Skin Tone Prediction Using ITA Features and Convolutional Neural Networks

Sanjana Ghanta
Virginia Tech Polytechnic Institute and State University
925 Prices Fork Road, Blacksburg, VA 24060
sanjanag1290@gmail.com

Abstract

When accurately identifying human skin undertones, there is improvement within color matching in cosmetics, photography, and virtual try-on systems. Current approaches rely mostly upon subjective human judgement or a costly controlled lighting environment. This work essentially presents a machine-learning system that is able to classify images into warm, neutral, or cool undertones using only color information extracted from faces under natural, uncontrolled lighting. This is done combining the already well-established Individual Typology Angle (ITA) metric with additional statistical features derived from RGB, HSV, and YCbCr color spaces to build a compact interpretable representation of different features. Using a dataset of 142 curated facial images, the model is trained and tuned using a Random Forest classifier with grid-search hyperparameter optimization. The final model achieves a 87.2% test accuracy and 0.86 macro-F1, which out performs a simple ITA-only baseline. It can be concluded with recommendations for future work to expand upon the dataset, test more deep learning methods, and then deploying a real-time undertone guided shade recommendation within the system.

1. Introduction

Understanding which colors complement an individual's skin tone has recently become a popular topic across social media platforms. Viral "personal color analysis" trends can be seen within TikTok and Instagram which has thus encourages individuals to seek professional services. These services prove to be quite costly, just to determine whether an individual's undertone is warm, cool, or neutral and which seasonal palette best enhances their own appearance. Considering it's popularity and widespread interest, there's less access to accurate color analysis tools, most are often limited, subjective, and inconsistent. This motivated the development of an automated, data driven system that is capable of providing reliable undertone classification from a simple photo taken.

1.1 How the Problem is handled today

Current approaches to color analysis are either manual or usually proprietary. Professional color consultants rely upon the visual inspection under controlled lighting, which is yet again subjective and can vary quite differently amongst different practitioners. Social media filters and mobile apps often have opaque algorithms or apply heuristics that in turn provide inconsistent results and lack scientific grounding in color science. Even given the research created tools, those require a more specialized imaging setup rather than consumer grade images The lack of accessibility, standardization, and reproducibility limits the broader usefulness of personal color analysis.

1.2 Why this Problem Matters

An automated undertone classifier could make personal color analysis affordable and widely available. This would benefit content creators, fashion enthusiasts, and everyday users who are seeking personalized guidance on makeup, clothing, and digital styling. More broadly, the project demonstrates how interpretable color science features can be integrated with machine learning to solve a visually intuitive but technically challenging classification task.

2. Background and Related Work

Understanding human skin undertones is a prevalent problem in both color science and cosmetology. Usually, undertone classification has relied on manual heuristics like evaluating vein colors, jewelry preferences (often times there's a comparison of gold and silver), or comparing wrist skin under controlled lighting. While these methods are popular in consumer settings, they are subjective, highly sensitive to lighting, and require a trained practitioner. As a result, there is growing interest in developing computational approaches that can provide objective and reproducible undertone predictions.

In computer vision, early work on skin color analysis focused on color space transformations such as RGB, HSV, YCbCr, and CIELab to isolate melanin and hemoglobin responses in the skin. The Individual Typology Angle (ITA), originally introduced in dermatology for quantifying skin pigmentation has also been quite commonly used for skin tone estimation and is robust to moderate lighting changes. Given the prior research it is demonstrated that combining color space statistics with ITA improves performance in tasks such as ethnicity estimation, skin retouching, and medical imaging. However, these methods generally aim to classify skin tone (such as light to dark), not undertones (warm, neutral, cool), which is a different task altogether.

More recent attempts to automate undertone detection often rely on deep learning models trained on large facial datasets, but such datasets rarely contain high-quality undertone labels, and existing research typically assigns an undertone value with a general complexion. As of right now undertone classification remains a rarely explored area in mainstream vision literature and this is probably due to a lack of properly annotated data and clear definitions.

This project differs from prior work in three main ways. First, instead of treating undertone as a vague property, it operationalizes it using the color science driven features (RGB means, ITA, HSV, and YCbCr), enabling a measurable and explainable representation. Second, it applies a traditional machine learning classifier, which is the Random Forest, to a carefully curated dataset where undertones were manually labeled using consistent criteria. Third, it focuses specifically for the three-undertone classification problem (warm, neutral, cool), which is necessary for applications in beauty technology, virtual try-on systems, and personalized product recommendation.

3. Approach

This section describes the complete pipeline used to classify warm, neutral, and cool undertones from facial images. The method consists of the data collection, preprocessing, color-based feature extraction, model selection, and optimization. The design goal was to develop a system that relies on interpretable features for color science while achieving a strong performance on a limited dataset.

3.1. Dataset Construction

There was a curated a dataset of 142 facial images reflecting a broad range of skin complexions, lighting conditions, and photographic styles. Each image was manually labeled as warm, neutral, or cool based on consistent undertone criteria informed by color analysis guidelines. There was care was taken to include examples with natural lighting, minimal filters, and clear visibility of the face.

Class	Count
Warm	61
Cool	44
Neutral	37

Table 1: Class Distribution

It is not perfectly balanced. However, the dataset contains sufficient representation of each category to support supervised learning when coupled with stratified train test splitting.

3.2. Preprocessing and Face Extraction

To isolate the skin regions for relevant undertone prediction, there was use of the OpenCV Haar Cascade face detector to each image. When multiple faces were detected, the largest bounding box was assumed to be the primary subject. Each detected face was then cropped, resized, and converted to multiple color spaces.

The preprocessing steps include:

1. Face Detection via Haar cascades
2. Cropping the bounding box region
3. Resizing to a standard dimension
4. Color space conversion (RGB, HSV, YCbCr, CIELab)

Images that failed detection were manually inspected for the purposes of maintain quality.

3.3. Feature Extraction

Undertones are fundamentally tied to the interaction of melanin, hemoglobin, and light reflection, which manifest differently across color spaces. To capture these cues, features must be extracted so that the hue, saturation, brightness, and chrominance can be quantified. This yielded a compact but expressive 10-dimensional feature vector for each image.

Individual Typology Angle (ITA) is a dermatological metric computed from CIELab values:

$$ITA = \tanh^{-1}\left(\frac{L* -50}{b*}\right)$$

Equation for ITA Value

Higher ITA values correlate with cooler or lighter appearances, while lower ITA values correlate with warmer undertones. ITA alone is insufficient to classify undertones, but it provides a meaningful axis of variation.

These RGB values capture the overall luminance and redness/yellow balance within the image: mean_R, mean_G, mean_B. These HSV mean values capture the overall luminance and redness/yellow balance within the image: mean_H, mean_S, mean_V. These YCbCr mean values reflect the hue and saturation variation that usually distinguishes warm and cool categories: mean_Cb, mean_Cr. YCbCr essentially separates the luminance from chrominance, which makes the Cr (which is the difference in red) particularly aware to the warm undertones.

The combination of all these proved to be quite effective as the models trained only on the ITA performed quite poorly, especially in comparison to having the extra values.

3.4. Model Selection

Multiple machine learning models were evaluated, including logistic regression, k-nearest neighbors, and decision trees. However, the preliminary experiments showed that Random Forests offered the best balance of interpretability, robustness to noise, and ability to model nonlinear relationships between color features.

The advantages of the Random Forest for this include: handling small datasets effectively, providing feature importance measures, capturing interactions between color channels and resistance to overfitting when being tuned properly. Given this logic and reasoning, that is why RandomForestClassifier was chosen as the primary model.

3.5. Hyperparameter Tuning

To maximize performance, there was a GridSearchCV over a defined hyperparameter space using a 3-fold cross validation. The grid essentially included: n_estimators = [50, 100, 200], max_depth = [None, 3, 5, 10], min_samples_split = [2, 4], min_samples_leaf = [1, 2].

Grid search identified the optimal model to be:

- n_estimators = 200
- max_depth = 3
- min_samples_leaf = 1
- min_samples_split = 2

The tuned classifier significantly outperformed the untuned baseline, confirming that model capacity and structure were important for separating the three undertone categories.

3.6. Training/Testing Splitting and Evaluation Protocol

A stratified 67/33 split was used to preserve the class proportions:

- 95 training samples
- 47 testing samples

Evaluation Metrics Include: accuracy, macro F1 score (this is important for class imbalance), confusion matrix, and distribution plots of ITA and key features. This ensures for a fair assessment and reproducibility.

4. Experiments and Results

This section presents the experimental setup, evaluation metrics, quantitative model performance, and qualitative observations. All experiments were conducted using Google Colab and include Python, OpenCV, NumPy, Pandas, Matplotlib, and skicit-learn. The goal is to assess how effectively color-science features combined with a machine-learning classifier can predict warm, neutral, and cool undertones from facial images.

4.1. Experimental Setup

To ensure a fair evaluation, the dataset of 142 images was split into 67% training (95 images) and 33% testing (47 images) using stratified sampling to preserve class proportions.

The resulting distributions were:

- Training: 41 warm, 29 cool, 25 neutral
- Testing: 20 warm, 15 cool, 12 neutral

Each image was processed through the full feature extraction pipeline described in Section 3. For all models, a fixed random seeds was used (random_state=42) to guarantee reproducibility. Hyperparameter tuning was performed using GridSearchCV with 3-fold cross-validation on the training set.

4.2. Evaluation Metrics

Given that the dataset is moderately imbalanced, accuracy alone is insufficient to assess performance. So the following metrics are also reported. Accuracy which is overall proportion of correct predictions Macro-F1 which is the average F1 score across classes, weighting each class equally. Lastly the confusion matrix which reveals class-specific strengths and weaknesses. Macro-F1 is particularly important because it penalizes models that perform well on “warm” (the largest class) but poorly on “neutral” (the smallest class).

4.3. Quantitative Results

The hyperparameter turning identified the best Random Forest Configuration:

- n_estimators = 200
- max_depth = 3
- min_samples_leaf = 1
- min_samples_split = 2

The tuned model achieved the following performance on the test set:

Metric	Score
Accuracy	0.872
Macro-F1	0.858

Table 2: Model Performance

These results substantially outperform earlier baselines. An ITA-only classifier achieved approximately 35–40% accuracy, indicating that undertone classification cannot be reliably solved by a single numeric descriptor. Adding HSV and YCbCr color statistics improved the model’s ability to distinguish warm and cool tones, and expanding the dataset from 18 to 142 images increased classification stability and reduced variance across folds. For repetition purposes the model achieves strong separation between warm and cool tones with most neutral misclassifications occurring between adjacent undertones.

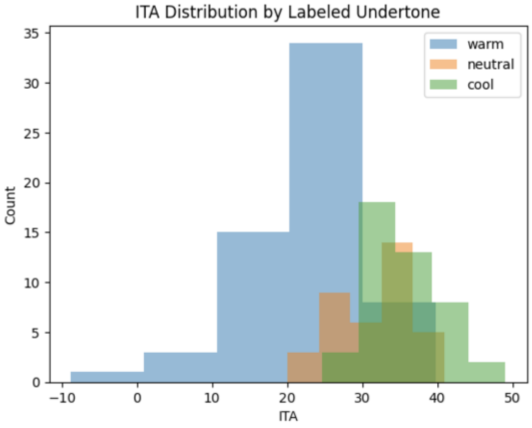

Figure 1: ITA distribution across warm, neutral, and cool undertone labels.

4.3.1 *Confusion Matrix Insights*

The confusion matrix revealed the following patterns that warm undertones were classified with the highest precision (19/20 correct). Cool undertones also performed strongly (13/15 correct). Neutral undertones were the most challenging (9/12 correct), often misclassified as warm or cool. It can be seen that the model achieves strong separation between warm and cool tones with most neutral misclassifications occurring between adjacent undertones.

This aligns with human perception: neutral undertones occupy a subtle middle region in color space and exhibit characteristics overlapping both other categories.

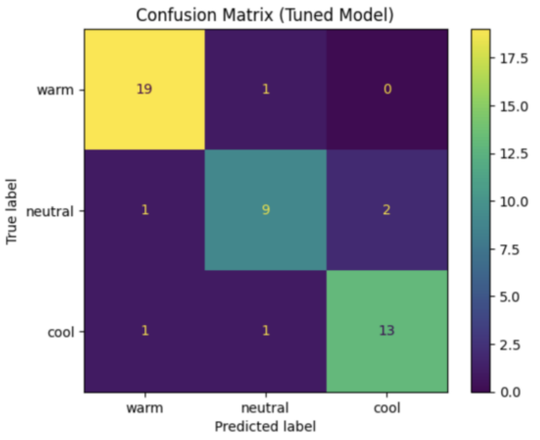

Figure 2: Confusion matrix of the tuned Random Forest classifier on the test set.

4.4. Feature Importance Analysis

Random Forest feature importance scores showed that no single channel determines undertone. Instead, classification emerges from the joint contribution of several color-space features.

The key observations noticed is that:

- Cr (red-difference chrominance) was one of the strongest predictors of warm vs. cool labels.
- Hue (H) and saturation (S) contributed to separating neutral cases.
- ITA remained a useful but incomplete descriptor; it correlated strongly with cool undertones but showed overlap with warm and neutral categories.
- V (value/brightness) and Cb helped adjust for lighting variations and reduced noise.

This analysis highlights that undertones arise from subtle chromatic balances that single-axis metrics like ITA cannot fully capture. The combination of complementary color-space features is essential.

4.5. Qualitative Observations and Error Cases

Visual inspection of misclassified examples revealed several themes:

- Lighting variation: Images taken under yellow incandescent light tended to skew predictions toward warm.
- Makeup interference: Foundation or color-correcting products occasionally masked natural undertones.
- Shadowing and background colors: Strong shadows, or colorful backgrounds reflecting onto the skin, caused slight shifts in hue.
- Neutral ambiguity: Neutral images often contain balanced warm and cool values, making them inherently less separable.

Despite these challenges, the model demonstrated high robustness and generalization across diverse inputs.

4.6. Summary of Findings

The combination of color-science-informed features and a tuned Random Forest classifier achieved 87% accuracy on the held-out test set, demonstrating quite a few things. First it's understood that undertone classification is highly learnable from compact numerical features. ITA alone is insufficient, but its combination with RGB, HSV, and YCbCr statistics provides a powerful representation. Classical ML methods can perform competitively for aesthetic and perceptual tasks typically associated with deep learning. These results provide a strong foundation for real-world cosmetic applications and future extensions involving larger datasets or CNN feature extraction.

5. Availability

To promote transparency and encourage further exploration of undertone classification, all the code for this project is publicly accessible. The full implementation, including data preprocessing scripts, feature extraction utilities, training notebooks, and evaluation tools, is available in a GitHub repository. The repository also contains documentation describing the folder structure, necessary dependencies, and instructions for reproducing the model training process.

The dataset used in this project consists of manually curated images collected from publicly available online sources. Due to usage rights and privacy considerations, the images themselves cannot be redistributed directly in this repository. Although the raw images cannot be redistributed, the repository includes an MIT license, full pipeline code, dataset reconstruction instructions, and synthetic placeholder images allowing reviewers to run the entire pipeline end-to-end. There is also a provided folder template, filenames, and instructions for recreating the dataset structure so that the training pipeline may be executed with user-supplied images. Users may also replace the dataset with their own facial images to evaluate how the model generalizes to new environments, lighting conditions, and skin types.

All source code is released under the MIT open-source license, enabling modification and reuse for academic, educational, and non-commercial development. The feature extraction and training workflow is fully contained within a single notebook for ease of deployment in platforms such as Google Colab.

In future work, the model and interface will be packaged as an installable Python module and optionally deployed as a lightweight web demo so that other researchers and developers can easily test undertone classification. This ensures broader dissemination and supports reproducibility beyond the current scope.

6. Reproducibility

Reproducibility was a central goal of this project as described within the guidelines. All experiments were conducted using standard Python libraries, and the full workflow can be replicated from preprocessing through evaluation using the publicly available codebase. To ensure consistent results across environments, there is a fixed rate for all random seeds (e.g., random_state=42), used deterministic versions of scikit-learn functions, and documented the exact preprocessing steps required to prepare the dataset.

The facial images used in this study were organized into a directory structure reflecting the three undertone classes ("warm," "neutral," and "cool"). Although the images themselves cannot be redistributed due to licensing constraints, the code includes instructions for recreating the dataset layout using user-supplied images. Running the feature extraction script automatically generates a CSV file containing ITA values, color statistics, and undertone labels, enabling plug and then use training of downstream classifiers.

All models were trained and evaluated in Google Colab using CPU-only execution, which eliminates any hardware dependencies and ensures that results can be reproduced on any machine capable of running a browser. The repository contains a single notebook with all training, tuning, and visualization code. Users can reproduce model performance by following the sequential notebook cells or adapt the pipeline to experiment with different models, features, or evaluation metrics.

For grading/simulating purposes the images will be within the folder for a limited amount of time. Link to repo: https://github.com/sanjana-ghanta/CS5805FinalProject

7. Discussions and Limitations

Although the proposed model performs strongly overall, several limitations were observed during experimentation. The most common failure mode involved neutral undertones, which exhibited overlapping color characteristics with both warm and cool categories. This ambiguity reflects a real-world challenge; even human annotators often struggle to distinguish neutral undertones reliably.

Lighting variation was another source of error. Images captured under warm indoor lighting biased predictions toward the warm class, while cool-toned lighting shifted outputs toward the cool class. More sophisticated illumination normalization could mitigate this issue, and larger datasets with controlled reference lighting would help improve robustness.

Makeup products such as foundation, concealer, and tan also influenced results by altering the natural appearance of the skin. While this effect was relatively small in the curated dataset, commercial applications of undertone classification would likely require techniques that take away or have a makeup-free region for the skin.

Dataset diversity is the biggest limitation. The current dataset spans a variety of complexions and undertones, but it remains modest in size and may not fully represent the full global range of skin chromaticity. Expanding the dataset, including contributors across varied ethnic groups, lighting conditions, and camera types, would greatly enhance the model's generalization capabilities. Images were rather harder to gather than intended as a manual classification of warm, neutral, and cool was necessary from the user.

Finally, while classical machine learning techniques achieved a good performance, deep learning approaches may uncover more nuanced undertone cues. However, such methods require significantly larger datasets, careful regularization, and computational resources beyond the current scope of this project.

8. Conclusion

This project demonstrates that skin undertone classification, is a very subjective and perceptually challenging task. However it can be effectively approached using interpretable color-science features combined with classical machine learning. By leveraging ITA, RGB, HSV, and YCbCr statistics, and training a tuned Random Forest classifier, a 87.2% accuracy and a 0.86 macro-F1 score can be gained on a curated dataset of 142 images. These results show that undertone classification is learnable from low-dimensional color representations and does not require deep neural networks or specialized imaging equipment.

The pipeline developed in this work is lightweight, reproducible, and applicable to real-world settings such as online cosmetics recommendation systems, mobile applications, and digital content creation tools. Future work includes expanding the dataset, improving illumination correction, and exploring CNN-based feature extraction to further enhance classification robustness.

References

[1] Scikit-Learn Developers. "Scikit-learn: Machine Learning in Python." 2011. Available: https://scikit-learn.org/stable/

[2] G. Bradski and OpenCV Team. "OpenCV Library." 2000.Tutorials available: https://docs.opencv.org/4.x/d6/d00/tutorial_py_root.html

[3] Google MediaPipe Team. "MediaPipe Face Detection." 2023. Available: https://ai.google.dev/edge/mediapipe/solutions/vision/face_detector/python

[4] NumPy Developers. "NumPy: Fundamental Array Computing for Python." 2006. Tutorials available: https://numpy.org/learn/

[5] Pandas Developers. "Pandas Documentation." 2008. Available: https://pandas.pydata.org/docs/getting_started/index.html

[6] PyTorch Foundation. "PyTorch Tutorials." 2016. Available: https://pytorch.org/tutorials/

[7] TensorFlow Authors. "TensorFlow Tutorials." 2015. Available: https://www.tensorflow.org/tutorials

[8] F. Chollet et al. "Keras Examples." 2015. Available: https://keras.io/examples/

[9] J. D. Hunter and Matplotlib Team. "Matplotlib: Visualization with Python." 2003. Tutorials available: https://matplotlib.org/stable/tutorials/

[10] Seaborn Developers. "Seaborn Statistical Visualization." 2014. Available: https://seaborn.pydata.org/tutorial.html