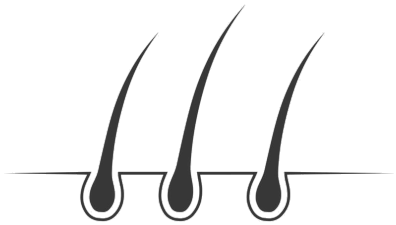


# Evaluation Report

## Students Names:


- Alam Bebar - 207415407
- Ryan Thawkho - 305070567


## Executive Summary

This report evaluates the results of our predictive modeling project for male pattern baldness classification by using both tabular and image-based datasets. The report reflects on the clarity and usefulness of our results, highlights key findings such as the improvement gained from segmentation and CLAHE filtering, and ranks models based on their real-world applicability. It also assesses whether the models meet the original goals set at the beginning of the project, from both a technical and practical perspective, identifies questions raised by our analysis, and suggests possible directions for future work. Furthermore, it includes a review of the models selected for final use. The first appendix provides supporting charts, classification reports, and visual summaries that reinforce the evaluation, including insights from merged survey data, top model comparisons, and per-class performance breakdowns, for both tabular and image based models. The report concludes with an additional appendix that outlines all the 5 datasets used in the project, including their sources and a brief description of each one.

# Table of Content

# 1. Evaluating the Results

After testing and comparing our models on both the tabular (Mendeley) and image-based (Roboflow) datasets, we assessed how well the results align with the original project goals, while improving the accuracy and practical usability of male pattern baldness stage prediction using machine learning and deep learning.

## Are your results stated clearly and in a form that can be easily presented?

The results are clearly presented in visual charts and confusion matrices in the Modeling Report. Key findings such as model accuracy, F1 scores, and class-specific insights were broken down for each algorithm. This structure makes it easy to communicate model performance both to technical and non-technical stakeholders.

- **Best tabular model**: XGBoost (accuracy = 0.8144)
- **Best image-based model**: ResNet18 (F1 = 0.7668, test accuracy = 0.8558)
- **Best preprocessing improvement**: Segmentation + CLAHE, which increased both accuracy and F1 score by ~15%

## Are there particularly novel or unique findings that should be highlighted?

- Applying segmentation to isolate the hair region, followed by CLAHE for image contrast, significantly improved classification quality, especially for stages with subtle scalp differences.
- Even lightweight models like EfficientNet-B0 and MobileNetV3-Large showed strong performance with proper preprocessing.
- Across nearly all CNN models, **Level 5** Norwood classification was a consistent challenge, which suggests a need to either merge this level with Level 4 or 6, or explore targeted oversampling or augmentation strategies.

## Model Ranking by Applicability

The models were ranked in the Modeling Report by a mix of accuracy and practical value. It is safe to say that these models can be recommended for further development or deployment depending on whether structured data or visual scalp data is available in a given clinical context.

**Tabular Models (Hair Fall Causes Dataset)**

| Model Name | Accuracy |
|---|---|
| XGBoost | 0.8144 |
| K-Nearest Neighbors (KNN) | 0.8041 |
| Random Forest (GridSearchCV) | 0.8041 |

**Image Models (Scalp Image Dataset)**

| Model Name | Accuracy | F1 Score |
|---|---|---|
| ResNet18 | 0.86 | 0.7668 |
| MobileNetV3-Large | 0.86 | 0.75 |
| EfficientNet-B0 | 0.85 | 0.75 |
| ConvNeXt-Tiny | 0.85 | 0.75 |

## How Well The Results Answer Business Goals

The core business goal was to build a working prototype that could predict hair loss stages and/or risk levels in men. Based on our results:

- **Tabular model (XGBoost)** is effective for quick assessments using survey or clinical data. Additionally, it offers the ability to customize and suggest personalized treatment plans or preventative measures for hair loss.
- **Image models** show clear potential for a diagnostic app, as they can classify scalp images with reasonably high accuracy.
- These findings collectively can facilitate early-stage screening and promote more tailored treatment plans or preventative approaches.

## Follow-up Questions Raised

- How can we improve predictions for Level 5 without increasing data volume? (e.g., smarter augmentation, merging stages)
- Is it possible to combine the predictions from tabular and image models to improve final decision-making?
- Would larger image datasets with more balanced stage representation lead to better classification of stages like Norwood 5-7?
- Can additional personal attributes (e.g., hormones, diet) improve tabular model performance (e.g. improve correlation between features)?

## Our Approved Models

Below is the list of approved models for the final report. These models align with both data science standards and business goals:

- **XGBoost**: Best tabular model in terms of accuracy and generalization. Recommended for numerical clinical intake forms.
- **ResNet18**: Strong overall image classifier with high F1 score and good handling of moderate and advanced stages.
- **EfficientNet-B0 (another option for an image deep learning model)**: Balanced accuracy with the benefit of being resource-efficient and fast.

# 2. Review Process

Looking back at our project, there were several things that worked well and a few areas we would approach differently in the future.

- One of the biggest successes was the decision to treat this as a multi-modal task. Using both tabular and image datasets gave us a broader perspective and let us explore different modeling techniques. This also made our final results stronger, since we could evaluate hair fall risk in two different ways: survey data and scalp images. Scalp images can be exclusively used for classifying male pattern baldness. Survey data can then be leveraged to recommend preventative measures, tailored to the predicted stage of baldness.
- Another key choice that paid off was applying **segmentation and CLAHE** preprocessing to the scalp images. At first, the models gave lower accuracy (around ~64%), but after improving the image quality and removing background noise, our top models reached ~86% test accuracy. That showed us the value of investing time in the preprocessing phase, especially for medical imaging.
- In terms of what could be improved, we spent a lot of time tuning deep learning models manually. Looking back, we could have automated the training pipeline a bit more to compare models faster. Also, **Level 5** Norwood was a consistent weak spot across models, and earlier exploration of class merging or smarter augmentation could have helped.
- Lastly, coordinating the merging of datasets and keeping labels aligned (especially between Norwood levels) took a while. For next time, we would document data transformations more thoroughly to save time.
- The experience helped us understand the strengths and limits of combining structured data with image classification. We gained hands-on practice with real-world data issues and are now more confident in applying machine learning to health-related use cases. Furthermore, earlier integration of business stakeholders might help align technical outcomes with practical expectations sooner.

## Level Grouping Rationale & Implementation (Levels 1-2 and 4-5)

### Why we merged

- **Levels 1-2:** Our initial dataset used **1-7** Norwood labels. As we expanded, we integrated a second dataset labeled **2-7** only. To harmonize labels across datasets and avoid discarding Level-1 samples, we map any model/output of **1** into the **"Levels 1-2"** bucket for presentation. This keeps cross-dataset training/evaluation consistent while preserving early-stage cases. **Levels 4-5:** Across model training and validation, the lowest per-class accuracy consistently appeared at **Levels 4 and 5** (Level-4 particularly weak at ~30-40%). These adjacent stages are visually similar and frequently confused. We therefore **merge Levels 4 and 5** into a single **"Levels 4-5"** bucket for user-facing results to reduce confusion and convey more reliable guidance.

## What's grouped vs. what stays the same

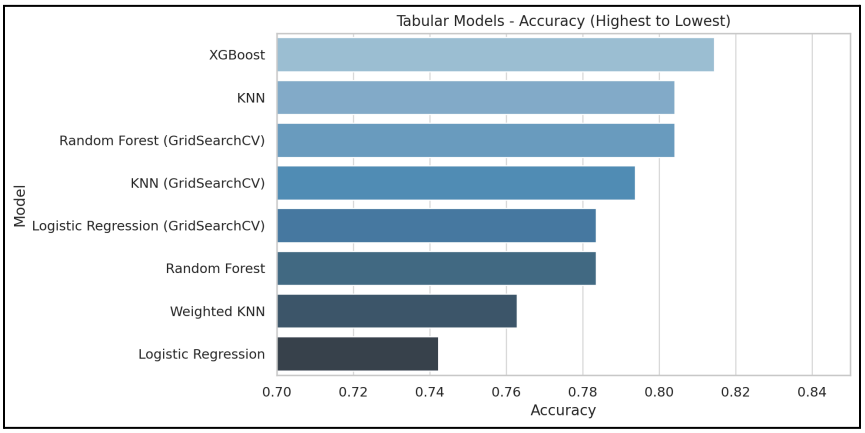
- **Model outputs (internal):** All five CNNs still predict **discrete levels (1-7)** and we still compute a standard **majority vote** on those raw levels.
- **User-facing display:** After majority vote, we **map** the final level into display buckets:
  - 1 or 2 : **"Levels 1-2"**
  - 3 : **"Level 3"**
  - 4 or 5 → **"Levels 4-5"**
  - 6 : **"Level 6"**
  - 7 : **"Level 7"**
- **Recommendations & estimates:** The **same combined recommendations, hair-fall guidance, and graft/price estimates** are shown for each merged bucket (1-2 share a set; 4-5 share a set). Level-specific content for 3, 6, 7 remains unchanged.
- Merging **1-2** resolves cross-dataset label alignment (1-7 vs. 2-7). Merging **4-5** addresses real-world model confusion and low class-specific accuracy by communicating a more dependable, clinically meaningful bucket, without changing how models are trained, voted, or tracked internally.

# Appendix: Key Graphs and Performance Tables

## Hair Fall Causes Dataset (Mendeley) + Merged Google Forms Responses

### Comparison Between Top 8 Tabular Models

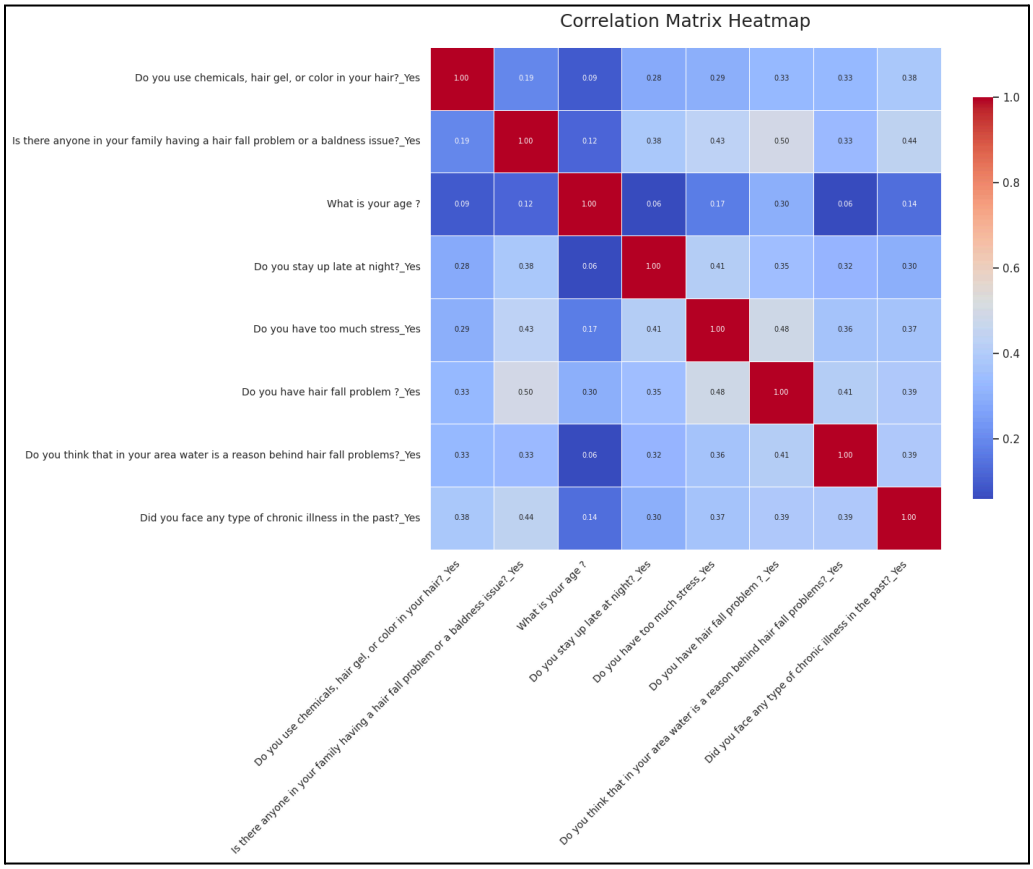
| Model | Accuracy | Notes |
|---|---|---|
| **XGBoost** | 0.8144 | Best performance overall |
| **KNN** | 0.8041 | Very good, minimal tuning |
| **Random Forest (GridSearchCV)** | 0.8041 | Good balance & interpretability |
| **KNN (GridSearchCV)** | 0.7938 | Slight improvement over KNN |
| **Logistic Regression (GridSearchCV)** | 0.7835 | High recall for 'Yes' |
| **Random Forest** | 0.7835 | Good but not optimized |
| **Weighted KNN** | 0.7629 | More balanced but lower |
| **Logistic Regression** | 0.7423 | Baseline model |




**Bar chart comparing the accuracy of the top 8 tabular models -** XGBoost achieved the highest accuracy (0.8144), followed closely by KNN and Random Forest with GridSearchCV. Logistic Regression (default) performed the weakest, serving as a baseline reference.

## Features Correlation Matrix (After Merging Google Form Responses)

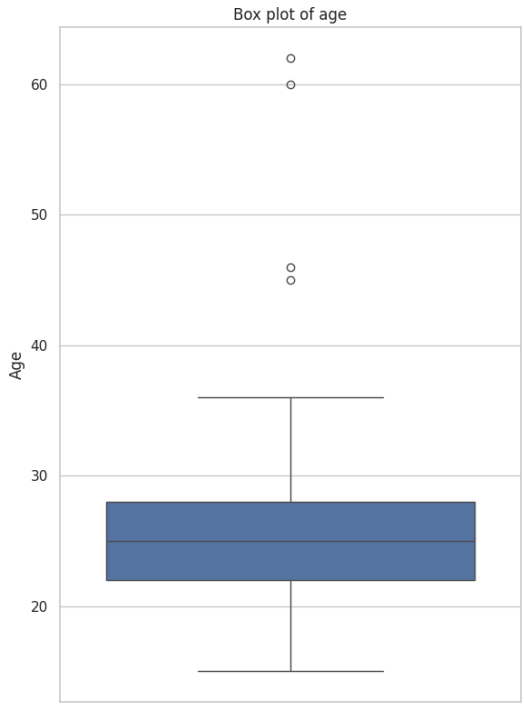
- The following charts and graphs present the values included in the final dataset after merging the records from the Google Forms responses. **Interestingly, after adding the new records, we observed a surprising decrease in the correlation values between the features.**

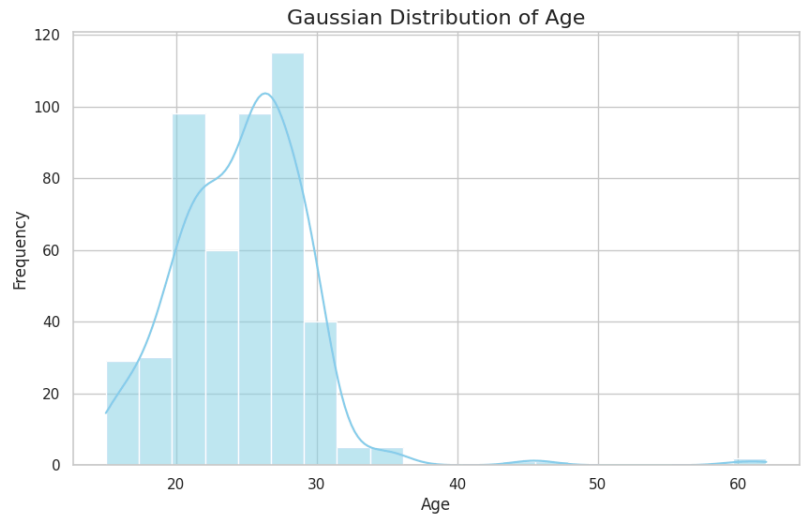



| Feature | Correlation with Target |
|---|---|
| Do you have hair fall problem ? | 1 |
| Is there anyone in your family having a hair fall problem or a baldness issue? | 0.497499 |
| Do you have too much stress? | 0.483251 |

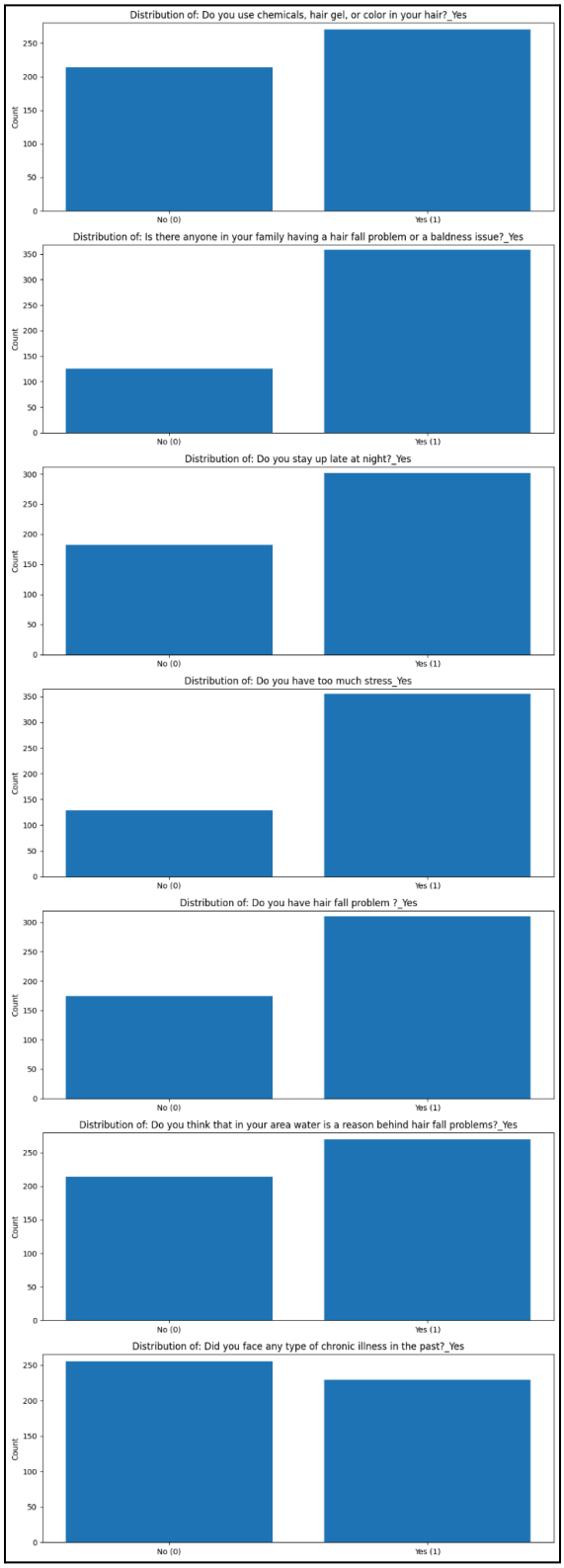
| | |
|---|---|
| Do you think that in your area water is a reason behind hair fall problems? | 0.408046 |
| Did you face any type of chronic illness in the past? | 0.390889 |
| Do you use chemicals, hair gel, or color in your hair? | 0.351738 |
| Do you stay up late at night? | 0.330020 |
| What is your age? | 0.296860 |

## Additional Visualizations

| Visualization | Description |
|---|---|
|   | This graph illustrates the distribution of ages **in the final dataset after merging it with the Google Forms responses,** using a box plot. It highlights the median age, interquartile range, and any potential outliers, which provides a clear overview of how the ages are spread within the dataset. |




This graph illustrates the distribution of participants' ages, following a bell-shaped Gaussian curve, with most responses concentrated around the 20s to early 30s.




This graph shows how many participants answered “Yes” or “No” for each question in the dataset, which helps seeing the distribution, as well as which factors were more common across the responses

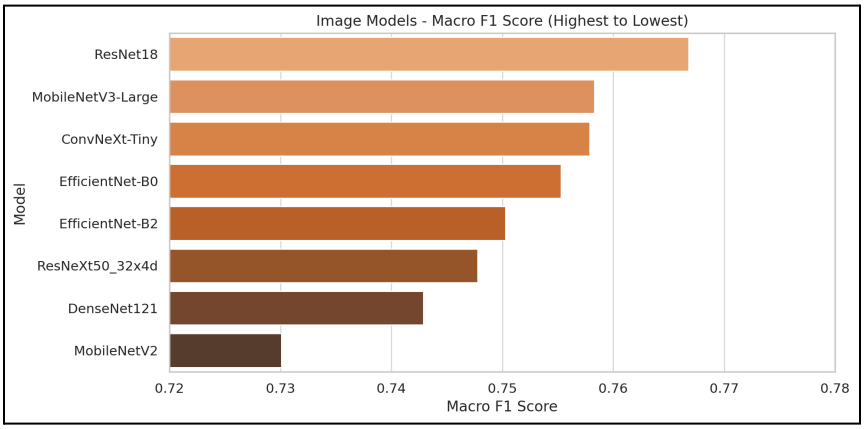
## Classification Report For XGBoost (Accuracy: 0.8144)

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **0 (No Hair Fall)** | 0.71 | 0.67 | 0.69 | 30 |
| **1 (Hair Fall)** | 0.86 | 0.88 | 0.87 | 67 |
| **Accuracy** | - | - | 0.81 | 97 |
| **Macro Average** | 0.78 | 0.77 | 0.78 | 97 |
| **Weighted Average** | 0.81 | 0.81 | 0.81 | 97 |

# Scalp Image Combined Datasets (Roboflow)

## Comparison Between Top 8 Images Models

| Model | Test Accuracy | Macro F1 | Notes |
|---|---|---|---|
| **ResNet18** | 0.86 | 0.7668 | Top performer, lightweight |
| **MobileNetV3-Large** | 0.8582 | 0.7583 | High balance and accuracy |
| **EfficientNet-B0** | 0.8606 | 0.7553 | Efficient and strong |
| **ConvNeXt-Tiny** | 0.8534 | 0.7579 | Strong on rare classes |
| **DenseNet121** | 0.8413 | 0.7429 | Balanced across stages |
| **ResNeXt50_32x4d** | 0.863 | 0.7478 | High accuracy, lower F1 |
| **MobileNetV2** | 0.8293 | 0.7301 | Lightweight, good generalization |
| **EfficientNet-B2** | 0.8438 | 0.7503 | Slightly lower accuracy |




**Bar chart comparing the macro F1 scores of the top 8 image models -** ResNet18 achieved the highest F1 score (0.7668), indicating strong overall performance across Norwood stages. MobileNetV3-Large, ConvNeXt-Tiny, and EfficientNet-B0 followed closely, while MobileNetV2 showed the lowest F1 score among the top models.

## Image Count per Norwood Level (Train/Valid/Test) After Combining Two Roboflow Datasets and Applying Segmentation + CLAHE

This table summarizes the number of images per Norwood level (2-7) in the **train**, **validation**, and **test** folders in our two combined Roboflow image datasets.

| Level | Train | Valid | Test | Total |
|---|---|---|---|---|
| **Level 2** | 3963 | 388 | 210 | 4561 |
| **Level 3** | 2139 | 210 | 95 | 2444 |
| **Level 4** | 1149 | 125 | 51 | 1325 |
| **Level 5** | 712 | 77 | 13 | 802 |
| **Level 6** | 572 | 42 | 20 | 634 |
| **Level 7** | 534 | 53 | 27 | 614 |
| **Total** | **9069** | **895** | **416** | **10380** |

## Improvement from Preprocessing Table (CLAHE + Segmentation)

| Stage | Test Accuracy | Macro F1 |
|---|---|---|
| **Before Segmentation** | ~0.64 | ~0.60 |
| **After Segmentation + CLAHE** | 0.86 | 0.76 |

## Classification Report For ResNet18 Per-Class F1 Scores (Overall F1-Score 0.7668)

| Norwood Level | F1-Score |
|---|---|
| **Level 2** | 0.932 |
| **Level 3** | 0.8308 |
| **Level 4** | 0.7475 |
| **Level 5** | 0.3429 |
| **Level 6** | 0.8696 |
| **Level 7** | 0.9057 |

# Appendix: Data Sources

For this project, we utilize **5 datasets** from different sources:

- **Tabular Model**
  - **Hair Fall Causes Dataset (Mendeley)**
    i. **Source:** Mendeley data repository
    ii. **Description:** This dataset consists of **717 survey responses from male and female participants** regarding hair care habits, health conditions, and lifestyle choices that may influence hair loss. The data was collected via Google Forms distributed across social media and online communities. The survey is available for downloading either as a CSV or XML file. The survey dataset is the sole content of the respiratory.
  - **Google Forms Survey Responses**
    **i.** **Source:** חיזוי נשירת שיער
    **ii.** **Description:** We created a Google Form mirroring the original dataset's features and structure to enrich it. The **24 additional male survey responses** were cleaned, formatted, and merged using Python. After preprocessing, removing irrelevant columns, and excluding female entries we identified strongly correlated features with the target variable. The remaining data resulted in **483 male responses (7 features and 1 target variable).**

- **Images Model**
  - **Scalp image dataset (Roboflow) 1**
    i. **Source:** Roboflow dataset repository
    ii. **Description:** This dataset includes a total of **2,317 images** of male scalps photographed from multiple angles, including front, back, left, right, and top-down views. The images are labeled according to Norwood stages 2-7 that represent various degrees of male pattern baldness. The dataset is split into training (87%), validation (9%), and test (4%) sets, resulting in 2,009 training images, 209 validation images, and 99 test images. To align the labels to our original dataset, we merged **Level 1 and Level 2** in the new dataset and then integrated it with our own.

- **Scalp image dataset (Roboflow) 2**
  - **i. Source:** Roboflow dataset repository
  - **ii. Description:** This dataset contains a total of **8,084 scalp images** captured from a top-down angle, each labeled with a Norwood stage 1-7. The dataset is pre-split into train (87%), validation (8%), and test (4%) sets, resulting in 7,060 training images, 686 validation images, and 338 test images.
- **PyTorch Hair Segmentation GitHub Repository**
  - **i. Source:** YBIGTA/pytorch-hair-segmentation GitHub repository
  - **ii. Description:** This dataset obtained from GitHub contains annotated images for hair segmentation tasks. Each image is paired with a corresponding segmentation mask that outlines the hair region, enabling us to train a model that detects and isolates the scalp area. We used this dataset to build a segmentation pipeline that improved classification performance by focusing the model on relevant scalp regions.