast = SPRING
with 28.1% confidence.

969 out of 3943 people in the training data share
this season (24.6% of the population).
```

Figure 17: Plain-language summary explaining the model's prediction in human-readable terms.

Figure 17 presents a Plain-Language Summary of how the decision is made, which translates the model's reasoning into an easy-to-understand explanation. It enables users to interpret and evaluate the result with confidence, highlighting the explanatory capabilities of the model.

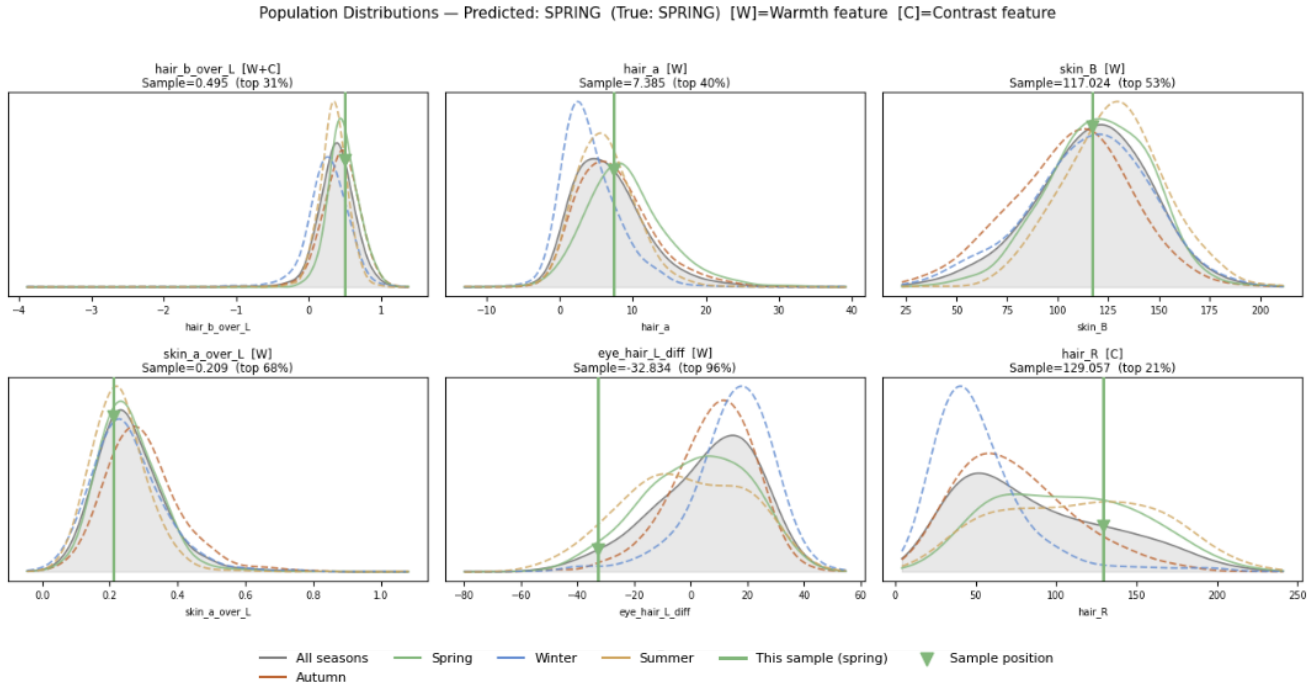


Figure 18: Population distributions of the most influential features, showing the position of the sample relative to each seasonal class.

The interpretability of the proposed framework is further enhanced in Figure 18, where the distribution of each feature across four seasons is illustrated. The vertical green line indicates the value of the sample being evaluated, allowing the users to immediately see where they fall relative to the learned distributions. These plots provide insight into why certain features support or contradict a particular season. For instance, the value of normalized skin redness (skin_a_over_L) falls on the peak of the Summer distribution, indicating a cool undertone. Although this particular feature does not support the true Spring label, the visualization clearly shows how the model sees

this but still chooses Spring as the final prediction because other features align more closely with it. This information helps users to understand the model's reasoning better.

In conclusion, Nearest Centroid, as well as other distance-based approaches, offer greater interpretability since their predictions are based on explicitly engineered features that correspond to established concepts in personal color analysis, such as warmth, contrast, and vividness. This allows model decisions to be traced directly to measurable and human-interpretable perceptual attributes, making the reasoning process more transparent. In contrast, CNN models learn high-dimensional latent representations that are not directly associated with human-interpretable concepts. Therefore, although they may achieve strong performance, it is difficult to determine which specific perceptual attributes influenced a certain decision and to verify if they do indeed rely on personal color analysis theory. Even if the same feature engineering was applied to a CNN — and technically this is possible — the provided explanations would still remain post-hoc approximations of how the model operated. An explicit representation of the decision-making process that actually corresponds to the model´s internal mechanism is still lacking. In other words, although we could adjust the CNN models so that they produce explanations that are interpretable for end users, the model itself remains uninterpretable, as it operates on numerous nonlinear transformations. In contrast, distance-based models are intrinsically interpretable because they rely on explicit distance comparisons to class centroids, which is transparent. As a result, every stage of the decision process is traceable and reproducible. In brief, there are two sources of interpretability, which are interpretable features and interpretable model, and it is the second element that CNN models lack.

7.6.3 Data efficiency

In terms of data efficiency, the CNN-based approach required approximately 4,000 labeled images to achieve its accuracy of 55.4% for 4-season classification and 31.8% for 12-subtype classification. Nearest Centroid requires a substantially smaller subset, specifically in the range of 100-1,000 labeled images to achieve an accuracy of 51%-53.97% for 4-season classification and 500-1500 labeled images for 28%-30.84% accuracy for 12-subtype classification. This demonstrates that Nearest Centroid has significantly higher data efficiency than CNN models. Furthermore, as Nearest Centroid mostly relies on the computation of distance to precomputed class centroids, its computational requirements are significantly lower than CNN models, which are typically resource-intensive.

7.6.4 Feature Space Analysis

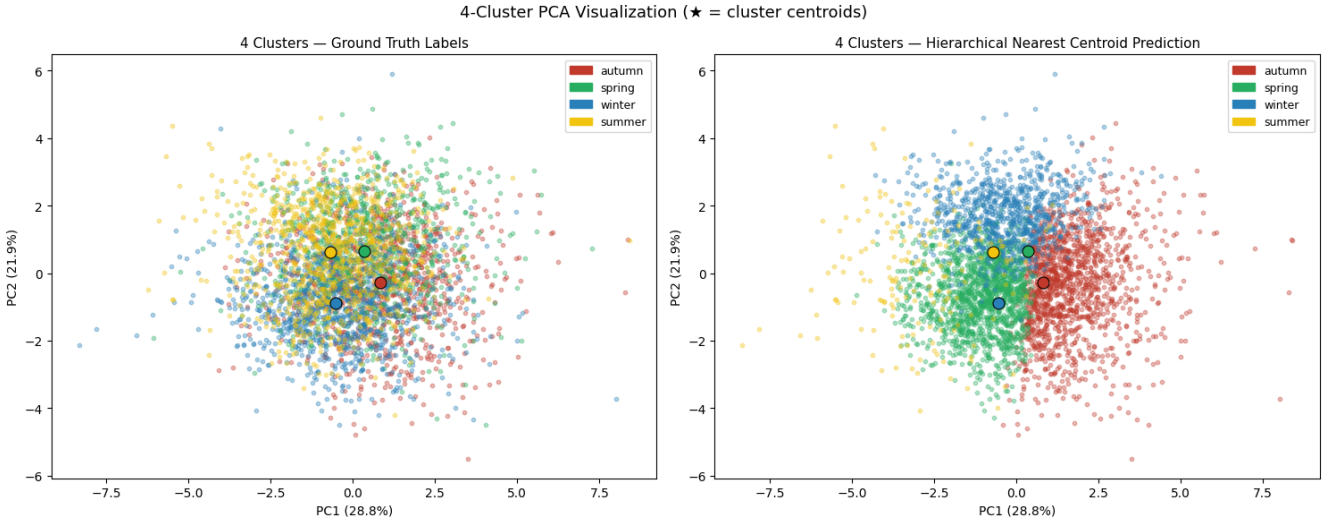


Figure 19: PCA Visualization of the 4-Season Classification Framework: Ground Truth Labels and Hierarchical Nearest Centroid Predictions

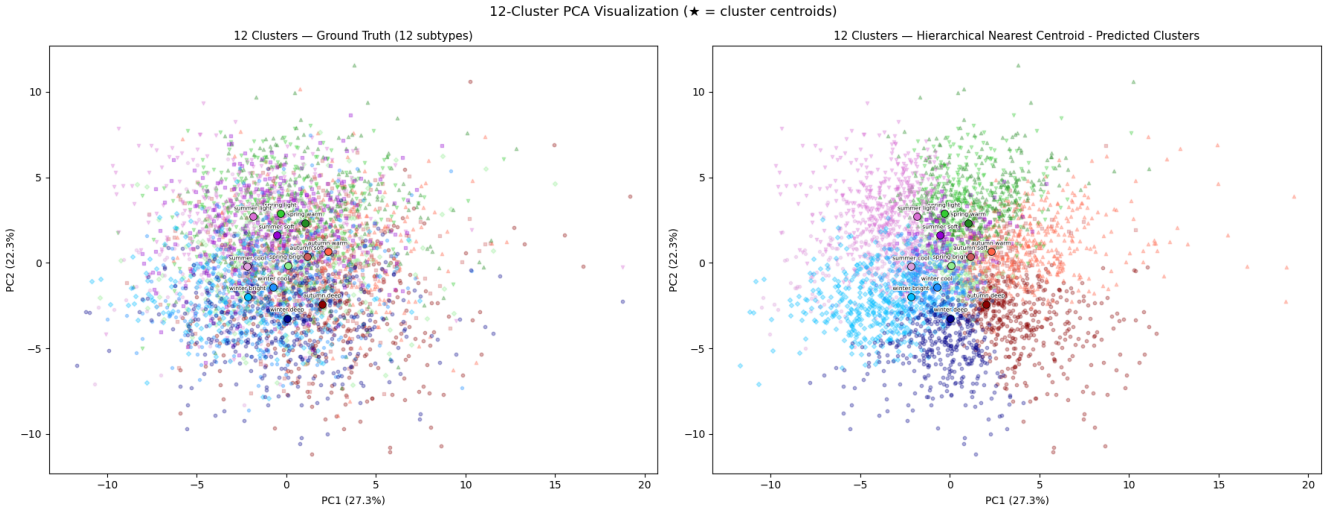


Figure 20: PCA Visualization of the 12-Subtype Classification Framework: Ground Truth Labels and Hierarchical Nearest Centroid Predictions

The left PCA plots for the ground truth in Figure 19 and Figure 20 show a substantial overlap between all seasons and subtypes. This indicates that there is, indeed, not much intrinsic separation between seasonal classes in the feature space. This conclusion is supported by the relatively low (55% and 31%) performance of the CNN-based models, indicating that the limitation in performance is primarily due to the data representation itself, not just the classifier capacity.

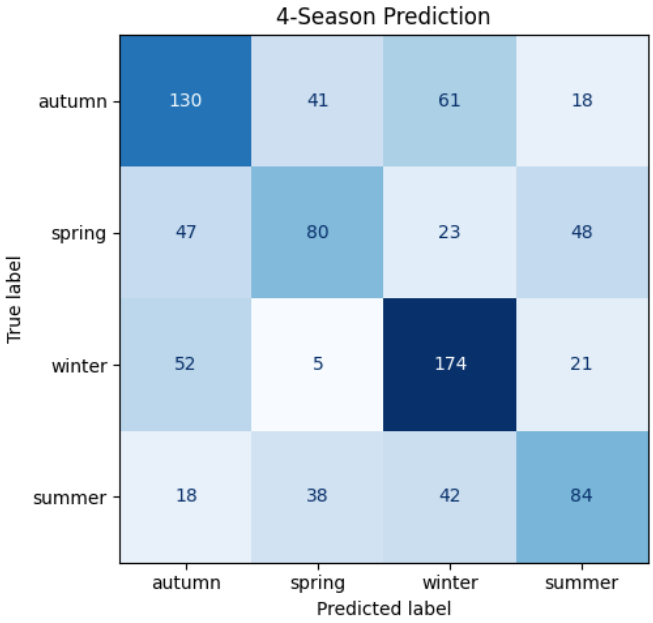


Figure 21: Confusion matrix for the 4-season classification task with Hierarchical Soft Nearest Centroid

This result is expected given the fact that personal color analysis is fundamentally an ambiguous task. As shown in the confusion matrix for 4-season classification in Figure 21, the strongest confusion occurs between Autumn and Winter, since these two seasonal palettes have display strong visual similarity, they both share deeper tones, lower brightness and strong contrast. From a color-analysis perspective, a deep autumn palette can resemble a deep winter palette, which caused the confusion. Similarly, Spring and Summer also cause considerable confusion, likely due to their shared high-brightness characteristics and relatively subtle differences in warmth and saturation. Winter has the highest correct predictions (174), indicating that it is the most visually distinctive season. On the other hand, Spring has the lowest correct predictions. This suggests that Spring lacks strong discriminative boundaries.

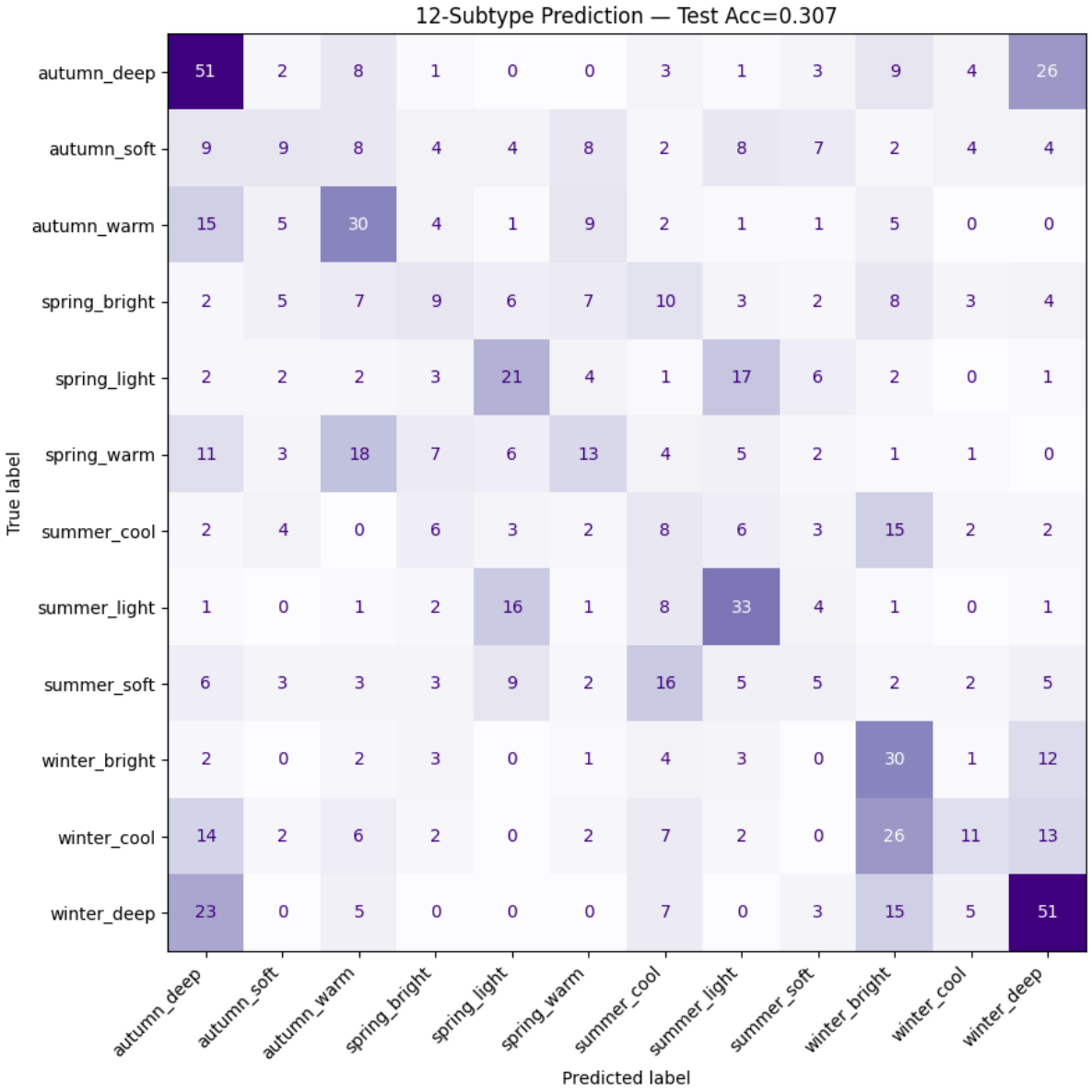


Figure 22: Confusion matrix for the 12-subtype classification task with Hierarchical Hard Nearest Centroid

The 12 subtype confusion matrix in Figure 22 also shows that there was the most confusion between adjacent visual subtypes (e.g., Autumn Deep vs. Winter Deep and Spring Light vs. Summer Light) indicating that many subtype distinctions rely on gradual variations in undertone, contrast and saturation rather than sharply separable visual boundaries. The model consistently performed better on subtypes with distinctive luminance or contrast features than on subtypes with muted properties.

8 Discussion

This thesis is not only concerned with identifying the most suitable approach for personal color analysis, but also provides insights into the structure of the task itself. Personal color analysis relies on subjective judgments by its nature, therefore, even human experts cannot achieve perfect consistency with their predictions. Due to this absence of a fully objective ground truth, label consistency is not guaranteed. As a result, it cannot be expected that a computational model reaches perfect performance, as its performance is ultimately evaluated against a reference standard that is subjective itself. Consequently, reported accuracy should be interpreted as agreement with a specific annotation protocol rather than an absolute measure of correctness.

Although personal color analysis is straightforward theoretically, in practice, it is quite challenging due to the subtle distinctions between classes and the heavy reliance on human perception. When the theory is implemented through the three-dimensional framework, a mismatch between textbook definitions and data-driven feature importance can be observed. For example, while undertone is theoretically defined by the coloration of the skin alone, empirical results show that factors such as lighting and inter-regional contrast also contribute to model performance. This mismatch does not necessarily indicate that the theoretical framework is flawed, it rather demonstrates how holistic and context-dependent human perception is. Therefore, the task is better seen as an approximation of human perceptual judgment, instead of a well-defined classification problem. As a result, although it can justify its predictions in terms of human-interpretable features, when these features exhibit counter-intuitive behavior—such as when seemingly irrelevant factors contribute to improved performance—the model cannot provide a deeper causal explanation beyond the observed statistical associations.

The inherent ambiguity is demonstrated in Section 7.6.4, where perceptual boundaries are not strictly separable. One possible explanation for the observed class overlap is the inherent ambiguity that exists in personal color analysis, where the perceptual boundaries between colors are not strictly defined, even for human annotators. However, an alternative explanation could be that the current feature representations may not accurately depict the best color space for modeling this task. Although color spaces such as HSV, LAB and engineered features all provide meaningful approximations, they may still be insufficient to provide a complete representation of human perception. This presents the possibility that the observed class overlap may not only be caused by the nature of the problem, but also by the choice of feature space. If a more expressive and perceptually aligned feature space was developed, it could potentially increase the separability of the classes. Therefore, what remains unexplained is the extent to which the color spaces correspond to true perceptual mechanisms.

8.1 Practical Implications

The proposed distance-based framework has multiple practical implications for automated personal color analysis.

Firstly, the proposed distance-based framework provides a much higher level of interpretability than the CNN-based framework. As seen in the Section 7.6.2, this method offers an opportunity to inspect and evaluate the feature selection procedure at each layer, allowing users and developers to examine the contributions of various facial regions or color properties explicitly. Furthermore, by decomposing the decision process into sequential stages, the method is able to provide intermediate

reasoning steps that reflect how predictions are formed. This transparency gives the users more confidence in the results and the system, which is difficult to achieve in conventional CNN-based models.

Secondly, this framework reduces the need for costly large-scale expert annotations, which is required by supervised deep learning methods. It was demonstrated that Nearest Centroid requires a relatively small number of labeled samples to obtain reliable centroids and meaningful class representations. Unlike CNN-based methods where a massive amount of labeled training data is necessary to achieve performance improvements, the distance-based methods can already establish the core class structure with limited labeled information. Additional unlabeled data can still be used to continue refining the feature space. Therefore, it is a much more viable approach for situations where resources to perform annotations are limited.

Finally, the distance-based models are significantly more lightweight compared to deep neural networks. This enables real-time inference without the need of any specialized hardware. It results in easy deployment, lower energy consumption and rapid experimentation with different model configurations.

These benefits are obtained with only a minor trade-off in predictive performance, specifically a few percentage points in accuracy.

8.2 Limitations

Despite the promising results of the proposed approach, there are some limitations that must be recognized.

Firstly, personal color analysis is an inherently subjective task. The ground truth labels used in this study are provided by human experts, however, there is no definitive, absolute objective standard for seasonal colors. As a result, some degree of inconsistency or ambiguity in the annotations may influence the performance of the model and the evaluation results of these models.

Secondly, variations in image brightness and quality can still introduce noise to the extracted features even though color spaces such as LAB and HSV are designed to reduce lighting sensitivity. Furthermore, natural color and brightness properties of one's face can also be modified or obscured due to the use of makeup. For instance, foundation can alter perceived skin undertones, eyeliners or eyeshadows can alter perceived contrasts. Human experts are often trained to infer underlying seasonal color through the layers of makeup to a certain extent, however, the proposed models fully rely on pixel-level features and are therefore unable to detect artificially modified color information. This fundamental mismatch between human expert labeling and computational feature extraction can introduce noise into the learning process and potentially limit model performance.

Thirdly, the proposed pipeline relies on accurate segmentation of the face to extract representative skin, hair and eye color values. If the segmentation system fails or produces poor masks, the extracted color values will not match the intended face regions, resulting in unreliable classifications. There was no observed segmentation failure during the experiments, likely due to the high quality and uniformity of the data set, however, this dependency is still a limitation of the proposed pipeline. In real application, the likelihood of segmentation failure could increase because of variations such as occlusions, extreme head positions, poor lighting, accessories, or odd hairstyles. Future work should explore the robustness of the pipeline under these conditions, and possibly incorporate quality checks to identify and deal with segmentation failure, prior to feature extraction.

Fourthly, the proposed framework assumes that personal color analysis can be split into 3 smaller

tasks: detecting warmth, detecting contrast, and detecting vividness. Although this decomposition aligns with human expert reasoning, there may be many more perceptual factors that impact personal color. This is reflected in previously observed results, where the models occasionally select features that appear counterintuitive from a human perspective to assign a certain attribute to a person. It suggests that these 3 layers alone may not fully capture the complexity of real-world personal color analysis.

Finally, the dataset (CelebA-Mask-HQ) used in this study has the potential to introduce bias, since it consists of celebrity images from Western media sources, mostly with lighter skin tones. This causes a demographic imbalance in the data distribution. Furthermore, there are fundamental differences between an average celebrity and an average person regarding facial characteristics and how the images are taken. Consequently, the generalizability of the model to broader real-world populations may be limited.

8.3 Future Work

An extension of this work should focus on finetuning the proposed framework to improve its performance and robustness. Future work may utilize the findings of this study to improve CNN-based approaches.

The first suggestion for future work is to use color-space-specific distance metrics. Although Euclidean distance provides a simple way to measure similarity, it may not accurately align with human perception of colors. Euclidean distance assumes that 1 unit change in the color space equals the same perceptual change for human vision. However, in reality, humans are more sensitive to some colors than others, and their perception changes depending on various factors such as hue and saturation. We therefore need distance metrics that measure how different two colors appear to humans, not just how far apart they are numerically. For LAB specifically, CIEDE2000 is the most advanced and perceptually accurate color difference formula. For HSV, a suggestion is to use circular distance for Hue since it is circular, then combine it with Saturation and Value Euclidean distance. A further step could be to implement different weights to the different color channels. For RGB, Euclidean is a standard choice, however, it can also benefit from a weighting scheme. Since Green contributes more to perception than the other two colors [3], increasing its weight can improve realism. Choosing the suitable distance metric for each color space can enhance the feature representation and accordingly improve model performance.

The second suggestion is to explore other more advanced techniques for soft prediction. The current soft prediction relies on inverse-distance normalization, which provides some uncertainty assessment in borderline cases. However, in some settings, it performs slightly worse and hard prediction. Since soft prediction has a stronger potential from an interpretability standpoint due to its ability to reflect continuous transitions between classes, it is worth being further investigated and improved, possibly with other probabilistic methods such as Gaussian kernel-based assignments or temperature-scaled softmax functions.

The third suggestion is to expand the current feature set. Previous experiments have shown that model performance is highly dependent on feature engineering and the quality of the selected features. Currently, the engineered feature set consists of 50 features. Given that greedy search can scale to a larger number of features, developing a more extensive and domain-specific set of features may lead to significant gains in accuracy.

The fourth suggestion is to make use of a more diverse and realistic dataset. The current dataset

consists of only celebrity images from Western media sources, causing a demographic imbalance in the data distribution. By using a more demographically balanced dataset, the generalizability of the model to broader real-world populations can be improved.

Lastly, the findings drawn from experimenting with interpretable models can be incorporated into the process of training deep learning models. Previous experiments have shown that LAB and HSV, as well as a hierarchical pipeline are more suitable for personal color analysis. These findings can be used to enhance deep learning models, which would allow potential performance gains.

References

[1] Rama Alyoubi, Taif Alharbi, Albatul Alghamdi, Yara Alshehri, and Elham Alghamdi. Colors matter: Ai-driven exploration of human feature colors. *arXiv preprint arXiv:2505.14931*, 2025.

[2] FacePerceiver. facer: Face analysis library. https://github.com/FacePerceiver/facer, 2023. Accessed: 2026-04-22.

[3] Joshua J Gooley, Shantha MW Rajaratnam, George C Brainard, Richard E Kronauer, Charles A Czeisler, and Steven W Lockley. Spectral responses of the human circadian system depend on the irradiance and duration of exposure to light. *Science translational medicine*, 2(31):31ra33–31ra33, 2010.

[4] Neil Hester and Eric Hehman. Dress is a fundamental component of person perception. *Personality and Social Psychology Review*, 27(4):414–433, 2023.

[5] Johannes Itten. *The Art of Color*. Van Nostrand Reinhold, New York, NY, 1973. Original work published 1961 as *Kunst der Farbe*.

[6] Carole Jackson. *Color Me Beautiful*. Random House Publishing Group, New York, NY, 1987.

[7] Renae Knapp and Dee Dorr. *Beyond the Color Explosion: The Color Key Program*. Rainy Day, Publisher location unknown, 1985. Based on Robert C. Dorr's Color Key System developed from 1928 onwards.

[8] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation, 2020.

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

[10] Lollipop77. Do you think a person can be more than one season? https://www.reddit.com/r/coloranalysis/comments/1fv87lx/do$_y ou_t hink_{ap} erson_c an_b e_m ore_t han_o ne_s eason/$ $2026 - 06 - 30$.

[11] Maklaan. Rgb color cube. https://commons.wikimedia.org/wiki/File:RGB$_c olor_c ube.svg, 2015. Wikimedia C$ $SA3.0. Accessed : 2026 - 06 - 30$.

[12] Jisoo Park, Hyungjoon Kim, Seonmi Ji, and Eenjun Hwang. An automatic virtual makeup scheme based on personal color analysis. In *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication*, pages 1–7, 2018.

[13] Pinterest. Pinterest pin 492862752993410675. https://www.pinterest.com/pin/492862752993410675/, n.d. Accessed: 2026-06-30.

[14] PSY222. Colorinsight: Personal color detection model. https://github.com/PSY222/Colorinsight, n.d. GitHub repository, accessed: 2026-04-17.

[15] Paolo Russo and Fabiana Di Ciaccio. Deep classification of microplastics through image fusion techniques. *IEEE Access*, PP:1–1, 01 2024.

[16] Lindsey Siegfried, Sophie M Bilik, Jamie Burgess, Paola Catanuto, Ivan Jozic, Irena Pastar, Rivka Stone, and Marjana Tomic-Canic. An optimized and advanced algorithm for the quantification of immunohistochemical biomarkers in keratinocytes. *JID Innovations*, 4:100270, 02 2024.

[17] Patricia Spires. Beyond undertone: Know your depth, saturation, & contrast. https://www.pinterest.com/pin/437693657549269250/, n.d. Pinterest. Accessed: 2026-06-30.

[18] Lorenzo Stacchio. Deep-armocromia: Seasonal color analysis for face images. https://github.com/lorenzo-stacchio/Deep-Armocromia, 2024. GitHub repository.

[19] Lorenzo Stacchio, Marina Paolanti, Francesca Spigarelli, and Emanuele Frontoni. Deep armocromia: A novel dataset for face seasonal color analysis and classification. In Alessio Del Bue, Cristian Canton, Jordi Pont-Tuset, and Tatiana Tommasi, editors, *Computer Vision – ECCV 2024 Workshops*, pages 352–367, Cham, 2025. Springer Nature Switzerland.

[20] Style with DC. Everything you need to know about color dimensions & color analysis. https://stylewithdc.com/blogs/blog/everything-you-need-to-know-about-color-dimensions-color-analysis, n.d. Accessed: 2026-06-30.