GenEthic Analysis: Building a Secure and Accessible Genetic Analysis Framework

By

Sapir Sharoni, Bachelor of Arts in biology

A thesis submitted to the Graduate Committee of

Ramapo College of New Jersey in partial fulfillment

of the requirements for the degree of

Master of Science in Data Science

Fall, 2024

Committee Members:

Dr. Debbie Yuster, Advisor

Dr. Paramjeet Bagga, Co-Advisor

Dr. Sourav Dutta, Reader

COPYRIGHT

Dedication

I want to dedicate this project to my husband, my unwavering rock throughout the past six years of my academic journey. Your endless support, encouragement, and belief in me have been my greatest strength. Through every challenge, your love and commitment carried me forward. This journey would not have been possible without you, and I am eternally grateful.

Acknowledgments

I would like to express my deepest gratitude to Dr. Debbie Yuster and Dr. Paramjeet Bagga for their invaluable guidance, continuous support, and constant encouragement throughout this research. Their knowledge and (most importantly) their patience have been a source of motivation and an essential part of this project.

I am also grateful to Dr. Sourav Dutta for his time, effort, and willingness to serve on my committee. Their review of my research and participation in this process have been greatly appreciated.

Table of Contents

Dedication 4
Acknowledgments 5
Table of Contents 6
List of Tables 7
List of Figures 8
List of Abbreviations 1
Chapter 1: Abstract 2
Chapter 2: Introduction 3
Chapter 3: Background 5
Chapter 4: Methodology 9
Overview 9
Section A: General Methodology 9
Literature Review and SNP Selection 9
Data Collection and Preparation 11
Data Pre-Processing and Standardization 12
Model Development and Evaluation 13
Section B: Trait-Specific Analysis and Modeling 14
Chapter 5: Analysis and Discussion 29
Chapter 6: Conclusions 43
References 46
Appendices 50
Appendix A: Data Resources 50
Appendix B: List of Model SNPs 51
Appendix C: Reported Eye and Hair Color Classifications 56
Appendix D: Populations Classifications into Subgroups 60
Appendix E: Classification Results for all models 63

List of Tables

Table 1. Replacement dictionary for standardizing genotype notations during preprocessing. .. 12
Table 2. Replacement dictionary for standardizing genotype for the SNP rs8176719 15
Table 3. Blood Type Model - Original class distribution and Adjusted Class Weights 16
Table 4. Eye Color Models - Original class distribution and Adjusted Class Weights 19
Table 5. Red Hair Color Model - Original class distribution and Adjusted Class Weights..... 21
Table 6. Hair Color Model - Original class distribution and Adjusted Class Weights 22
Table 7. Ancestry Model - Original class distribution and Adjusted Class Weights..... 28
Table 8. Classification report for the biological gender prediction model..... 29
Table 9. Gradient Boosting classification report for ABO blood type prediction. 32
Table 10. Gradient Boosting classification report for three-class eye color predictions. 34
Table 11. Gradient Boosting classification report for Blue/Blue-Mixed vs. Green/Light-Mixed eye color predictions. 35
Table 12. Gradient Boosting classification report for Blue/Blue-Mixed vs. Dark-Mixed/Brown eye color predictions. 36
Table 13. Gradient Boosting classification report for red vs. non-red hair color prediction. 38
Table 14. Gradient Boosting classification report for light vs. dark hair color prediction. 39
Table 15. Classification report for ancestry prediction using the Neural Network (MLP) model. 41
Table 16. Data Sources for Genetic Analysis. 50
Table 17. Biological Gender prediction SNPs 51
Table 18. Blood Type Prediction SNPs 52
Table 19. Eye Color Prediction SNPs 53
Table 20. Red Hair vs. Non-Red Hair Prediction SNPs..... 54
Table 21. Light vs. Dark Hair Color Prediction SNPs 54
Table 22. Models' classification results for blood type prediction. 63
Table 23. Models' classification results for the 3-colors hair color prediction dataset. 64
Table 24. Models' classification results for the Blue/Blue-Mixed vs. Green/Light-Mixed subset. 65
Table 25. Models' classification results for the Blue/Blue-Mixed vs. Dark Mixed/Brown subset. 66
Table 26. Models' classification results for red vs. non-red hair color..... 67
Table 27. Models' classification results for Light vs. Dark hair colors. 68
Table 28. Models' classification report for Ancestry..... 69

List of Figures

Figure A. A GWAS search result for ABO blood group associations. 10
Figure B. Study accession details for the ABO blood group (O vs. non-O) trait. 10
Figure C. Comparison of Blood Type Distribution in the Dataset vs. Global Distribution. 16
Figure D. Eye Color Prediction Flowchart 19
Figure E. Hair Color Prediction Flowchart 23
Figure F. ALFA frequency distributions for the alternative allele across major populations (African, Asian, and European) for selected SNPs. 24
Figure G. Dendrogram of the hierarchical relationship between population classes based on the Euclidean distance between them. 27
Figure H. Confusion Matrix for Gender Prediction Model. 30
Figure I. Confusion matrix for Gradient Boosting model in blood type prediction. 32
Figure J. Confusion matrix for Gradient Boosting model in eye color prediction. 34
Figure K. Confusion matrix for Gradient Boosting model in Blue vs. Green eye color prediction. 36
Figure L. Confusion matrix for Gradient Boosting model – Blue vs. Brown eye color prediction. 36
Figure M. Confusion matrix for Gradient Boosting model – red vs. non-red hair color prediction. 38
Figure N. Confusion matrix for Gradient Boosting model - light vs. dark hair color prediction. 39
Figure O. Confusion matrix for MLP Model in ancestry prediction. 41

List of Abbreviations

4x	ALFA	Allele Frequency Aggregator
11x	DNA	Deoxyribonucleic acid
2x	DTC	Direct-to-consumer
5x	GWAS	Genome-wide association studies
3x	HGDP	Human Genome Diversity Project
4x	MLP	Multilayer perceptron (neural network)
3x	MLR	Multinomial logistic regression
13x	rsID/s	A unique identifier used by researchers and databases for each SNP.
2x	SGDP	Simons Genome Diversity Project
82x	SNP/s	Single nucleotide polymorphism/s
8x	SRY	Sex-determining region Y
1x	VCF	Variant Call Format
1x	WSL	Windows Subsystem for Linux

Chapter 1: Abstract

The latest advances in genetic research have paved the way for innovative new applications of genetic data in areas such as ancestry research, forensic science, and medicine. However, the current Direct-To-Consumer (DTC) genetic platforms often have limited accessibility and utility, posing significant challenges for researchers and other professionals. Furthermore, concerns about the privacy and security of data within popular DTC companies persist among users.

To address these limitations, a framework was developed for a predictive genetic analysis tool that prioritizes privacy, security, and user-friendliness. This study focused on predicting observable traits, including ancestry, biological sex, blood type, and eye and hair color, using single nucleotide polymorphisms (SNPs). A machine-learning-driven methodology was employed, integrating data preprocessing, standardized genotype encoding, and model evaluation. Models such as Gradient Boosting and Neural Networks were used to predict traits, demonstrating high accuracy across categories, including blood type and population groups (96%). The results demonstrate that using the proposed framework it is feasible to create a genetic analysis tool capable of bridging the gap between privacy and security and practical usability. It is important to note that the framework presented is adaptable, enabling its application across various industries. While this study focused on observable traits, future research can extend to various domains.

Keywords: genomics, genetic platforms, direct-to-consumer (dtc), predictive genetic analysis, privacy, data security, observable traits, ancestry, machine learning, single nucleotide polymorphisms (snps), gradient boosting, neural networks, blood type, forensic science.

Chapter 2: Introduction

The field of genomics has been growing rapidly in the last decade, leading to numerous innovations in various fields like healthcare, ancestry research, and forensic science. Applications of genetic research are at the forefront of everything from agriculture and food security to cancer research and criminal investigations [1]. Part of these advancements are Direct-To-Consumer (DTC) genetic testing services, which began to emerge in the early 2000s [2]. The growing availability of private and relatively inexpensive genetic testing enabled many individuals to explore their ancestry and genetic makeup. However, while this allows individuals the opportunity to peek into their genetics, genetic analysis tools disproportionately revolve around customer use despite their numerous applications in various fields. Existing platforms and services such as 23andMe, AncestryDNA, MyHeritage, and even Promethease are heavily restricted in terms of data availability, authorized usage, and the information they provide [3–6]. Given that these platforms hold massive amounts of DNA samples and analyzed data, such restrictions pose an obstacle for other domains, including research, forensics, and healthcare. In addition, some tools present overly complex reports, requiring prior knowledge in genetics to interpret them.

Tools like Promethease, while technically elaborate, are often inaccessible to those with limited knowledge of genetics. Promethease generates reports that contain tens of thousands of entries, presenting raw, uncontextualized genetic information (approximately 25,000 genetic associated traits)[7] that can overwhelm users without a strong scientific background. Additionally, the interface of Promethease is criticized for its lack of user-friendliness, failing to present findings in a clear or actionable format. Compounding these issues is the platform's failure to

adequately disclose the limitations and uncertainties of its analyses, which may lead to misinterpretation or misuse of the information provided [7–9].

Beyond technical complexity, there are many concerns regarding data privacy, management, and usage after being collected by the analyzing service. The rise of consumer genetic analysis companies has raised significant privacy and data protection concerns. For example, in October 2023, 23andMe suffered a major data breach, with nearly 7 million of its users' genetic information being stolen [10]. Concerns about privacy also apply to the ownership and storage of data. In 2019, for example, MyHeritage acquired Promethease and the database it was associated with, SNPedia, leading to the transfer and storage of users' genetic profiles into MyHeritage's centralized databases [11,12]. More recently, this year (2024), 23andMe was reported to face a possible purchase following financial struggles, which raises alarm over the possible transfer of customers' genetic data to the acquiring company [13,14].

This research addresses the critical need for a transparent, user-friendly genetic analysis platform by developing a framework for genetic analysis that prioritizes privacy and security. The proposed framework is designed to ensure that user data remains private and secure, never stored beyond the analysis process, thereby alleviating concerns of unauthorized use or breaches. Although this study was focused on the analysis of observable and measurable traits, which are currently predominantly applicable in forensic contexts, the underlying framework has the potential to be extended to other industries, including healthcare, nutrition, and ancestry research.

Throughout this paper, I hope not only to demonstrate the feasibility of such a framework but also to establish a foundation for future genetic analysis research that would foster a balance between privacy, utility, and ethical data management.

Chapter 3: Background

The study of genes and heredity has transformed our understanding of biology, medicine, and human diversity. The identification and interpretation of DNA variations that contribute to observable traits and biological functions, are a key part of this field. Among these variations, single nucleotide polymorphisms (SNPs) have emerged as critical markers for understanding human genetics. SNPs represent variations at a single nucleotide position in the genome and provide valuable insights into the genetic basis of inherited traits, susceptibility to diseases, and individual responses to medications.

Human heredity occurs through the transmission of genetic material from parents to offspring, encoded within the structure of DNA. DNA is composed of nucleotides, each consisting of a sugar molecule, a phosphate group, and one of four nitrogenous bases: adenine (A), thymine (T), cytosine (C), or guanine (G). The sequence of these bases form genes, which encode instructions for protein synthesis. Proteins are constructed from amino acids, each specified by a three-letter codon in the DNA sequence, with specific codons signaling the start and stop of protein synthesis. Single nucleotide polymorphisms (SNPs) are variations at a single nucleotide position in the genome resulting from different types of mutations, including substitutions, insertions, or deletions, which may be synonymous (silent), missense (changing an amino acid), or nonsense (introducing a stop codon). These variations can influence protein structure and function, leading to differences in traits or susceptibility to disease. An allele refers to one of the alternative forms of a gene at a specific location (also called locus), often differing by one or more base pairs, which can result in varying effects on traits or phenotypes. Humans are diploid organisms, meaning they inherit two copies of each chromosome, one from each

parent. Therefore, humans carry two alleles at most locations (loci, plural for locus). If the two alleles are identical, the individual is considered homozygous at that locus and if the two alleles are different, they are considered heterozygous.

Genome-wide association studies (GWAS) leverage SNP data to identify genetic variants associated with specific traits or diseases [15]. Databases like dbSNP and initiatives such as the 1000 Genomes Project have significantly enriched our knowledge of human genetic diversity [16,17]. dbSNP provides a repository of known SNPs, while the 1000 Genomes Project catalogs genetic variations across global populations, offering a foundation for exploring genetic contributions to health and disease. Together, these resources facilitate genetic studies that link SNPs to observable phenotypes, including physical traits, disease susceptibility, and pharmacogenomic responses.

Single nucleotide variations play a significant role in determining a wide range of phenotypes, including observable phenotypes such as blood type, height, skin tone, hair color, and eye color [18]. These phenotypes can generally be categorized as single-gene traits or polygenic traits. Single-gene traits are determined primarily by variations in a single gene, whereas polygenic traits often result from cumulative interactions between variations in multiple genes. Specific SNPs have been linked to these traits through genome-wide association studies (GWAS), providing insights into the genetic mechanisms underlying human genetic diversity [19].

One example of single-gene inheritance in humans is the ABO blood group system. ABO blood groups are determined by the presence or absence of specific antigens on red blood cells. These antigens, encoded by the ABO, RH, and other blood group systems, are influenced by SNPs that alter glycosyltransferase activity [20]. Studies have identified SNPs in the ABO gene as critical determinants of blood type [21]. Similarly, the Rhesus (Rh) factor, an inherited protein that

can be found on the surface of the red blood cell, has been linked to SNP variations, demonstrating the straightforward inheritance patterns of single-gene traits [22].

Alternatively, physical phenotypes such as hair and eye color are classic examples of polygenic traits. These traits are influenced by interactions between several genes associated with variations in melanin synthesis and transport. Some of the key genes include OCA2 and HERC2, which are often associated with eye color; and MC1R, which effects the production and ratio of eumelanin (dark pigment) to pheomelanin (light pigment) in hair and skin [23,24]. Other genes, such as SLC24A5 and SLC24A4, have been shown to effect intracellular transporters important for the regulation and synthesis of melanin [25]. The diversity of colors, from blond to black hair and blue to brown eyes, arises from variations in these genes, often linked to specific SNPs, other variants, their interactions with regulatory elements, and environmental factors.

A broader view of genetic variation and SNP distribution can be seen in population genetics studies. Population genetics studies examine how genetic composition varies across populations under the influence of evolutionary processes such as natural selection, mutation, genetic drift, and gene flow [26]. In humans, the distribution of SNPs among populations can help us better understand the diversity of human ancestry, health, and appearance. It can also help us uncover patterns of environmental adaptation, ancestral migration, and the genetic foundation of population-specific traits. Key resources, such as the 1000 Genomes Project, the Human Genome Diversity Project (HGDP), and the Simon Genome Diversity Project, have established comprehensive databases of global genetic variation [17,27,28]. These resources can be further supplemented by other tools like the Allele Frequency Aggregator (ALFA), which provides comparative allele frequency data across major population groups [29].

Beyond ancestry, traits, and health, genetic research also has significant applications in forensics, including forensic genetic genealogy. In forensics, DNA analysis serves as a powerful tool, with applications that include identifying unknown individuals, tracking missing persons, and assisting in criminal investigations. In forensic genetic genealogy, SNP data is combined with ancestral records to trace familial lineages and use them to identify unknown individuals.

Yet, despite the potential of these methods, the use of consumer genetic platforms by law enforcement remains extremely limited. Private DNA analysis companies rarely and reluctantly cooperate with law enforcement, despite research indicating that, depending on the case type, the majority of surveyed adults support police access to private DNA databases [30,31]. Ironically, despite their stringent data-sharing guidelines and reluctance to disclose information with law enforcement, these companies have a history of security and privacy issues and have shown that they are willing to transfer user data completely during business acquisitions [32,33].

Consumer genetic platforms such as 23andMe and AncestryDNA have popularized genetic testing by offering users insights into their ancestry, health predispositions, and traits. Despite having made genetic data more widely available, these platforms usually operate inside restricted frameworks, which restricts the use of genetic data by forensic labs and law enforcement. Additionally, storage and sharing of genetic data continues to raise ethical concerns, including privacy issues, potential misuse, and security vulnerabilities.

These limitations highlight the growing need for a framework that bridges the gap between the complexity of genetic data and its practical applications in secure, private, and user-friendly formats. This thesis aims to demonstrate the feasibility of designing a transparent, secure, and intuitive framework for a genetic analysis tool, which can be adapted to serve versatile functions across different fields.

Chapter 4: Methodology

Overview

Designed to generate insights based on genetic data, this thesis initially focused on physical characteristics, such as ancestry, biological sex, eye color, hair color, and blood type. These predictions have potential applications in areas such as forensic investigations, where reconstructing physical traits can aid in solving cases. Importantly, the methods discussed in this chapter can be applied to a wide range of attributes, enabling future revisions to incorporate genetic insights relevant to other fields like healthcare, nutrition, or personalized wellness.

To develop the app, a quantitative approach was employed, integrating GWAS, population genetics, and machine learning. The methodology was carried out in three distinct phases: literature review and SNP selection, data collection and pre-processing, and data processing and model development.

Section A: General Methodology

Literature Review and SNP Selection

The first phase of the study involved identifying SNPs associated with traits of interest. A detailed review of existing research was conducted, drawing from GWAS results and population genetics studies. Using GWAS, trait-specific SNPs were searched via the main search engine. Included in the search results (figure A) is the “Study Accession,” which links the study reporting this association (Figure B).

Figure A. *A GWAS search result for ABO blood group associations.*

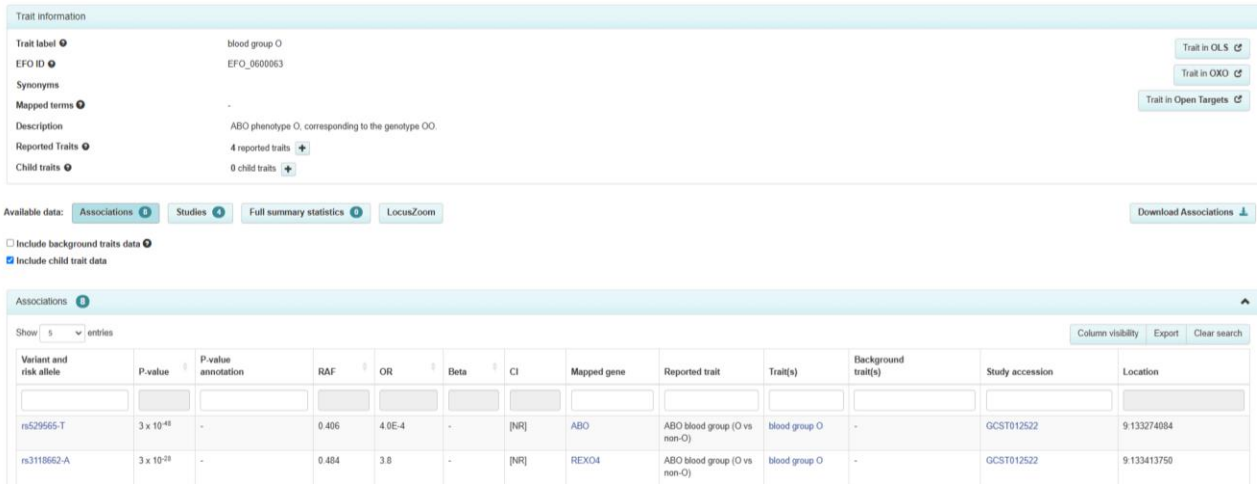
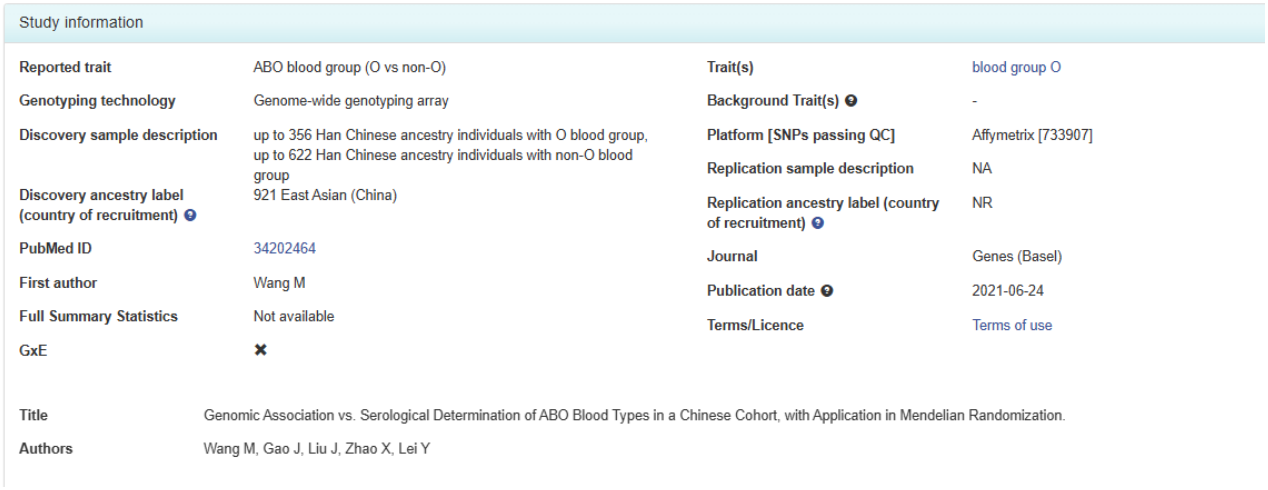


Figure B. *Study accession details for the ABO blood group (O vs. non-O) trait.*



Reporting studies were reviewed for clarification and validity and key SNPs were selected based on their established associations with specific traits. For instance, SNPs in the ABO gene were identified for blood type determination, while MC1R, HERC2, and OCA2 were included for hair and eye color. For ancestry, SNPs associated with population-specific markers were identified by comparing allele frequencies in different populations using ALFA. Gender determination relied on SNPs located on the Y chromosome, particularly within the SRY region. A comprehensive SNP reference file was compiled for each trait, containing rsIDs, effect alleles, other alleles, and the associated phenotypes. The effect allele was defined as the allele associated

with the reference phenotype and used as the basis for encoding. Conversely, the other allele refers to the alternative allele that is either negatively associated with, or not associated with, the reference phenotype. For example, in the case of red hair color, the reference phenotype was set as red hair. The effect allele in this context is the allele that contributes to or is associated with the presence of red hair, while the other allele would be the alternative allele not associated with red hair. To ensure accuracy, the effect and other alleles were cross-referenced with dbSNP to verify consistency in rsID assignments and notations [34].

Data Collection and Preparation

Genetic data samples were obtained from publicly available sources. For ancestry analysis, datasets from the 1000 Genomes Project, the Simons Genome Diversity Project (SGDP), and the Human Genome Diversity Project (HGDP) were utilized. These datasets are provided in Variant Call Format (VCF) and include rsIDs, genotypes, and population group information for each sample [17,27].

User-contributed datasets from openSNP were used to obtain data on hair color, eye color, and blood type [35]. These datasets are uploaded in text format and combine genotype information from platforms such as 23andMe, Ancestry.com, and Illumina with self-reported phenotypes [35]. Since the majority of the samples were taken from 23andMe, user-contributed samples were standardized to conform to the 23andMe format and converted into a CSV format.

To prepare the data for analysis, raw genetic files were organized into a directory structure by user ID. SNP genotypes were extracted for each individual and combined with phenotype data in a unified format. The resulting datasets included user IDs, reported phenotypes, and genotypes for relevant SNPs, forming the foundation for subsequent processing.

Data Pre-Processing and Standardization

Data preprocessing was a crucial step to ensure the consistency and quality of genetic samples. In DNA, the "forward (+) strand" refers to the sequence read in the 5′-to-3′ direction from left to right, while the "reverse (–) strand" refers to the complementary sequence written in the 5′-to-3′ direction from right to left.

To maintain uniformity, all genotype samples were standardized to the forward strand notation, ensuring alignment with standardized genetic notations. For each SNP, the rsID and alleles were validated against the rsID, effect allele, and other allele listed in the SNP reference file. Missing values were marked as 'NN' to simplify data processing. Additionally, the order of alleles for each SNP was standardized to ensure consistent representation across samples, enhancing readability for upcoming analyses. Table 1 describes the allele notations used for standardization. These notations represent diploid genotypes, which refer to the pair of alleles present at the same locus on homologous chromosomes. Homologous chromosomes contain the same genes in the same order but may exhibit allele variation between them. A diploid genotype can be homozygous, where both alleles are identical (e.g. AA), or heterozygous, where the alleles differ (e.g. AG). As the models are based on the quantity of alleles associated with the reference phenotype, the writing order is inconsequential. AG and GA, for instance, are equivalent notations. This standardization guarantees consistent representation across datasets.

Table 1. Replacement dictionary for standardizing genotype notations during preprocessing.

Original value	***Standardized value***
--, 0, ??	NN
GA	AG
GC	CG
TG	GT
TC	CT
TA	AT
CA	AC

Columns with more than 1/2 to 1/3 missing data were removed to ensure the reliability of the analyses and prevent skewing due to excessive missing values. In addition, rows with invalid genotypes, missing phenotypes, or excessive missing data (>33% for most traits) were also filtered out.

Since no predefined options were established, users of openSNP were able to freely enter any value for their phenotype. Therefore, to minimize variability and avoid duplication, reported phenotypes were standardized and categorized into broader classes.

Model Development and Evaluation

The data was encoded to allow for machine learning analysis. For most traits, homozygotes for the effect allele were encoded as 2, heterozygotes as 1, and homozygotes for the other allele as 0. Gender determination used a binary encoding system: SNPs in the SRY region were encoded as 1 if present and 0 otherwise.

Five machine learning models were implemented to predict traits: Logistic Regression, Random Forest, Gradient Boosting, Decision Tree, and a Neural Network. For imbalanced datasets, weighted sampling was used to distribute the class weights equally. Class weights were calculated using the compute_class_weight function from sklearn.utils.class_weight, which computes the weights for each class and assigns higher weights to underrepresented classes and lower weights to overrepresented ones [36]. Class weights were calculated using the formula:

$$w_i = \frac{n_{samples}}{n_{classes} * n_i}$$

Where $w_i =$ the weight for class i, $n_{samples} =$ the total number of samples, $n_{classes} =$ the total number of classes, and $n_i =$ the number of samples in class i

Using weighted sampling ensured that each class contributes equally during the model training process regardless of its representation in the dataset, helping preserve the overall

distribution of genotypes. Alternative balancing techniques, including Synthetic Minority Oversampling Technique (SMOTE) and Undersampling, were assessed. However, these yielded inferior outcomes in comparison to weighted sampling. A 10-fold stratified Cross-validation was used to evaluate the models. Each model was evaluated based on the F1 scores and accuracy metrics. The best-performing model for each trait was selected based on its performance across these metrics.

Section B: Trait-Specific Analysis and Modeling

I. Biological gender determination

Biological gender was determined by identifying SNPs located in the SRY (sex-determining region Y) region on the Y chromosome. This region plays a critical role in male differentiation, and the presence of key SNPs within this region indicates a biologically male individual. Conversely, the absence of such SNPs confirms a biologically female individual. SNPs were selected using the dbSNP, filtering for SNPs within the SRY region (see Appendix, B table 12). Four SNPs were included to provide alternative options in case of different mapped SNPs.

Genetic samples and reported gender were obtained from openSNP, resulting in an initial dataset of 1016 samples.

The data was filtered to remove samples from Males that did not have any data associated due to incomplete data files (21). In addition, samples with unclear reports of biological gender were removed (2). Overall, twenty-three samples were filtered out, yielding a dataset of 993 samples – 523 Biological Males and 470 Biological Females.

II. Blood type prediction:

SNP Selection: Blood type prediction was based on 15 SNPs. Originally, 27 SNPs were considered, but twelve were excluded due to insufficient data (see Appendix B, table 13).

Data Collection: Genetic samples and reported blood types were obtained from openSNP, resulting in an initial dataset of 868 samples. Data cleaning included the removal of columns with over 50% missing data and filtering out users without critical SNP data (rs8176719 or rs505922). Since most of the samples were missing data for Rh group (+/-) prediction, blood types were standardized into four classes: O, A, B, and AB, with 5 users excluded due to unclear reported values. After cleaning and grouping the classes, the dataset was reduced to 726 samples.

Standardization: Genotypes were standardized as described in the Data Pre-Processing and Standardization section, using the replacement dictionary provided in Table 1. For rs8176719, where the effect allele is a deletion versus insertion, a specific encoding technique was used.

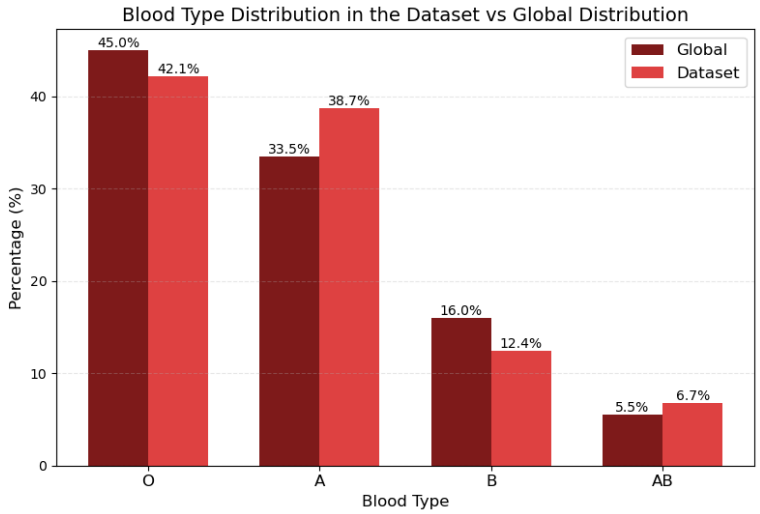
Table 2. *Replacement dictionary for standardizing genotype for the SNP rs8176719*

Original Value	Standardized Value
TT	DI
TC	II
CC	II
-C	II
-T	DD
T	DD
C	II
--	NN

The distribution of blood types within the dataset closely aligned with the global distribution (Figure C), which highlights the dataset's diversity and representation of real-world blood type distribution [37].

However, to lessen biases against overrepresented classes, weighted sampling was employed. By giving less weight to overrepresented classes and more to underrepresented ones, weighted sampling adjusts for class imbalances, ensuring that every class is fairly considered by the model, regardless of its representation in the dataset.

Figure C. *Comparison of Blood Type Distribution in the Dataset vs. Global Distribution.*



Model Training and Evaluation: Genotypes were encoded as described in section A, using O and A blood types as reference classes associated with the "effect allele." Five machine learning models were trained on the weighted training set (Table 3): Multinomial Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, and Neural Network.

Table 3. *Blood Type Model - Original class distribution and Adjusted Class Weights*

Type	Original class distribution	Adjusted Class Weights
A	281	0.65
AB	49	3.70
B	90	2.02
O	306	0.59

Model performance was evaluated using F1 scores and accuracy metrics on a 10-fold stratified cross-validation.

III. Eye color prediction:

SNP Selection: Eye color prediction utilized 36 SNPs, predominantly located on chromosome 15, a region strongly associated with eye color determination (see Appendix B, table 14). Following data cleaning, SNPs with more than one-third of missing genotype data were excluded, reducing the set to 22 SNPs.

Data Collection: Genetic data samples and self-reported eye colors were obtained from openSNP. The initial dataset contained 1,805 samples. However, after filtering out invalid genotypes and rows with more than one-third of the genotypes missing, 1,441 samples remained.

Standardization: Eye colors were self-reported without predefined categories, requiring classification into three distinct classes based on the Martin-Schultz Scale [38]:

- Class 1: Blue/Blue Mixed (Martin-Schultz Scale 1–5)
    - Class 1 included reported colors such as: “Blue with a yellow ring of flecks that make my eyes look green depending on the light or my mood.” and “Ice blue mixed with slate blue, with an amber pupil burst in both eyes and a brown spot adjacent to lower left pupil. eyes were green into my 20's.”
- Class 2: Green/Light-Mixed (Martin-Schultz Scale 6–8)
    - Class 2 included reported colors such as: “Green with blue halo” and “Green with amber burst and gray outer ring.”
- Class 3: Dark-Mixed/Brown (Martin-Schultz Scale 9–16)
    - Class 3 included reported colors such as: “Black” and “Brown-(green when external temperature rises).”

The complete reported eye color classifications can be reviewed in appendix C.

Samples with inconsistent or invalid color entries were removed, resulting in a final dataset of 1,432 samples. Additionally, since eye colors are highly complex and not always distinct, two subsets were created, one comparing only brown and blue eye colors and another comparing blue and green eye colors.

Class Distributions:

- Dark-Mixed/Brown: 713
- Blue/Blue-Mixed: 517
- Green/Light-Mixed: 202

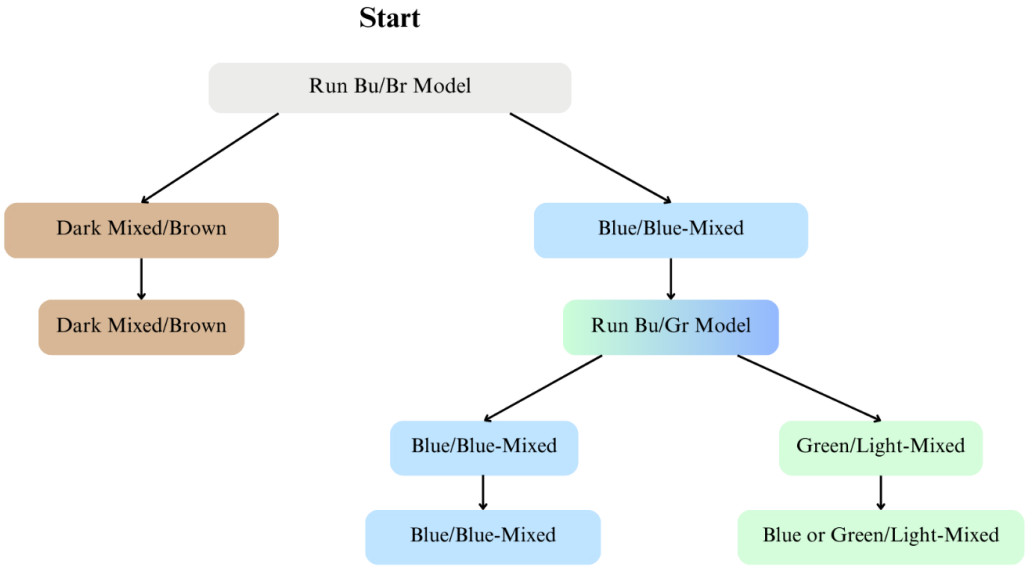
Model Training and Evaluation: three models were created: (1) A model containing all three color-classes (BuGrBr model), (2) Blue/Blue-Mixed vs. Green/Light-Mixed (Bu/Gr model), and (3) Blue/Blue-Mixed vs Dark Mixed/Brown (Bu/Br model). with the Bu/Br model being executed first.

Results from the two eye color models (Bu/Gr and Bu/Br) are meant to be integrated as follows:

1. If the result from the Bu/Br model is "*Dark Mixed/Brown"*, the prediction is "*Dark Mixed/Brown*" regardless of the Bu/Gr model's results.
2. If the result is "*Blue/Blue-Mixed*" in both models, the prediction is "*Blue/Blue-Mixed.*"
3. If the result is "*Blue/Blue-Mixed*" in the Bu/Br model and "*Green/Light-Mixed*" in the Bu/Gr model, the prediction should be "*Blue or Green.*"

Figure D below provides a visual outline of the process of eye color prediction.

Figure D. *Eye Color Prediction Flowchart*



Genotypes were standardized as described in section A. The dataset with all three classes was encoded with reference to the darker color, as it is often more dominant. Similarly, the blue and brown subset was encoded with reference to brown. However, the Green vs. Blue subset was encoded with reference to blue, as green phenotypes are more complex and often consist of a mixture of several colors. In addition, just ~2% of people worldwide have green eyes, making them extremely rare [39]. Weighted sampling was used to address imbalances between classes (Table 4), assigning higher weights to underrepresented classes.

Table 4. *Eye Color Models - Original class distribution and Adjusted Class Weights*

Model set	Original class distribution	Adjusted class weights
3-color classes set	Blue/Blue-Mixed: 517	Blue/Blue-Mixed: 0.92
	Dark Mixed/Brown: 713	Dark Mixed/Brown: 0.67
	Green/Light-Mixed: 202	Green/Light-Mixed: 2.36
Blue vs. Green subset	Blue/Blue-Mixed: 517	Blue/Blue-Mixed: 0.70
	Green/Light-Mixed: 202	Green/Light-Mixed: 1.78
Blue vs. Brown subset	Blue/Blue-Mixed: 517	Blue/Blue-Mixed: 1.19
	Dark Mixed/Brown: 713	Dark Mixed/Brown: 0.86

Five machine learning models were trained and evaluated: Multinomial Logistic Regression, Random Forest, Gradient Boosting, Decision Tree, and Neural Network. Performance was assessed using F1 scores, and accuracy metrics for a 10-fold stratified cross-validation.

IV. Hair Color Prediction

Hair color prediction was initially divided into two parts, with the first focusing on the classification of red hair versus non-red hair. This is due to the specific genes associated with red hair.

Red vs. Non-Red Hair:

SNP Selection: Thirteen SNPs associated with red hair were selected for this analysis (see Appendix B, table 15).

Data Collection: Samples were obtained from openSNP, including genotype data and user reported phenotypes.

Standardization: The data was cleaned and standardized as described in section A. Invalid genotypes and rows with more than 50% missing data were removed. Overall, 32 samples were removed from the dataset, making the final sample size 160.

User-reported responses regarding red hair were standardized into two categories: (1) Yes and (2) No, ensuring consistency in phenotype reporting.

The distribution of red versus non-red hair in the dataset was as follows:

- Red Hair (Yes): 49 samples
- Non-Red Hair (No): 111 samples

Model Training and Evaluation: Genotypes were encoded as described in the Data Pre-Processing and Standardization sub-chapter, with the "effect allele" representing red hair. Weighted sampling was applied to address class imbalances during training (Table 5).

Table 5. *Red Hair Color Model - Original class distribution and Adjusted Class Weights*

Red presence	Original class distribution	Adjusted class weights
No	111	0.72
Yes	49	1.63

Five machine learning models were trained and evaluated: Multinomial Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, and Neural Network. Performance was evaluated using F1 scores and accuracy metrics using a stratified 10-fold cross-validation.

Brown vs. Blonde Hair:

SNP Selection: Forty-five SNPs associated with hair color were initially selected, including many that are linked to key genes such as MC1R, one of the most studied genes associated with hair color [40]. After filtering SNPs with more than one-third of the data missing, 40 SNPs were retained (see Appendix B, table 16).

Data Collection: Genetic samples and reported phenotypes were obtained from openSNP, yielding an initial dataset of 1178 samples. Rows with invalid genotypes, incomplete phenotype reporting, or more than one-third of missing genotype data were filtered out. Red-haired individuals' samples were excluded from this analysis. Overall, 70 rows were excluded from the dataset, yielding a final sample size of 1066.

Standardization: The dataset was standardized as described in Section A. User-reported hair colors were not pre-set to specific categories, resulting in many diverse descriptors. To address this, colors were grouped into two broad categories: (1) Brown to Dark Brown/Black, and (2) Blonde to Dark Blonde

Class Distribution:

- Brown to Dark Brown/Black: 874 samples
    - This class included reported colors such as: “Black” and “Brown going to white in early 40s”
- Blonde to Dark Blonde: 192 samples
    - This class included reported colors such as: “Blond as a child. dark blond as an adult.” and “Dishwater blonde”

The complete reported hair color classifications can be reviewed in appendix C.

Model Training and Evaluation: Genotypes were encoded using the "effect allele," with lighter hair (blonde) as the reference class. Weighted sampling was applied during model training to address the class imbalance and ensure fair representation of both hair color categories (Table 6).

Table 6. *Hair Color Model - Original class distribution and Adjusted Class Weights*

Hair Color	Original class distribution	Adjusted class weights
Blonde to Dark Blonde	192	2.78
Brown to Dark Brown/Black	874	0.61

Five machine learning models were trained and evaluated: Multinomial Logistic Regression, Random Forest, Decision Tree, and Gradient Boosting. Model performance was evaluated on a 10-fold stratified cross-validation, using F1 scores and accuracy metrics.

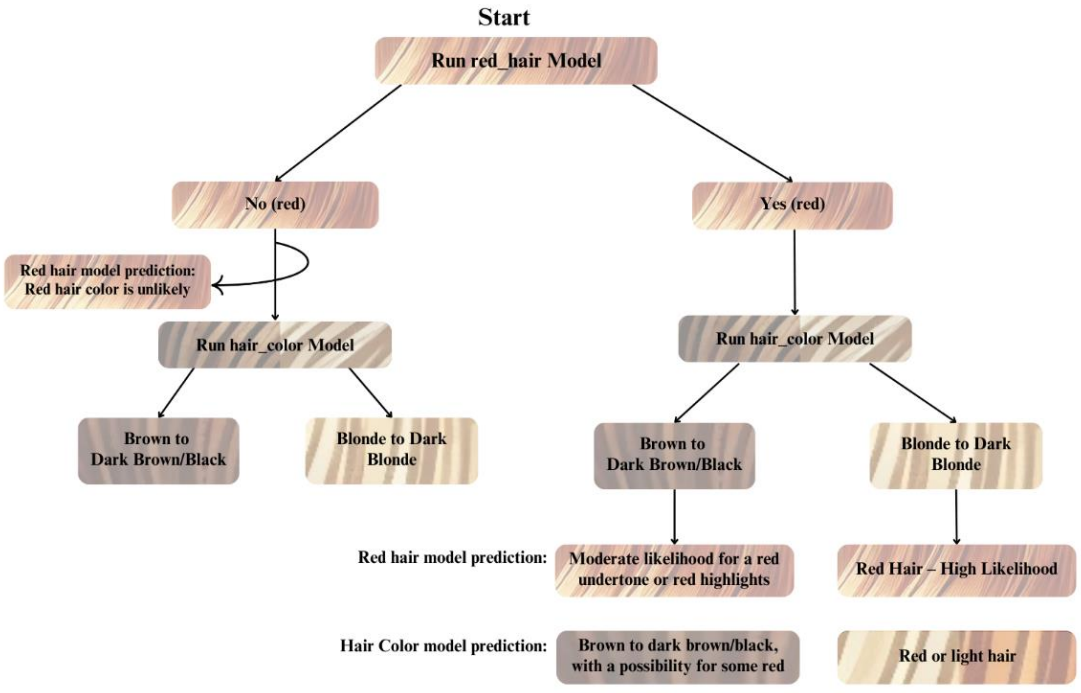
Prediction order of Red vs. Blonde/Brown hair colors:

Red hair is determined first via the red_hair model, followed by the hair_color model for blond vs. brown predictions:

1. If the red hair result is "*non-red*," the prediction for red hair should be "*Red hair color is unlikely*" and the hair_color model result is the determining one.
2. If the red_hair model's result is "*Yes*" and the hair_color model predicts *Blonde to Dark Blonde*, the prediction for red hair should be "*Red Hair – High Likelihood*" and the prediction for hair color should be "*Red or Light hair color*"
3. If the red hair result is "*Yes*" and the hair_color model predicts *Brown to Dark Brown/Black*, the prediction for red should be "*Red – Moderate likelihood for a red undertone or red highlights,"* and the prediction for hair color should be "*Brown to Dark Brown/Black, with a possibility for some red*"

Figure E below provides a visual outline of the process for hair color prediction.

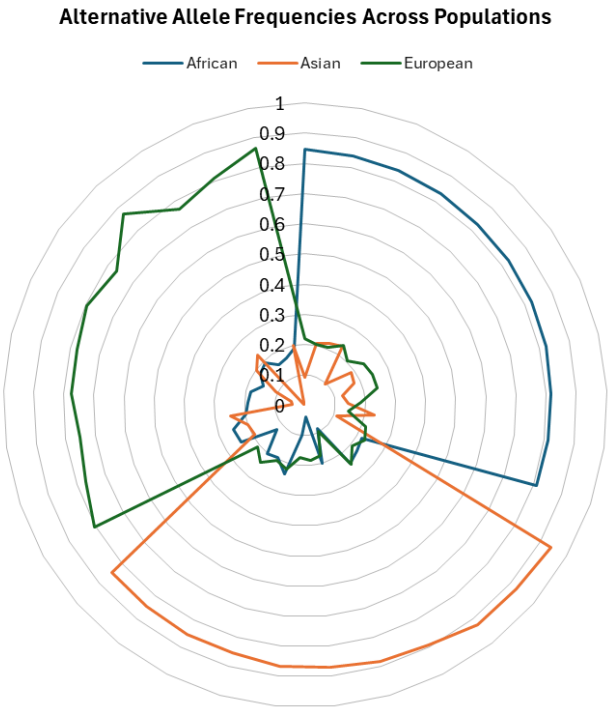
Figure E. *Hair Color Prediction Flowchart*



V. Ancestry Prediction:

SNP Selection: Ancestry prediction was based on an initial set of 169 SNPs. After cleaning and standardization, 164 SNPs were retained for analysis. The selection process focused on SNPs with the highest potential to distinguish populations based on their alternative allele frequencies across these groups. Using the ALFA database, allele frequency data across three major populations was obtained: African, Asian, and European. This enabled the identification of SNPs that exhibit significant frequency differences between populations. Figure B illustrates the distribution of alternative allele frequencies for selected SNPs across major populations (African, Asian, and European). This radar plot illustrates the differences in allele frequencies, which was used to select SNPs that maximize genetic distance.

Figure F. *ALFA frequency distributions for the alternative allele across major populations (African, Asian, and European) for selected SNPs.*



Data Collection: Genetic data was obtained from three publicly available resources: the 1000 Genomes Project, Simons Genome Diversity Project (SGDP), and Human Genome Diversity Project (HGDP). The dataset consisted of 4221 samples, each containing sample ID, genotype information, group (broad population category such as America and Central-South Asia), and population (specific country or region such as Colombia and Pakistan - Balochi). To address the high specificity of the population column, a new column called subgroup was added. Subgroup classifications were intermediate, reflecting more specificity than the broad group but less than the population column (for example, Central and South America and South Asia). Subgroups were assigned based on the population column, considering geographic and genetic similarities.

Standardization: The dataset was standardized as described in Section A. SNP columns with more than one-third of the missing data were removed, resulting in the removal of 5 SNPs. Rows with more than one-third of missing genotype data were also removed. Overall, 394 samples were removed, yielding a dataset of 3827 samples.

The class distribution within subgroups was as follows:

Subgroup	count
African	1030
East Asia and China	909
European	688
South Asian	583
Central and South American	277
Middle East	163
Native Americans	86
Northeast Asia and Siberia	42
Oceania	30
Eastern European and West Eurasia	19

To evaluate the genetic distinction among subgroups, the Euclidean distance was calculated, as it was found to approximately reflect genetic relationships between populations [41]. The Euclidean distance is a measure of the straight-line distance between two points. In the context of this study, the genetic distance was calculated based on the difference in pairwise allele proportions between subgroups. The proportions referred to are the proportions of each pairwise allele combination within a subgroup compared to the other combinations within that subgroup. The difference in proportions was calculated for each SNP and squared to avoid negative values. The squared values across all SNPs were then summed and the square root of this sum was taken to produce a single distance value, which represents the genetic difference between two subgroups. Distances were computed using the following formula [42]:

$$\delta_{AB} = \left\{\sum_i (x_{Ai} - x_{Bi})^2\right\}^{\left(\frac{1}{2}\right)} \qquad [Formula\ A]^{41}$$

Where $A = population\ subgroup\ A, B = population\ subgroup\ B$, and x_{Ai} and x_{Bi} are the proportions of pairwise allele combinations for SNP i of two different subgroups. Calculation example for two SNPs (rs10079352 and rs10145636) between African and Central and South American subgroups:

rsid	Allele combinations	African	Central and South American
rs10079352	AA (%)	81.826	19.134
	AG (%)	9.043	47.292
	GG (%)	9.131	33.574
	NN (%)	0	0
rs10145636	AA (%)	52.766	29.964
	AG (%)	40.035	47.653
	GG (%)	7.112	22.383
	NN (%)	0.0878	0

The first step was to convert the percentages into proportions:

$$\frac{alleles_{SNP}\%}{100}$$

Next, the squared difference between the proportions of pairwise alleles was calculated:

$$\left(\text{rs10079352}_{\text{Afr(AA)}} - \text{rs10079352}_{\text{centamerica(AA)}}\right)^2$$

rsid	Allele combinations	Squared difference
rs10079352	AA	0.393
	AG	0.146
	GG	0.060
	NN	0.000
rs10145636	AA	0.052
	AG	0.006
	GG	0.023
	NN	0.000

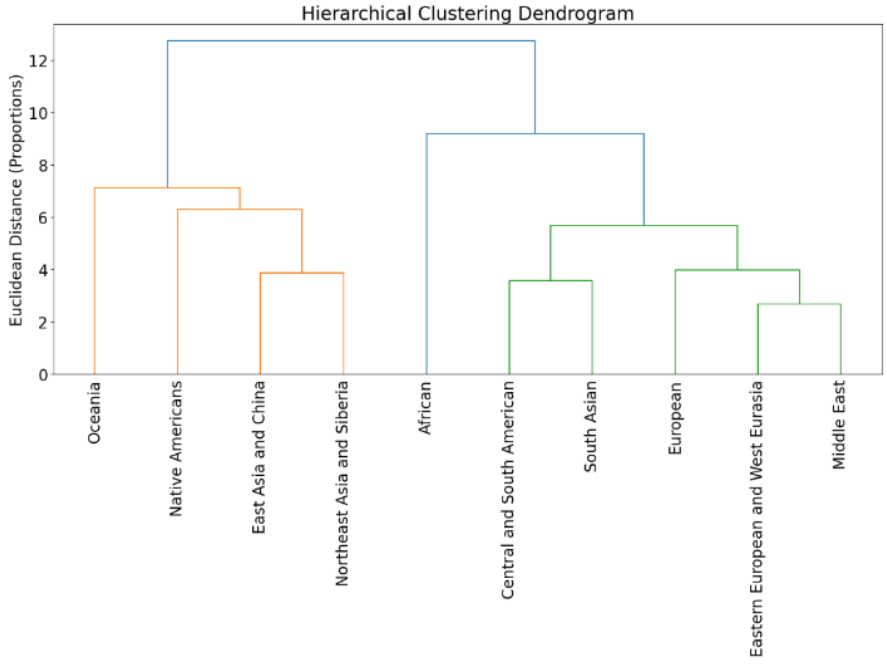
Then based on Formula A above, the square root of the sum of all distances was taken to generate the distance between the two population subgroups:

$$\delta_{Afr,centamerica} = \sqrt{0.680} = 0.825$$

This calculation method was applied to all SNPs via Python, using SciPy's *pdist* [43].

Using the calculated distance matrix, a dendrogram was generated to visually represent the relationships between subgroups based on genetic similarity (Figure G).

Figure G. *Dendrogram of the hierarchical relationship between population classes based on the Euclidean distance between them.*



Model Training and Evaluation: Alleles were encoded as described in Section A, with encoding based on the alternative allele. Weighted sampling was applied during model training to address the class imbalances (Table 7).

Table 7. *Ancestry Model - Original class distribution and Adjusted Class Weights*

Subgroup	Original class distribution	Adjusted class weights
Africa	1139	0.34
Central and South America	277	1.38
East Asia and China	800	0.48
Eastern Europe and West Eurasia	19	20.14
Europe	688	0.56
Middle East	163	2.35
Native Americans	85	4.50
Northeast Asia and Siberia	43	8.90
Oceania	30	12.76
South Asia	583	0.66

Five machine learning models were trained and evaluated, Logistic Regression, Random Forest, Gradient Boosting, Neural Network (MLP) and Decision Tree Classifier.

Model evaluation was done using 10-fold stratified cross-validation. The models were assessed based on F1 scores and accuracy metrics.

Chapter 5: Analysis and Discussion

The overall goal of this study was to measure the feasibility of developing a private, secure, and user-friendly computational framework for genetic analysis. This research concentrated on predicting observable traits, including ancestry, hair color, eye color, blood type, and biological gender, based on genetic data by utilizing genome-wide association studies (GWAS) and single nucleotide polymorphisms (SNPs). The results of this research demonstrated the potential for accurately linking genetic variations to phenotypic traits. In this chapter, I analyze the results of the predictive model developed for each trait, examine their implications, and discuss the context and challenges of genetic analysis.

I. Biological gender:

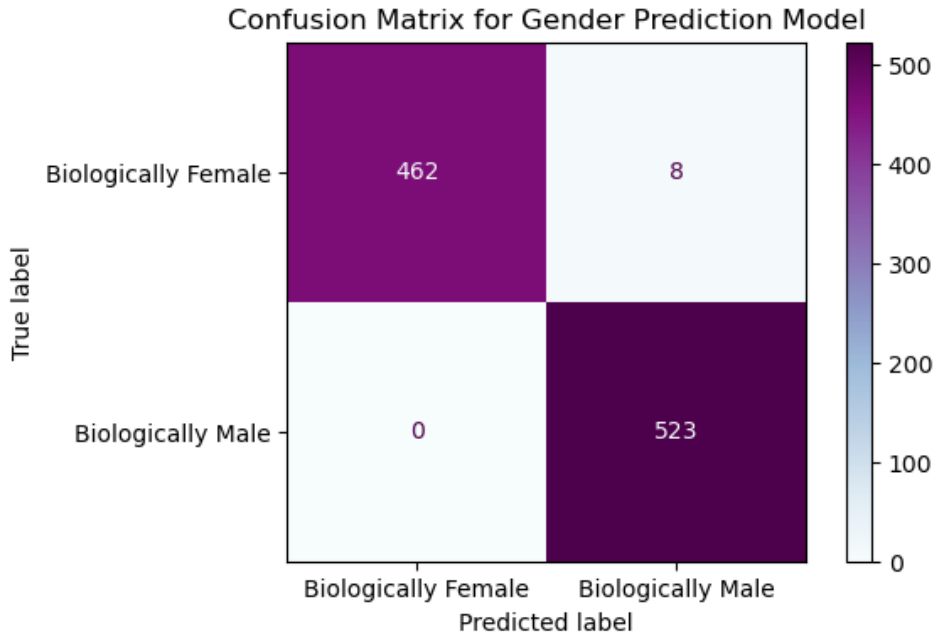
Biological gender prediction relied on identifying the presence or absence of SNPs located on the Y chromosome, particularly within the SRY (Sex-Determining Region Y) gene. The SRY gene is a reliable genetic marker for distinguishing biological males from biological females. This straightforward approach uses the genetic basis of gender determination, as the presence of the SRY gene is enough to classify an individual as male biologically.

The model's performance was evaluated, yielding an impressive F1 score and accuracy of 0.99, as shown in the classification report below.

Table 8. *Classification report for the biological gender prediction model.*

	Precision	**Recall**	**F1-Score**	**Support**
Female Biologically	1.00	0.98	0.99	470
Male Biologically	0.98	1.00	0.99	523
Macro avg	0.99	0.99	0.99	993
Weighted avg	0.99	0.99	0.99	993
Accuracy	**0.99**			**993**

Figure H. *Confusion Matrix for Gender Prediction Model.*



Upon reviewing the eight misclassified results, the following observations were made:

1. Three samples had discrepancies between their predicted gender and their reported biological gender, suggesting that the gender may have been incorrectly reported.
2. Five samples were reported as female but were misclassified as male due to the presence of a single allele associated with the Y chromosome. This anomaly might require further investigation to understand why these alleles are present in individuals classified as female.

Despite these minor discrepancies, the model demonstrated exceptional predictive power, with a 99% accuracy. Its simplicity, relying on just five SNPs from the SRY region, shows the power of using well-defined genetic markers for phenotype predictions. Future research could focus on exploring the causes of misclassifications and include markers that can predict disorders of sex development.

II. Blood Type Prediction

The prediction of blood types was approached using a machine learning model based on 15 SNPs identified as markers for ABO blood types. Blood type is determined by a single gene with a classic dominant-recessive inheritance pattern, where the O blood type is typically recessive. Among the selected SNPs, rs8176719 and rs505922 were set as required for prediction, as they play an important role in distinguishing O types from non-O types [44,45]. These SNPs specifically target the insertion/deletion polymorphism (rs8176719) in the ABO gene, which is key to determining whether an individual carries the O allele.

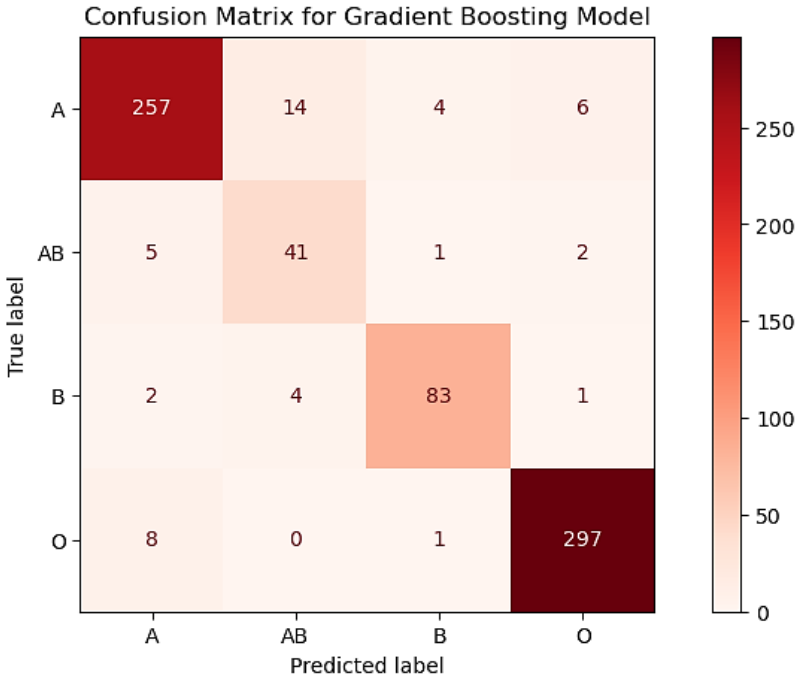
The dataset used for this study replicated real-world blood type distributions, indicating that the samples were representative of the broader population. For instance, O blood type accounted for 42% of the dataset, closely matching its global prevalence of approximately 45%. This representativeness ensures the model's applicability to diverse populations. However, the lack of data for Rh group prediction (+/-) limited the model's scope to ABO types only. Incorporating Rh-related SNPs in future datasets could expand the model's predictive capabilities to include Rh prediction.

All models evaluated during the study demonstrated high predictive capabilities, with F1 scores and accuracy metrics exceeding 94% on a 10-fold stratified cross-validation. Gradient Boosting was selected as the best-performing model due to its slightly better performance compared to other models. The Gradient Boosting model's classification report (Table 4) highlights its strong performance across all blood types.

Table 9. *Gradient Boosting classification report for ABO blood type prediction.*

Blood Type	Precision	Recall	F1-Score	Support
A	0.95	0.95	0.96	281
AB	0.93	0.84	0.88	49
B	0.94	0.97	0.95	90
O	0.97	0.98	0.97	306
Macro avg	0.95	0.93	0.94	726
Weighted avg	0.96	0.96	0.96	726
Accuracy	**0.96**			**726**

Figure I. *Confusion matrix for Gradient Boosting model in blood type prediction.*



The highest precision and recall scores were observed for the O blood type. This is largely due to its straightforward genetic determination by the rs8176719 deletion/insertion polymorphism, which provides a clear distinction between O and non-O blood types. In contrast, the AB blood type showed the lowest precision and recall. This was somewhat expected, as AB is a combination of the A and B alleles, introducing additional genetic complexity that makes

accurate classification more challenging. As we can see from the confusion matrix, the model showed some discrepancies between A and AB and B and AB types. Despite this complexity, the model's performance for AB still performed relatively well, with an F1 score of 0.87. These results suggest that while distinguishing AB from other types may be more complex, the model still captures the genetic interactions that are needed for prediction.

While the current model is effective for predicting ABO blood types, its scope could be expanded to include Rh group prediction. Integrating RhD-related SNPs into the dataset would allow for a more comprehensive blood typing model, which could be invaluable for clinical and forensic applications. Additionally, further refinements could be made to improve the classification of the AB blood type, perhaps by incorporating additional SNPs associated with glycosyltransferase activity in the ABO gene.

III. Eye color prediction

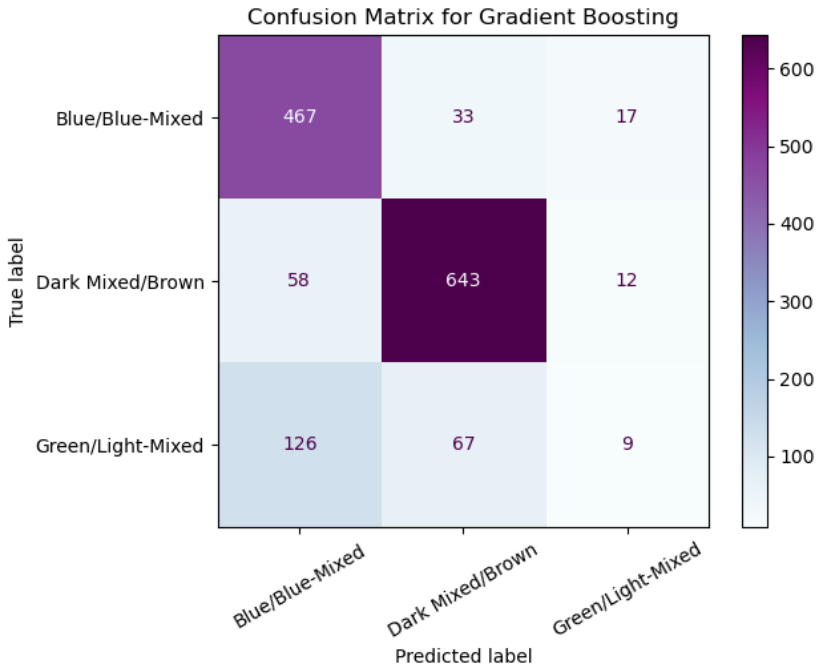
Eye color prediction proved to be one of the most complex traits to model due to the highly polygenic nature of this trait. Eye color is influenced by many genes, many of which interact with each other in complex ways to produce the spectrum of colors we see in eyes. This makes lighter and intermediate eye colors, such as green, particularly challenging to predict accurately. To account for this, three approaches were implemented: (1) prediction using a full dataset containing three classes (Blue/Blue-Mixed, Green/Light-Mixed, and Dark Mixed/Brown), (2) a subset comparing Blue/Blue-Mixed vs. Green/Light-Mixed, and (3) a subset comparing Blue/Blue-Mixed vs. Dark Mixed/Brown.

The Gradient Boosting model was selected for the three-class dataset due to its overall performance across the classes. The classification report (Table 5) illustrates the model's ability to distinguish between the three-color groups.

Table 10. *Gradient Boosting classification report for three-class eye color predictions.*

Eye Color	Precision	Recall	F1-Score	Support
Blue/Blue-Mixed	0.72	0.90	0.80	517
Dark Mixed/Brown	0.87	0.90	0.88	713
Green/Light-Mixed	0.26	0.05	0.08	202
Macro avg	0.61	0.62	0.59	1432
Weighted avg	0.73	0.78	0.74	1432
Accuracy	**0.78**			**1432**

Figure J. *Confusion matrix for Gradient Boosting model in eye color prediction.*



The model performed well for Dark Mixed/Brown eye colors, with an F1 score of 0.88, and relatively well for Blue/Blue-Mixed, with an F1 score of 0.80. However, it struggled to predict Green/Light-Mixed eyes, with an F1 score of only 0.07. This poor performance highlights the difficulty in distinguishing intermediate colors such as green and hazel, which are influenced by a complex combination of genes. While other models, such as MLR (F1 = 0.31) and Decision Tree (F1 = 0.24), performed slightly better for the Green/Light-Mixed group,

Gradient Boosting was ultimately selected due to its better performance for the other two groups. Nevertheless, the overall limitations of the three-class model are substantial, and its application should be considered carefully.

To refine the prediction, two subset models were created: (1) Blue/Blue-Mixed vs. Green/Light-Mixed (Bu/Gr model), and (2) Blue/Blue-Mixed vs Dark Mixed/Brown (Bu/Br model). This modeling process was designed to predict traits sequentially, intending to improve the model's ability to distinguish between colors.

The Gradient Boosting model performed relatively well for blue and brown eyes on both subsets (Tables 6-7). However, in the Bu/Gr model, the model still struggled with distinguishing the Green/Light-Mixed group from Blue/Blue-Mixed, with an F1 score of 0.44. This can be due to the very nuanced genetics related to intermediate colors such as green and hazel, which are often a combination between lighter color genes and darker ones that create those colors. As a result, the prediction of Blue/Blue-Mixed also had an F1 score that is lower (0.85) than that of the Bu/Br model (0.91). The Bu/Br model achieved high performance, with F1 scores of 0.91 for Blue/Blue-Mixed and 0.93 for Dark Mixed/Brown.

Table 11. *Gradient Boosting classification report for Blue/Blue-Mixed vs. Green/Light-Mixed eye color predictions.*

Eye Color	Precision	Recall	F1-Score	Support
Blue/Blue-Mixed	0.78	0.93	0.85	517
Green/Light-Mixed	0.64	0.33	0.44	202
Macro avg	0.71	0.63	0.64	719
Weighted avg	0.74	0.76	0.73	719
Accuracy	**0.76**			**719**

Figure K. *Confusion matrix for Gradient Boosting model in Blue vs. Green eye color prediction.*

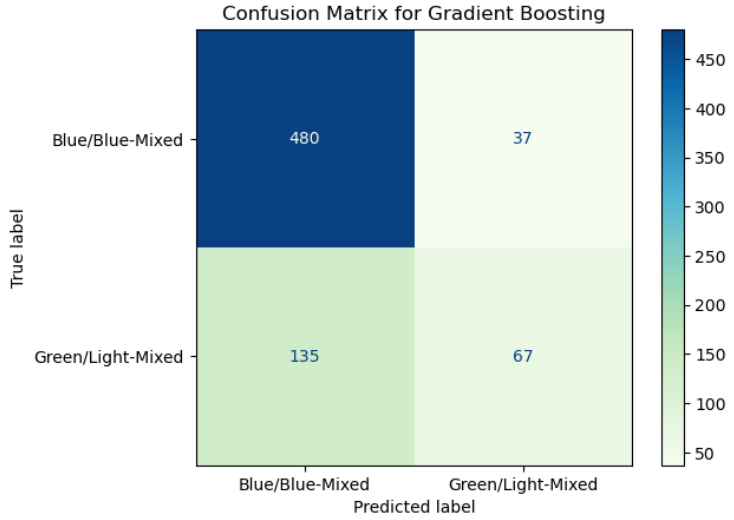
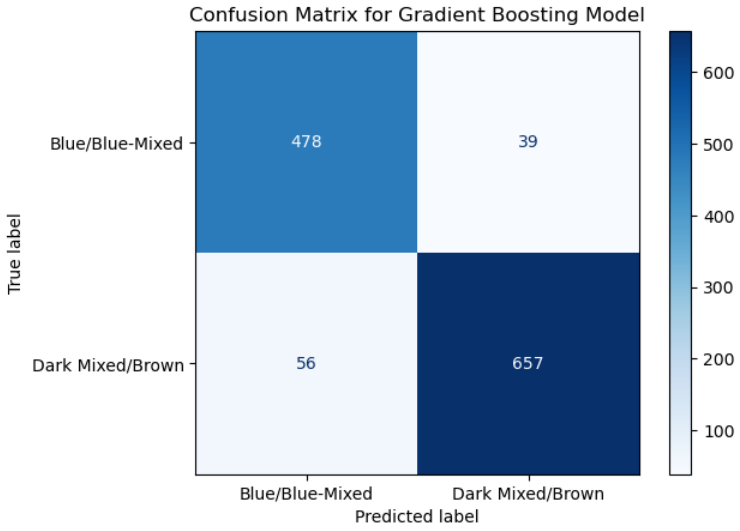


Table 12. *Gradient Boosting classification report for Blue/Blue-Mixed vs. Dark-Mixed/Brown eye color predictions.*

Eye Color	Precision	Recall	F1-Score	Support
Blue/Blue-Mixed	0.90	0.92	0.91	517
Dark Mixed/Brown	0.94	0.92	0.93	713
Macro avg	0.92	0.92	0.92	1230
Weighted avg	0.92	0.92	0.92	1230
Accuracy	**0.92**			**1230**

Figure L. *Confusion matrix for Gradient Boosting model – Blue vs. Brown eye color prediction.*



The absence of intermediate colors in this subset allowed the model to distinguish the genetic differences between the two groups more effectively. This highlights the importance of simplifying classification problems when dealing with traits influenced by complex genetic interactions.

One significant limitation of this study was the inconsistency in user-reported eye colors within the dataset. Since users were allowed to enter any value, the data included diverse and sometimes ambiguous descriptors, such as "black," which is not a naturally occurring eye color in humans [46]. Additionally, the lack of empirical measurement for eye color (using a validated scale or spectrophotometer) may have introduced biases based on the user's perception. Future studies could address this limitation by collecting more controlled and validated datasets, ensuring consistent reporting of eye colors.

Despite these challenges, the Gradient Boosting model demonstrated strong predictive performance for the Dark vs. Light subset, which shows the potential for machine learning to model complex traits such as eye color. However, it also highlights the importance of consistent data collection and preprocessing techniques to reduce biases and resulting errors.

IV. Hair color prediction:

Hair color prediction was divided into two distinct models: (1) red vs. non-red hair and (2) light hair (blonde) vs. dark hair (brown/black). This division was necessary due to the distinct genetics and biochemical processes underlying these traits. Red hair is determined by the presence of pheomelanin, a red/yellow pigment, while the lightness or darkness of hair depends on the level of eumelanin, a brown/black pigment. Given that different SNPs are involved in regulating these pigments, separate models were necessary for accurate predictions.

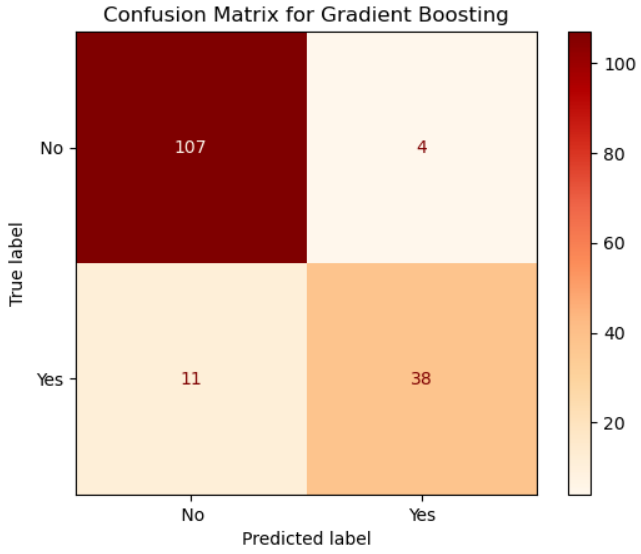
Red vs. Non-Red Hair Color Prediction

For red vs. non-red hair, Gradient Boosting was selected as the best-performing model, achieving high accuracy and balanced performance across both categories.

Table 13. *Gradient Boosting classification report for red vs. non-red hair color prediction.*

Hair Color	Precision	Recall	F1-Score	Support
No (non-red)	0.91	0.96	0.93	111
Yes (Red)	0.90	0.78	0.84	49
Macro avg	0.91	0.87	0.89	160
Weighted avg	0.91	0.90	0.92	160
Accuracy	**0.91**			**160**

Figure M. *Confusion matrix for Gradient Boosting model – red vs. non-red hair color prediction.*



The model achieved strong performance in predicting non-red hair, with an F1 score of 0.93. For red hair, the F1 score was slightly lower at 0.84. While weighted sampling was applied to address class imbalances, the rarity of red hair and limited genotypes in the sample still caused challenges for the model. Still, the model demonstrated strong overall accuracy at 91%, showing good capability to reliably distinguish between red and non-red hair colors.

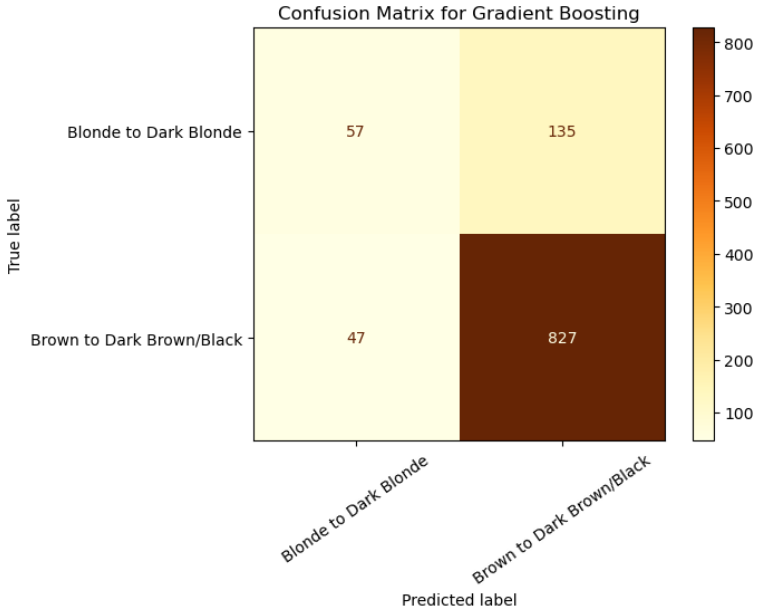
Light vs. Dark Hair Color Prediction

The second model focused on distinguishing between light hair (blonde to dark blonde) and dark hair (brown to dark brown/black). Gradient Boosting was the selected model, as it showed high predictive power for darker colors. For lighter hair colors, performance across all models was significantly lower, with an F1 score of 0.39 for the Gradient Boosting model. While MLR slightly outperformed Gradient Boosting for the lighter class (F1 = 0.52 vs. 0.39), Gradient Boosting was chosen due to its superior performance for darker hair, which comprised most of the dataset.

Table 14. *Gradient Boosting classification report for light vs. dark hair color prediction.*

Hair Color	Precision	Recall	F1-Score	Support
Blonde to Dark Blonde	0.55	0.31	0.39	192
Brown to Dark Brown/Black	0.86	0.95	0.90	874
Macro avg	0.71	0.63	0.65	1066
Weighted avg	0.80	0.83	0.81	1066
Accuracy	**0.83**			**1066**

Figure N. *Confusion matrix for Gradient Boosting model - light vs. dark hair color prediction.*



Similarly to eye-color, the most significant limitations in hair color prediction was the variability in user-reported data. Even more than eye colors, reported hair colors in the dataset were highly diverse, with descriptions ranging from "Grey and brown" to "Strawberry blonde as a child, now dark auburn brown." This variability made it challenging to group the data into consistent and meaningful categories. Furthermore, it is very likely that subjective reporting and differences in individual perception introduced errors into the dataset.

Another limitation was the overlap between certain hair colors, such as light brown and dark blonde. These colors are often difficult to differentiate without precise measurement tools, leading to potential misclassification. To address this issue in future studies, hair pigmentation should be measured using a standardized scale, such as spectrophotometric color analysis, which provides objective and quantifiable data on hair pigmentation.

Despite these challenges, the Gradient Boosting model demonstrated good performance for darker hair colors and red hair, showing high potential for predicting certain phenotypes. However, further refinement of the dataset and color measurement techniques is needed to improve prediction accuracy for lighter and intermediate hair colors.

V. Ancestry prediction

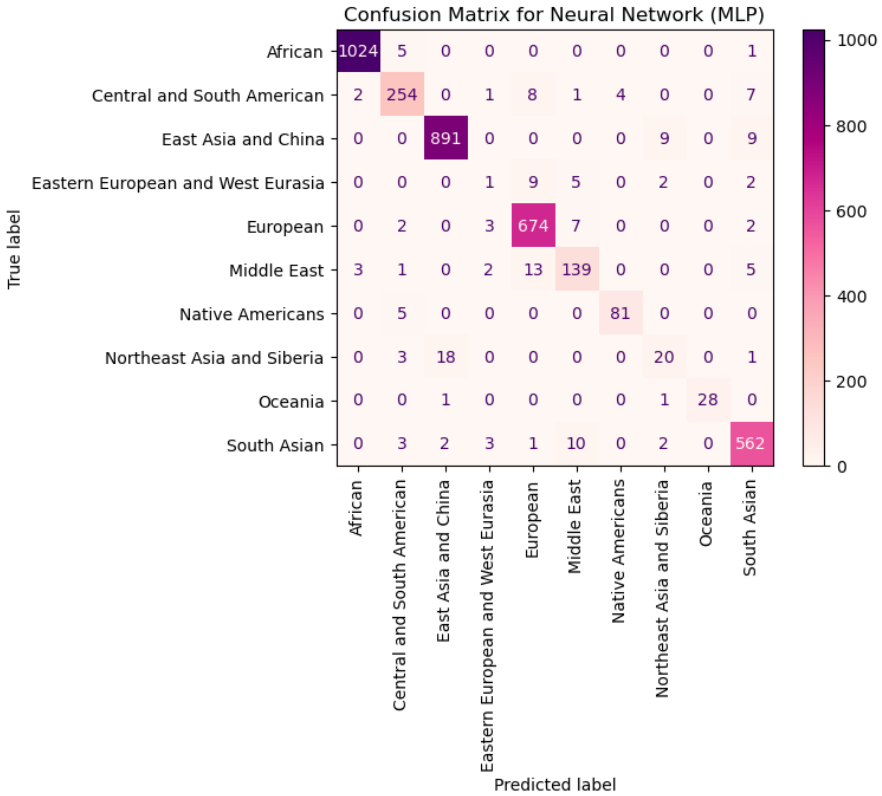
Ancestry prediction relies on genetic markers capable of distinguishing population groups based on their unique allele frequency distributions. For this model, the SNPs were selected based on their ability to maximize genetic distance between populations.

For ancestry prediction, the Neural Network (MLP) model was the top-performing model, with an accuracy of 96% and an F1 score of 0.96 across 10-fold stratified cross-validation.

Table 15. *Classification report for ancestry prediction using the Neural Network (MLP) model.*

	precision	recall	f1-score	support
African	0.99	1.00	1.00	1030
Central and South American	0.92	0.90	0.91	277
East Asia and China	0.97	0.99	0.98	909
Eastern European and West Eurasia	0.29	0.11	0.15	19
European	0.95	0.98	0.97	688
Middle East	0.84	0.85	0.84	163
Native Americans	0.96	0.92	0.94	86
Northeast Asia and Siberia	0.96	0.55	0.70	42
Oceania	0.97	0.97	0.97	30
South Asian	0.95	0.96	0.96	583
Macro avg	0.849	0.810	0.824	3827
Weighted avg	0.956	0.958	0.956	3827
Accuracy	**0.96**			**3827**

Figure O. *Confusion matrix for MLP Model in ancestry prediction.*



The classification report (Table 10) reveals consistently high precision and recall for most subgroups, particularly for African (F1 = 1.00), East Asia and China (F1 = 0.98), European (F1 = 0.97), and South Asian (F1 = 0.96) populations. However, the model struggled with smaller and less genetically distinct groups, such as Eastern European and West Eurasian (F1 = 0.15) and Northeast Asia and Siberia (F1 = 0.70). This low prediction capabilities may be due to the small sample sizes and genetic overlap with neighboring populations. While weighted sampling was used to reduce the model's biases towards overrepresented populations, the number of samples remains low and might not reflect genetic diversity well. In addition, the similarity between population as a result of migration patterns and common ancestors might force these groups to be merged into other population groups based on their genetic proximity.

The dendrogram (Figure C) shows that clusters align with geographical and genetic proximities. For example, African populations were distinctly clustered, reflecting their significant genetic distance from non-African populations, consistent with human migration patterns. Smaller subgroups, such as Northeast Asia and Siberia, appeared closely linked to East Asia and China, indicating genetic similarities despite their geographic locations.

It is worth investigating whether the current subgroup classifications can be refined. One possible approach is to classify populations into subgroups based on migration patterns rather than genetic distances. Additionally, merging subgroups with very small sample sizes or incorporating additional SNPs with greater distances from other populations might improve model performance for these groups.

While the model achieved excellent overall performance, the SNP selection process could be further refined to include markers with higher discriminatory power and reassess current markers that do not contribute to the model. A deeper review of allele frequency distributions and their potential for distinguishing specific populations from each other could help resolve some of the current limitations.

Chapter 6: Conclusions

The field of genomics has seen tremendous growth in the past decade, but many existing genetic analysis platforms remain limited in their scope and usability. Current tools, such as 23andMe and Promethease fall short in addressing critical issues such as data privacy and security, ease of use, and broader applicability beyond consumer contexts. This research aimed to address these issues by creating a framework that would enable a locally run genetic analysis that can be modified and customized to meet the unique requirements of different sectors. While the initial focus was on traits such as biological gender, blood type, and eye and hair color, the framework is flexible and can be adapted to incorporate various traits, making it applicable to different contexts, such as healthcare, pharmacogenomics, and nutrition.

Throughout this study, several challenges emerged, indicating areas for future improvements and research. One of the primary challenges was the inherent complexity of predicting less common phenotypes, such as green eye color or red hair. Such phenotypes are often underrepresented in the dataset. While this can be corrected using methods such as weighted sampling or oversampling, there is still an issue with diversity within those groups. As a result, while they share a trait, the small sample size might not be representative of the general population.

Another challenge that arose was the lack of sufficient data from certain population groups in the ancestry dataset. This issue not only restricts the inclusivity of the framework but also raises ethical concerns. as populations with limited representation in genetic research can have disadvantages in healthcare, genetic disease discovery, and treatment options. This highlights the importance of improving the representation of diverse populations in genetic

datasets, not only for predictive purposes, but also to ensure that genetic diversity is preserved across all research domains.

Another notable limitation was the inconsistency in user-reported traits, such as eye and hair color, which varied widely in description and were subjected to personal biases. Standardizing trait reporting and employing objective measurement techniques, such as spectrophotometric analysis for color predictions, could help minimize these issues and enhance the reliability of genetic predictions.

Yet, despite these limitations, the framework allowed for the generation of models with strong predictive capabilities across most categories, supporting its feasibility as a private, secure, and broadly applicable genetic analysis platform. Additionally, another finding that emerged during the research was the strong performance of the Gradient Boosting model for predictive genetics. While some additional research might be required, Gradient Boosting consistently performed well across most trait predictions, including more complex traits like eye and hair color, indicating that it can potentially be used as a "default" model for predicting phenotypes from genotypes.

Importantly, the framework proposed in this study is adaptable for various industries, from healthcare to personalized nutrition and ancestry research, by refining the data collected and expanding the scope of traits analyzed. The findings of this research contribute to the broader body of knowledge by illustrating a clear pathway for developing ethical, private, and practical genetic analysis platforms; it introduces methods for predictive genetic analysis, which demonstrate high capability and applicability; and it shows the high predictive power of models such as Gradient Boosting for simple and complex phenotype predictions.

In conclusion, while additional research is still necessary, this study successfully demonstrated the practicality of a novel framework for a genetic analysis tool that emphasizes privacy, security, and usability, setting it apart from existing platforms. Furthermore, the framework's adaptability enables it to be customized or expanded to address more specific or complex trait analyses, making it a valuable foundation for future research. By addressing current limitations and continuing to refine the methodology, this framework has the potential to make genetic analysis a more inclusive, ethical, and impactful domain. Future work could explore incorporating PCA for dimensionality reduction, which may enhance the framework by improving computational efficiency and focusing on the most significant genetic features.

References

(1) *15 Ways Genomics Influences Our World*. https://www.genome.gov/dna-day/15-ways (accessed 2024-11-27).

(2) Allyse, M. A.; Robinson, D. H.; Ferber, M. J.; Sharp, R. R. Direct-to-Consumer Testing 2.0: Emerging Models of Direct-to-Consumer Genetic Testing. *Mayo Clin. Proc.* **2018**, *93* (1), 113–120. https://doi.org/10.1016/j.mayocp.2017.11.001.

(3) 23andMe. *DNA Genetic Testing For Health, Ancestry And More - 23andMe*. https://www.23andme.com/ (accessed 2024-12-14).

(4) *Ancestry® | Family Tree, Genealogy & Family History Records*. https://www.ancestry.com/?o_xid=115784&o_lid=115784&o_sch=Paid+Search+Brand&ancid=4vmopenrg0&ds_rl=1286410&pgrid=120522386862&ptaid=kwd-29052520&s_kwcid=ancestry&gad_source=1&gclid=Cj0KCQiA0--6BhCBARIsADYqyL8sLbqObHkRsS9jfDJv0UkBnlWmWdPoK8HjRN2zywKStvKVEDINIMIaApx1EALw_wcB&gclsrc=aw.ds (accessed 2024-12-14).

(5) *Free Family Tree, Genealogy, Family History, and DNA Testing*. MyHeritage. https://www.myheritage.com (accessed 2024-12-14).

(6) *Promethease*. https://promethease.com/ (accessed 2024-12-14).

(7) Watson, B. *Promethease and Xcode Life health reports: How to choose 23andme raw data health reports [2023 Update]*. Xcode Life. https://www.xcode.life/23andme-raw-data/promethease-xcode-life-23andme-raw-data-analysis-health-reports/ (accessed 2024-11-27).

(8) Nelson, S. C.; Bowen, D. J.; Fullerton, S. M. Third-Party Genetic Interpretation Tools: A Mixed-Methods Study of Consumer Motivation and Behavior. *Am. J. Hum. Genet.* **2019**, *105* (1), 122–131. https://doi.org/10.1016/j.ajhg.2019.05.014.

(9) *Promethease Review: Everything You Need To Know - LifeDNA*. https://lifedna.com/dna-blog/promethease-review-everything-you-need-to-know/ (accessed 2024-11-19).

(10) DeGeurin, M. Hackers Got Nearly 7 Million People's Data from 23andMe. The Firm Blamed Users in 'Very Dumb' Move. *The Guardian*. February 15, 2024. https://www.theguardian.com/technology/2024/feb/15/23andme-hack-data-genetic-data-selling-response (accessed 2024-11-24).

(11) *MyHeritage Acquires Promethease and SNPedia*. https://www.businesswire.com/news/home/20190907005012/en/MyHeritage-Acquires-Promethease-and-SNPedia (accessed 2024-11-24).

(12) Skwarecki, B. *If You Ever Used Promethease, Your DNA Data Might Be on MyHeritage Now*. Lifehacker. https://lifehacker.com/if-you-ever-used-promethease-your-dna-data-might-be-on-1841327595 (accessed 2024-11-24).

(13) Allyn, B. 23andMe Is on the Brink. What Happens to All Its DNA Data? *NPR*. October 3, 2024. https://www.npr.org/2024/10/03/g-s1-25795/23andme-data-genetic-dna-privacy (accessed 2024-11-24).

(14) Enright, M. *Inside the fall of 23andMe*. CNBC. https://www.cnbc.com/2024/10/23/inside-the-fall-of-23andme.html (accessed 2024-11-24).

(15) *GWAS Catalog*. https://www.ebi.ac.uk/gwas/efotraits/EFO_0003924 (accessed 2024-11-09).

(16) *rs9463733 RefSNP Report - dbSNP - NCBI*. https://www.ncbi.nlm.nih.gov/snp/rs9463733 (accessed 2024-11-14).

(17) *1000 Genomes | A Deep Catalog of Human Genetic Variation*. https://www.internationalgenome.org/home (accessed 2024-11-27).

(18) 23andMe. *23andMe - Genetics 101: What are SNPs?* https://www.23andme.com/gen101/snps/ (accessed 2024-11-27).

(19) Rahim, N. G.; Harismendy, O.; Topol, E. J.; Frazer, K. A. Genetic Determinants of Phenotypic Diversity in Humans. *Genome Biol.* **2008**, *9* (4), 215. https://doi.org/10.1186/gb-2008-9-4-215.

(20) Jajosky, R. P.; Wu, S.-C.; Zheng, L.; Jajosky, A. N.; Jajosky, P. G.; Josephson, C. D.; Hollenhorst, M. A.; Sackstein, R.; Cummings, R. D.; Arthur, C. M.; Stowell, S. R. ABO Blood Group Antigens and Differential Glycan Expression: Perspective on the Evolution of Common Human Enzyme Deficiencies. *iScience* **2022**, *26* (1), 105798. https://doi.org/10.1016/j.isci.2022.105798.

(21) Groot, H. E.; Villegas Sierra, L. E.; Said, M. A.; Lipsic, E.; Karper, J. C.; van der Harst, P. Genetically Determined ABO Blood Group and Its Associations With Health and Disease. *Arterioscler. Thromb. Vasc. Biol.* **2020**, *40* (3), 830–838. https://doi.org/10.1161/ATVBAHA.119.313658.

(22) Flegel, W. A. The Genetics of the Rhesus Blood Group System. *Blood Transfus.* **2007**, *5* (2), 50–57. https://doi.org/10.2450/2007.0011-07.

(23) *Is eye color determined by genetics?: MedlinePlus Genetics*. https://medlineplus.gov/genetics/understanding/traits/eyecolor/ (accessed 2024-11-27).

(24) Branicki, W.; Brudnik, U.; Wojas-Pelc, A. Interactions between HERC2, OCA2 and MC1R May Influence Human Pigmentation Phenotype. *Ann. Hum. Genet.* **2009**, *73* (2), 160–170. https://doi.org/10.1111/j.1469-1809.2009.00504.x.

(25) Meyer, O. S.; Lunn, M. M. B.; Garcia, S. L.; Kjærbye, A. B.; Morling, N.; Børsting, C.; Andersen, J. D. Association between Brown Eye Colour in Rs12913832:GG Individuals and SNPs in TYR,

TYRP1, and SLC24A4. *PLOS ONE* **2020**, *15* (9), e0239131. https://doi.org/10.1371/journal.pone.0239131.

(26) *Population genetics - Latest research and news | Nature*. https://www.nature.com/subjects/population-genetics (accessed 2024-11-28).

(27) *Simons Genome Diversity Project*. Simons Foundation. https://www.simonsfoundation.org/simons-genome-diversity-project/ (accessed 2024-11-28).

(28) *CRB du CEPH*. https://www.fjd-ceph.org/crb-du-ceph (accessed 2024-11-28).

(29) *ALFA: Allele Frequency Aggregator*. https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/ (accessed 2024-11-28).

(30) Quinton, A. R.; Kelty, S. F.; Scudder, N. Attitudes towards Police Use of Consumer/Private DNA Databases in Investigations. *Sci. Justice* **2022**, *62* (3), 263–271. https://doi.org/10.1016/j.scijus.2022.02.009.

(31) Russell, K.; Kelty, S. F.; Scudder, N. Public and Family Support and Concerns for Providing DNA to Law Enforcement in Long-Term Missing Person Cases. *Sci. Justice* **2023**, *63* (2), 149–157. https://doi.org/10.1016/j.scijus.2022.12.004.

(32) *How 23andMe Responds to Law Enforcement Requests for Customer Information*. 23andMe Customer Care. https://customercare.23andme.com/hc/en-us/articles/212271048-How-23andMe-Responds-to-Law-Enforcement-Requests-for-Customer-Information (accessed 2024-11-27).

(33) Whittaker, Z. *Ancestry says it fought two police requests to search its DNA database*. TechCrunch. https://techcrunch.com/2021/02/10/ancestry-police-warrant-dna-database/ (accessed 2024-11-30).

(34) *Home - SNP - NCBI*. https://www.ncbi.nlm.nih.gov/snp/ (accessed 2024-11-30).

(35) *openSNP*. https://opensnp.org/ (accessed 2024-11-30).

(36) *compute_class_weight*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html (accessed 2024-12-16).

(37) *World Population By Percentage of Blood Types*. WorldAtlas. https://www.worldatlas.com/articles/what-are-the-different-blood-types.html (accessed 2024-12-01).

(38) *Martin–Schultz scale*. Academic Dictionaries and Encyclopedias. https://en-academic.com/dic.nsf/enwiki/11703157 (accessed 2024-12-01).

(39) *Eye colors: Most common and percentages*. https://www.medicalnewstoday.com/articles/eye-color-percentage (accessed 2024-12-02).

(40) *Is hair color determined by genetics?: MedlinePlus Genetics*. https://medlineplus.gov/genetics/understanding/traits/haircolor/ (accessed 2024-12-03).

(41) Wang, C.; Zöllner, S.; Rosenberg, N. A. A Quantitative Comparison of the Similarity between Genes and Geography in Worldwide Human Populations. *PLOS Genet.* **2012**, *8* (8), e1002886. https://doi.org/10.1371/journal.pgen.1002886.

(42) Cox, M. A. A.; Cox, T. F. Multidimensional Scaling. In *Handbook of Data Visualization*; Chen, C., Härdle, W., Unwin, A., Eds.; Springer: Berlin, Heidelberg, 2008; pp 315–347. https://doi.org/10.1007/978-3-540-33037-0_14.

(43) *pdist — SciPy v1.14.1 Manual*. https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.pdist.html (accessed 2024-12-16).

(44) Li-Gao, R.; Carlotti, F.; de Mutsert, R.; van Hylckama Vlieg, A.; de Koning, E. J. P.; Jukema, J. W.; Rosendaal, F. R.; Willems van Dijk, K.; Mook-Kanamori, D. O. Genome-Wide Association Study on the Early-Phase Insulin Response to a Liquid Mixed Meal: Results From the NEO Study. *Diabetes* **2019**, *68* (12), 2327–2336. https://doi.org/10.2337/db19-0378.

(45) Paterson, A. D.; Lopes-Virella, M. F.; Waggott, D.; Boright, A. P.; Hosseini, S. M.; Carter, R. E.; Shen, E.; Mirea, L.; Bharaj, B.; Sun, L.; Bull, S. B.; the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group. Genome-Wide Association Identifies the ABO Blood Group as a Major Locus Associated With Serum Levels of Soluble E-Selectin. *Arterioscler. Thromb. Vasc. Biol.* **2009**, *29* (11), 1958–1967. https://doi.org/10.1161/ATVBAHA.109.192971.

(46) *What Is the Rarest Eye Color?*. Verywell Health. https://www.verywellhealth.com/what-is-the-rarest-eye-color-5087302 (accessed 2024-12-04).

(47) *NCBI Variation Services*. https://api.ncbi.nlm.nih.gov/variation/v0/ (accessed 2024-12-10).

(48) *Index of /genetics/reich_lab/sgdp/vcf_variants*. https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/vcf_variants/ (accessed 2024-12-10).

(49) *SPSmart*. http://spsmart.cesga.es/search.php?dataSet=ceph_stanford (accessed 2024-12-10).

(50) *BigQuery – Google Cloud console*. https://console.cloud.google.com/bigquery?p=bigquery-public-data&page=table&t=1000_genomes_phase_3_variants_20150220&d=human_genome_variants&pli=1&inv=1&invt=AbjxDg&project=applied-dialect-422121-g5&ws=!1m5!1m4!4m3!1sbigquery-public-data!2shuman_genome_variants!3s1000_genomes_phase_3_variants_20150220 (accessed 2024-12-10).

Appendices

Appendix A: Data Resources

Table 16. *Data Sources for Genetic Analysis.*

Dataset	Source
ALFA Frequency and dbSNP	data was obtained via python through NCBI API, using RefSNP services: https://api.ncbi.nlm.nih.gov/variation/v0/ [29,47]
SGDP	https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/vcf_variants/ [48]
HGDP	http://spsmart.cesga.es/search.php?dataSet=ceph_stanford [49]
1000 Genomes Project	https://console.cloud.google.com/bigquery?p=bigquery-public-data&page=table&t=1000_genomes_phase_3_variants_20150220&d=human_genome_variants [50] Or https://www.internationalgenome.org/data/ [17]
OpenSNP	https://opensnp.org/ [35]

Appendix B: List of Model SNPs

The following tables include terms used in this paper. Their meanings are clarified below to aid in interpretation:

- **effect_allele**: Denotes the “effect allele”, which is the allele variation associated with the reference phenotype and serves as the basis for encoding.
- **other_allele**: Denotes the “other allele”, which is the allele variation that is either negatively associated with, or not associated with the reference phenotype,

Table 17. *Biological Gender prediction SNPs*

rsid	effect_allele	other_allele	ref_phenotype
rs757619452	G	A	Male
rs569336697	T	C	Male
rs11575897	G	A	Male
rs2534636	C	T	Male

Table 18. *Blood Type Prediction SNPs*

rsid	effect_allele	other_allele	ref_phenotype
rs8176719	D	I	O
rs505922	T	C	O
rs657152	A	C	O
rs8176704	G	A	O
rs612169	G	A	O
rs529565	C	T	O
rs8176693	C	T	O
rs514659	C	A	O
rs8176645	A	T	O
rs512770	G	A	O
rs688976	C	A	O
rs8176672	C	T	A
rs8176741	G	A	A
rs8176722	C	A	A
rs8176720	T	C	A
rs7853989	C	G	A
rs8176743	C	T	A
rs8176747	G	C	A
rs635634	C	T	A
rs507666	A	G	A
rs687289	A	G	A
rs8176746	G	T	A
rs8176749	C	T	A
rs7030248	A	G	AB
rs41302905	T	C	O
rs590787	C	A	RhD (Rh+)
i4001527	C	T	RhD (Rh+)

Table 19. *Eye Color Prediction SNPs*

rsid	associated_color	effect_allele	other_allele
rs12913832	blue	G	A
rs16891982	blue	G	C
rs12203592	blue/green	T	C
rs12896399	blue	T	G
rs6119471	brown (G)	C	G
rs35866166	brown (C)	T	C
rs62538956	brown (C)	T	C
rs1289469	brown (C)	A	C
rs1126809	brown (G)	A	G
rs1426654	brown (G)	A	G
rs1800407	blue/gray	T	C
rs1393350	blue	A	G
rs1408799	blue	C	T
rs1800401	blue	A	G
rs7174027	light	G	A
rs3794604	light	C	T
rs4778241	darker	C	A
rs1129038	blue	T	C
rs1667394	blue	T	C
rs916977	light	C	T
rs11636232	light	T	C
rs7174027	dark	G	A
rs7495174	light	A	G
rs7183877	blue	C	A
rs1800411	light	A	G
rs1498519	light	G	T
rs977588	lighter color	A	C
rs12593929	light	A	G
rs3935591	light	C	T
rs7170852	light	A	T
rs2238289	blue	A	G
rs3940272	blue	G	T
rs8028689	blue	T	C
rs2240203	blue	T	C
rs11631797	blue	G	A
rs35264875	light	T	A

***Table 20**. Red Hair vs. Non-Red Hair Prediction SNPs*

rsid	effect_allele	other_allele	ref_phenotype
rs1805007	T	C	red hair
rs1805008	T	C	red hair
i3002507	C	G	red hair
rs1805006	A	C	red hair
rs11547464	A	G	red hair
rs1805009	C	G	red hair
rs3212379	T	C	red hair
rs34474212	C	T	red hair
rs34158934	T	C	red hair
rs201326893	A	C	red hair
rs555179612	D	I	red hair
rs200000734	T	C	red hair
rs368507952	A	G	red hair

***Table 21.** Light vs. Dark Hair Color Prediction SNPs*

rsid	associated_color	effect_allele	other_allele
rs35264875	blonde	T	A
rs4911414	blonde	T	G
rs12896399	blonde	T	G
rs8033165	blonde	T	C
rs12913832	blonde	G	A
rs12203592	blonde	T	C
rs28777	blonde	A	C
rs6918152	blonde	G	A
rs12821256	blonde	C	T
rs1667394	blonde	T	C
rs12931267	blonde	G	C
rs16891982	blonde	G	C
rs291671	Blonde	G	A
rs1129038	Blonde	T	C
rs8045560	Blonde	T	C
rs1110400	Blonde	C	T
rs1268789	Blonde	C	T
rs1393350	Blonde	A	G

rs885479	blonde	A	G
rs3829241	blonde	A	G
rs1042602	blonde	C	A
rs1800407	blonde	T	C
rs17783630	Blonde	C	A
rs7603664	Blonde	C	T
rs974455	Blonde	A	G
rs7495174	Blonde	A	G
rs916977	Blonde	C	T
rs11636232	Blonde	T	C
rs4778241	Blonde	C	A
rs7174027	Blonde	G	A
rs26722	Light	C	T
rs1015362	Blonde	C	T
rs3794604	Light	C	T
rs9782955	Light	C	T
rs1126809	brown (G)	A	G
rs1426654	Light	A	G
rs1408799	Light	C	T
rs1800401	light	G	A
rs4911442	light	G	A
rs312262906	light	D	I
rs796296176	light	D	I
rs201326893	light	A	C
rs4959270	light	A	C
rs2402130	light	A	G
rs683	light	A	C

Appendix C: Reported Eye and Hair Color Classifications

I. Eye color classifications:

Class 1: Blue/Blue Mixed (Martin-Schultz Scale 1–5)
'blue', 'Blue', 'Dark blue', 'Dark Blue', 'Blue spot of brown', 'blue spot of brown', 'Split - one side dark blue / other side light blue and green', ‘Ice blue mixed with slate blue, with an amber pupil burst in both eyes and a brown spot adjacent to lower left pupil. eyes were green into my 20's.’, 'Light gray/blue.', 'Light Gray Blue', 'Gray-blue', 'blue-grey', 'gray-blue', 'Blue-grey', 'Blue/gray', 'Blue grey', 'Blue/gray', 'Dark Grayish-Blue Eyes (like a stone)', 'Blue-grey with central heterochromia', 'Light gray/blue. amber/med brown on sphincter. gray ring around outer edge. flecks (nevi).', 'Light Gray/Blue. Amber/Med Brown on Sphincter. Gray ring around outer edge. Flecks (Nevi).', 'Light gray/blue. amber/med brown on sphincter. gray ring around outer edge. flecks (nevi).', 'Blue with yellow parts', 'Blue with yellow inner ring', 'blue-brown heterochomia', 'blue-brown heterochromia ', 'blue-brown heterochromia', 'blue-green', 'Blue-green', 'blue-green ', 'Blue/Green', 'Blue/green', 'Blue-green ', 'Light blue-green', 'Blue-grey; broken amber collarette', 'Blue with a yellow ring of flecks that make my eyes look green depending on the light or my mood '.

Class 2: Green/Light-Mixed (Martin-Schultz Scale 6–8)
'Light green', 'Light-mixed Green', 'Light-mixed green', 'green', 'Green', 'Green ', 'Green with blue halo', 'Green-gray', 'Grey and amber', 'Grey and Amber', 'green-blue outer ring and brown flecks around iris', 'Green-blue outer ring and brown flecks around iris', 'Green with brown freckles', 'Green yellow', 'green yellow', 'Blue/green/gold', 'Green with amber burst and gray outer ring', 'Ambar-Green', 'Ambar-green', 'hazel light green', 'Hazel light green', 'Green-hazel', 'Green-Hazel', 'Blue-green; amber collarette, and gray-blue ringing', 'Blue-green-grey', 'Hazel, olive green with amber starburst', 'Dark gray, blue, green (central heterochromia), yellow/brown ring around pupil, ', 'Hazel green', 'one brown one green', 'blue, grey, green, changing', 'Blue, grey, green, changing', 'Changes blue/green/grey', 'Changes with mood blue/grey/green', 'Blue/green/grey - changes with lighting and clothing', 'Changes with mood blue/grey/green', 'Blue/green/grey - changes with lighting and clothing', 'Blue/Green/Grey - changes with lighting and clothing'.

Class 3: Dark-Mixed/Brown (Martin-Schultz Scale 9–16)
'Black', 'Brown/black', 'Dark brown', 'dark brown', 'Dark brown', 'Brown', 'brown', 'Olive-Brown ringing Burnt Umber-Brown', 'Olive-brown ringing burnt umber-brown', 'indeterminate brown-green with a subtle grey caste', 'Indeterminate brown-green with a subtle grey caste', 'Hazel', 'hazel', 'Hazel/Light Brown', 'Hazel/light brown', 'Hazel (brown/green)', 'Hazel dark green', 'Hazel (light brown, dark green, dark blue)', 'Hazel/Yellow', 'Hazel/yellow', 'light brown with dark green tint', 'Light brown with dark green tint', 'green-brown', 'Brown-green', 'Green-brown', 'brown-green', ' green brown', 'Brown green starburst', 'Amber/Brown', 'Amber/brown', 'Brown-Amber', 'Brown-amber', 'Amber - (yellow/ocre brown)', 'Brown-(green when external temperature rises)', 'Brown - Brown and green in bright sunlight', 'Brown - brown and green in bright sunlight', 'Grey brown', 'Brown with blue outer ring', 'Brown inner, dark green outer', 'Brown center starburst, amber and olive green, with dark gray outer ring', 'Losing eye pigment as i age, currently in the light brown almost green phase'.

II. Hair color classifications:

Brown to Dark Brown/Black
'Black', 'Black ', 'Brown-Black', 'Brown-black', 'Blackish brown', 'Brown-black/Brown', 'Brown-black/Dark brown', 'Black/Black (very slight tint of red)', 'Black/Dark brown', 'Brown-black/Very dark brown', 'Black/Brown', 'Darkest brown to black ', 'Brown-Black/Darkest brown to black ', 'Brown-Black/Very dark brown', 'Darkest brown to black', 'Dark brown almost black', 'Brown-black/Dark brown almost black', 'Brown-Black/Dark brown almost black', 'Black/Medium golden brown', 'Very dark brown', 'Dark Brown/Very dark brown', 'Dark brown/Brunette', 'Brown to Dark brown', 'Dark Brown', 'dark brown', 'Dark brown', 'Brown to dark brown', 'Dark brown to brown', 'Brown/Dark brown', 'Dark brown/Brown', 'Medium to dark brown', 'Brown', 'brown', 'Brown going to white in early 40s', 'Brown and silver', 'Grey and brown', 'Brunette', 'Brown,red,blond', ‘Reddish brown', 'Strawberry blonde/Reddish Medium Brown', 'Medium brown', 'Medium Brown', 'Brown/Medium brown', 'med brown', 'Dark brown; blonde highlights', 'Medium brown with highlights', 'Reddish-brown/Very dark brown', 'Brown/Reddish brown', 'Dark brown; red highlights', 'Dark brown; red highlights/Medium to dark brown', 'medium brown, red highlights', 'Medium brown, red highlights', 'Brown/Reddish medium brown', 'Blondish reddish brown', 'Auburn (Reddish-Brown)', 'Auburn', 'Auburn (reddish-brown)', 'Reddish-brown', 'Reddish medium brown', 'Dark auburn', 'Between dark blonde and light brown/Dark auburn', 'Auburn (reddish-brown)/Medium golden brown', 'Auburn (reddish-brown)/Brown', 'Chestnut', 'Chestnut brown', 'strawberry brown', 'Strawberry brown', 'Strawberry blond as a child, now dark auburn brown', 'Dirt-Brown', 'Dirt-brown/Medium brown', 'Dirt-brown', 'Light to Medium brown', 'Light to medium brown', 'Medium Brown/Light brown', 'Medium golden brown', 'Medium Golden Brown', 'Light brown/Medium golden brown', 'medium brown/Light brown', 'Light brown', 'Light ashy brown', 'Chestnut/Light brown', 'Light Brown', 'light brown', 'light ashy brown ', 'Brown/Light brown', 'Toe head to dark reddish brown', 'blond born, today dark brown', 'Blond born, today dark brown', 'Dark blonde as a child, dark brown as an adult', 'Blond born, today dark brown/Grey and Brown', 'Hair darkening with age, starting blonde, ending dark brown', 'Brown/Blond as child. started turning dark brown after puberty', 'Blond as child. started turning dark brown after puberty', 'Blond born, today dark

brown/Brown', 'Blond born, today dark brown/Medium brown', 'Blonde to light brown as child, medium brown as adult with blonde highlights from sun', 'Dark blonde/Blonde to light brown as child, medium brown as adult with blonde highlights from sun', 'Dark Blond/Blonde to light brown as child, medium brown as adult with blonde highlights from sun', 'Light Brown/Blonde to light brown as child, medium brown as adult with blonde highlights from sun', 'very light blonde as child to med brown as adult', 'Blonde as a child, to brown as an adult', 'Brunette/Blonde as a child, to brown as an adult', 'Dark blonde as a child, chestnut brown as an adult', 'Blond as a child and light brown as an adult', 'Blonde as a child, light brown as an adult', 'Very light blonde as child to med brown as adult'.

Blonde to Dark Blonde
'Blonde', 'Blond', 'Blond grey', 'Blonde/Dark blonde ', 'Blond/Light blonde', 'Light brown/Blonde', 'light blonde as a child and medium blonde as an adult.', 'Blonde as child, ash blonde as adult, early white', 'Blond as a child. Dark blond as an adult.', 'Dark blonde/dark ash blonde, lightens in sun very easily. platinum blonde as child', 'Dark blonde/Blonde as a child, to brown as an adult', 'Blond as a child. dark blond as an adult.', 'Dark blonde ', 'Dark blonde', 'Dark Blond', 'Dark blond', 'Blonde/Dark blonde-light brown', 'Between dark blonde and light brown', 'Dark blonde with a little of every colour but black.', 'Dark blonde (light brown)', 'Dark blonde/light brown', 'Dirt-Blonde', 'Dirt-blonde', 'Dark blonde / Dirt-blonde', 'Dirt-blonde/Dark blonde', 'Dishwater blonde', 'Dirty blonde, light brown, something?', 'Dirt-blonde/Dark blonde ', 'Dirty Blond, Dark Red Beard', 'Strawberry blonde', 'Dark blonde, strawberry', 'Strawberry Blond', 'Light Brown/Dark blonde, ', 'Dirt-brown/Dark Blond'.

Appendix D: Populations Classifications into Subgroups

1. **Africa**

   African Ancestry in Southwest US, Americans of African Ancestry in SW USA, C. African Republic - Biaka Pygmy, D. R. of Congo - Mbuti Pygmy, Esan in Nigera, Gambian in Western Divisons in The Gambia, Kenya - Bantu, Luhya in Webuye, Kenya, Mende in Sierra Leone, MKK, Namibia - San, Nigeria - Yoruba, Nigeria - YRI HapMap, Senegal - Mandenka, South Africa - Bantu, Yoruba in Ibadan, Nigeria, CHD, African Carribbeans in Barbados, 03. Nigeria - Yoruba, 04. Namidia - San, 06. Central African Republic - Biaka Pygmies, BotswanaOrNamibia - BantuTswana, Congo - Mbuti, Dinka-3, Kenya - BantuKenya, Kenya - Luhya, Kenya - Luo, Namibia - Ju_hoan_North, SouthAfrica - Khomani_San, Sudan - Dinka, 06. Central African Republic - Biaka Pygmies

2. **Central and South America**

   Puerto Rican, Puerto Ricans from Puerto Rico, Mexican Ancestry, Mexican Ancestry from Los Angeles USA, Colombian, Colombians from Medellin Colombia, Peruvians from Lima Peru (detected admixture)

3. **Middle East**

   Algeria (Mzab) - Mozabite, Israel (Carmel) - Druze, Israel (Central) - Palestinian, Israel (Negev) - Bedouin, Iraq - Iraqi_Jew, Israel(Central) - Palestinian, Jordan - Jordanian, Western Sahara (Morocco) - Saharawi, Yemen - Yemenite_Jew, Iran - Iranian

4. **East Asia and China**

   Brunei - Dusun, Cambodia - Cambodian, in Ho Chi Minh City, Vietnam, Thai-1, Thai-2, Agta-1, Agta-2, Agta-3, Bajo-17, Bajo-19, Bajo-21, Balkars-1, Batak-1, Batak-2, Batak-3, Dusun-10, Dusun-11, Dusun-12, Dusun-14, Dusun-4, Dusun-5, Dusun-7, Dusun-8, Igorot-1, Igorot-2, Igorot-3, Igorot-3, Igorot-4, Igorot-4, Igorot-5, Igorot-6, Lebbo-1, Lebbo-2, Lebbo-3, Lebbo-4, Luzon-2, Luzon-6, Murut-11, Murut-13, Murut-19, Murut-3, Murut-4, Murut-5, Murut-6, Philippines - Igorot, Taiwan - Ami, Vietnamese_central-1, Vietnamese_central-2, Vietnamese_north-1, Vietnamese_north-2, Vietnamese_north-3, Vietnamese_south-1, Vietnamese_south-2, Vietnamese_south-3, Vietnamese_south-4, Vietnamese_south-5, Vizayan-1, Vizayan-3, Japan - Japanese, Japan - JPT HapMap, Japanese in Tokyo, Japan, 43. Japan - Japanese, Korea - Korean, Korean-2, China - Dai, China - Daur, China - Han, China - Hezhen, China - Lahu, China - Miaozu, China - Naxi, China - Oroqen, China - She, China - Tu, China - Tujia, China - Uygur, China - Xibo, China - Yizu, Chinese Dai in Xishuangbanna, China, Han Chinese, She-1, She-2, Southern Han Chinese, 31. China - Hezhen

5. **Eastern Europe and West Eurasia**

Russia (Caucasus) - Adygei, Kyrgyzystan - Kyrgyz, Bashkirs-10, Bashkirs-2, Bashkirs-3, Bashkirs-4, Bashkirs-6, Belarusians-1, Belarusians-2, Belarusians-3, Belarusians-4, Bulgaria - Bulgarian, Bulgaria - Bulgarian, Chuvashes-1, Chuvashes-2, Chuvashes-3, Croats-1, Croats-2, Croats-3, Croats-6, Czechoslovia(pre1989) - Czech, Estonia - Estonian, Estonians-1, Estonians-2, Estonians-3, Estonians-4, Estonians-5, Estonians-6, Finland - Saami, Finland - Saami, Greece - Crete, Greece - Crete, Greece - Greek, Greece - Greek, Hungarians-1, Hungarians-2, Hungary - Hungarian, Hungary - Hungarian, Ingrians-1, Ingrians-2, Ingrians-3, Latvians-1, Latvians-2, Latvians-3, Lithuanians-1, Lithuanians-2, Lithuanians-3, Maris-1, Maris-2, Maris-3, Maris-4, Moldavians-2, Moldavians-3, Mordvins-1, Mordvins-2, Mordvins-3, Poles-1, Poles-2, Poles-3, Poles-4, Russia(Caucasus) - Adygei, Russia(Caucasus) - Adygei, Udmurds-1, Udmurds-2, Udmurds-3, Udmurds-4, Ukrainians_east-1, Ukrainians_east-2, Ukrainians_east-3, Ukrainians_north-1, Ukrainians_west-1, Ukrainians_west-2, Ukrainians_west-3, Vepsas-1, Vepsas-2, Vepsas-3, Vepsas-4, Kumyks-1, Kumyks-2, Kumyks-3, Kryashen-Tatars-4, Kryashen-Tatars-5, Kryashen-Tatars-8, Tatars-1, Tatars-2, Tatars-3, Abkhazia - Abkhasian, Abkhazians-1, Abkhazians-5, Abkhazians-6, Albania - Albanian, Albanians-1, Albanians-2, Albanians-3, Armenia - Armenian, Armenia - Armenian, Armenians-1, Armenians-2, Armenians-3, Armenians-4, Armenians-5, Armenians-7, Avars-1, Avars-12, Avars-9, Azerbaijanis-13, Azerbaijanis-14, Azerbaijanis-24, Balkars-2, Balkars-4, Circassians-1, Circassians-2, Circassians-3, Cossacks_Kuban-1, Cossacks_Kuban-2, Cossacks-1, Cossacks-2, Georgia - Georgian, Georgia - Georgian, Georgians-1, Georgians-2, Kabardins-1, Kabardins-2, Kabardins-3, Kabardins-4, Kazakhs-1, Kazakhs-2, Kazakhs-3, Kyrgyz_Tdj-1, Kyrgyz_Tdj-2, Kyrgyz_Tdj-3, Kyrgyz-1, Kyrgyz-2, Kyrgyz-3, Kyrgyz-4, Lezgins-1, Lezgins-2, Lezgins-3, Lezgins-4, Poland - Polish, Rushan-Vanch-1, Rushan-Vanch-2, Russia - Abkhasian, Russia - Chechen, Russia - Lezgin, Russia - Lezgin, Russia - North_Ossetian, Russia-North_Ossetian, Tabasarans-4, Tabasarans-5, Tabasarans-7, Tajikistan - Tajik, Tajikistan - Tajik, Tajiks-1, Turkey - Turkish, Turkey - Turkish, Turkmens-2, Uzbek-1, Uzbek-2, Uzbek-3

6. **Europe**

British, British in England and Scotland, Finnish in Finland, France - Basque, France - French, Iberian population in Spain, Iberian populations in Spain, Icelandic-1, Icelandic-2, Italy - from Bergamo, Italy - Sardinian, Italy - Tuscan, Norwegian-1, Orkney Islands - Orcadian, Utah residents (CEPH) with Northern and Western European ancestry, Germans-1, Germans-2, Germans-3, Saami-4, Saami-5, Saami-6, Swedes-1, Swedes-2, Toscani in Italia, Toscani in Italy, Russia - Russian, Russia, Russians-Central-1, Russians-North-2, Russia, Russians-North-1, Russia

7. **Native Americans**

   Argentina - Chane, Brazil - Karitiana, Brazil - Surui, Colombia - Piapoco, Colombia - Piapoco and Curripaco, Mexico - Maya, Mexico - Mayan, Mexico - Mixe, Mexico - Mixtec, Mexico - Pima, Mexico - Zapotec, Puerto Rican - Quechua

8. **Northeast Asia and Siberia**

   Russia - Tlingit, China - Mongola, Russia - Aleut, Russia - Altaian, Russia - Chukchi, Russia - Eskimo_Sireniki, Russia - Even, Russia - Ulchi, Russia - Yakut, Siberia - Yakut, Altaians-1, Altaians-2, Altaians-3, Altaians-4, Altaians-5, Altaians-6, Buryats-11, Buryats-318, Buryats-336, Buryats-350, Buryats-355, Buryats-361, Buryats-383, Buryats-398, Buryats-406, Buryats-530, Buryats-561, Buryats-578, Buryats-6, Buryats-636, Buryats-639, Buryats-640, Eskimo-11, Eskimo-2, Eskimo-20, Eskimo-3, Evenks-1, Evenks-1, Evenks-14, Evenks-16, Evenks-2, Evenks-21, Evenks-22, Evenks-31, Evenks-35, Evenks-40, Evenks-41, Evenks-55, Evenks-62, Evens_Magadan-1, Evens_Magadan-2, Evens_Magadan-3, Evens_Magadan-3, Evens_Magadan-5, Evens_Sakha-1, Evens_Sakha-2, Evens_Sakha-3, Koryaks-1, Koryaks-10, Koryaks-11, Koryaks-12, Koryaks-13, Koryaks-14, Koryaks-15, Koryaks-16, Koryaks-2, Koryaks-3, Koryaks-4, Koryaks-5, Koryaks-6, Koryaks-7, Koryaks-8, Koryaks-9, Mongolians-1, Mongolians-2, Mongolians-3, Mongolians-4, Mongolians-5, Mongolians-6, Russia - Eskimo_Chaplin, Russia - Eskimo_Naukan, Russia - Eskimo_Naukan, Russia - Itelman, Russia - Mansi, Russia - Mansi, Yakut-K4, Yakuts-1, YakutS4, YakutS8, Yakuts-K1, Yakuts-K2, Yakuts-K3, Yakuts-M1

9. **Oceania**

   Bougainville - NAN Melanesian, New Guinea - Papuan, New Zealand - Maori, PapuaNewGuinea - Papuan, Australia - Australian, PapuaNewGuinea - Bougainville, PapuaNewGuinea - Bougainville,USA - Hawaiian

10. **South Asia**

    Bengali from Bangladesh, Gujarati Indians in Houston, TX, Indian Telugu from the UK, Pakistan - Balochi, Pakistan - Brahui, Pakistan - Burusho, Pakistan - Hazara, Pakistan - Kalash, Pakistan - Makrani, Pakistan - Pathan, Pakistan - Sindhi, Punjabi from Lahore, Pakistan, Sri Lankan Tamil from the UK, India - Brahmin, Nepal - Kusunda, Burmese-12, Burmese-14, Burmese-15, Burmese-20, Burmese-3

Appendix E: Classification Results for all models

Blood Type Prediction

Table 22. *Models' classification results for blood type prediction.*

Logistic Regression				
	Precision	Recall	F1-score	Support
A	0.947	0.954	0.950	281.000
AB	0.784	0.816	0.800	49.000
B	0.876	0.944	0.909	90.000
O	0.976	0.941	0.958	306.000
Macro avg	0.896	0.914	0.904	726.000
Weighted avg	0.940	0.938	0.938	726.000
Accuracy	0.938			
Random Forest				
	Precision	Recall	F1-score	Support
A	0.960	0.947	0.953	281.000
AB	0.816	0.816	0.816	49.000
B	0.944	0.944	0.944	90.000
O	0.968	0.980	0.974	306.000
Macro avg	0.922	0.922	0.922	726.000
Weighted avg	0.952	0.952	0.952	726.000
Accuracy	0.952			
Gradient Boosting				
	Precision	Recall	F1-score	Support
A	0.957	0.954	0.955	281.000
AB	0.932	0.837	0.882	49.000
B	0.946	0.967	0.956	90.000
O	0.968	0.980	0.974	306.000
Macro avg	0.951	0.934	0.942	726.000
Weighted avg	0.958	0.959	0.958	726.000
Accuracy	0.959			
Neural Network (MLP)				
	Precision	Recall	F1-score	Support
A	0.944	0.954	0.949	281.000
AB	0.844	0.776	0.809	49.000
B	0.912	0.922	0.917	90.000
O	0.967	0.967	0.967	306.000
Macro avg	0.917	0.905	0.910	726.000
Weighted avg	0.943	0.944	0.943	726.000
Accuracy	0.944			
Decision Tree Classifier				
	Precision	Recall	F1-score	Support
A	0.951	0.904	0.927	281
AB	0.683	0.837	0.752	49
B	0.923	0.933	0.928	90
O	0.968	0.974	0.971	306
Macro avg	0.881	0.912	0.895	726
Weighted avg	0.937	0.933	0.934	726
Accuracy	0.933			

Eye color prediction

Table 23. *Models' classification results for the 3-colors hair color prediction dataset.*

Logistic Regression				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.770	0.758	0.764	517
Dark Mixed/Brown	0.907	0.839	0.872	713
Green/Light-Mixed	0.273	0.356	0.309	202
macro avg	0.650	0.651	0.648	1432
weighted avg	0.768	0.742	0.754	1432
Accuracy	0.742			
Random Forest				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.712	0.805	0.756	517
Dark Mixed/Brown	0.864	0.899	0.881	713
Green/Light-Mixed	0.274	0.144	0.188	202
macro avg	0.617	0.616	0.608	1432
weighted avg	0.726	0.758	0.738	1432
Accuracy	0.758			
Gradient Boosting				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.716	0.903	0.799	517
Dark Mixed/Brown	0.867	0.902	0.884	713
Green/Light-Mixed	0.237	0.045	0.075	202
macro avg	0.607	0.617	0.586	1432
weighted avg	0.723	0.781	0.739	1432
Accuracy	0.781			
Neural Network (MLP)				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.722	0.814	0.765	517
Dark Mixed/Brown	0.842	0.854	0.848	713
Green/Light-Mixed	0.175	0.109	0.134	202
macro avg	0.580	0.592	0.583	1432
weighted avg	0.705	0.735	0.718	1432
Accuracy	0.735			
Decision Tree Classifier				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.680	0.665	0.673	517
Dark Mixed/Brown	0.833	0.774	0.802	713
Green/Light-Mixed	0.205	0.267	0.232	202
macro avg	0.573	0.569	0.569	1432
weighted avg	0.689	0.663	0.675	1432
Accuracy	0.663			

Table 24. *Models' classification results for the Blue/Blue-Mixed vs. Green/Light-Mixed subset.*

Logistic Regression				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.818	0.721	0.767	517
Green/Light-Mixed	0.452	0.589	0.512	202
macro avg	0.635	0.655	0.639	719
weighted avg	0.715	0.684	0.695	719
Accuracy	0.684			
Random Forest				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.769	0.841	0.803	517
Green/Light-Mixed	0.464	0.351	0.400	202
macro avg	0.616	0.596	0.602	719
weighted avg	0.683	0.704	0.690	719
Accuracy	0.704			
Gradient Boosting				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.779	0.928	0.847	517
Green/Light-Mixed	0.641	0.327	0.433	202
macro avg	0.710	0.628	0.640	719
weighted avg	0.740	0.759	0.731	719
Accuracy	0.759			
Neural Network (MLP)				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.771	0.861	0.814	517
Green/Light-Mixed	0.493	0.347	0.407	202
macro avg	0.632	0.604	0.610	719
weighted avg	0.693	0.716	0.699	719
Accuracy	0.716			
Decision Tree Classifier				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.763	0.723	0.743	517
Green/Light-Mixed	0.376	0.426	0.399	202
macro avg	0.569	0.575	0.571	719
weighted avg	0.654	0.640	0.646	719
Accuracy	0.640			

Table 25. *Models' classification results for the Blue/Blue-Mixed vs. Dark Mixed/Brown subset.*

Logistic Regression				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.887	0.942	0.914	517
Dark Mixed/Brown	0.956	0.913	0.934	71
macro avg	0.922	0.928	0.924	1230
weighted avg	0.927	0.925	0.925	1230
Accuracy	0.925			
Random Forest				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.887	0.913	0.900	517
Dark Mixed/Brown	0.936	0.916	0.926	71
macro avg	0.911	0.914	0.913	1230
weighted avg	0.915	0.915	0.915	1230
Accuracy	0.915			
Gradient Boosting				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.895	0.925	0.910	517
Dark Mixed/Brown	0.944	0.921	0.933	71
macro avg	0.920	0.923	0.921	1230
weighted avg	0.923	0.923	0.923	1230
Accuracy	0.923			
Neural Network (MLP)				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.880	0.890	0.885	517
Dark Mixed/Brown	0.919	0.912	0.915	71
macro avg	0.899	0.901	0.900	1230
weighted avg	0.903	0.902	0.903	1230
Accuracy	0.902			
Decision Tree Classifier				
	Precision	Recall	F1-score	Support
Blue/Blue-Mixed	0.834	0.838	0.836	517
Dark Mixed/Brown	0.882	0.879	0.881	71
macro avg	0.858	0.858	0.858	1230
weighted avg	0.862	0.862	0.862	1230
Accuracy	0.862			

Hair color prediction

Table 26. *Models' classification results for red vs. non-red hair color.*

Logistic Regression				
	Precision	Recall	F1-score	Support
No	0.929	0.829	0.876	111
Yes	0.689	0.857	0.764	49
macro avg	0.809	0.843	0.820	160
weighted avg	0.856	0.838	0.842	160
Accuracy	0.838			
Random Forest				
	Precision	Recall	F1-score	Support
No	0.889	0.937	0.912	111
Yes	0.837	0.735	0.783	49
macro avg	0.863	0.836	0.847	160
weighted avg	0.873	0.875	0.873	160
Accuracy	0.875			
Gradient Boosting				
	Precision	Recall	F1-score	Support
No	0.907	0.964	0.934	111
Yes	0.905	0.776	0.835	49
macro avg	0.906	0.870	0.885	160
weighted avg	0.906	0.906	0.904	160
Accuracy	0.906			
Neural Network (MLP)				
	Precision	Recall	F1-score	Support
No	0.898	0.955	0.926	111
Yes	0.881	0.755	0.813	49
macro avg	0.890	0.855	0.869	160
weighted avg	0.893	0.894	0.891	160
Accuracy	0.894			
Decision Tree Classifier				
	Precision	Recall	F1-score	Support
No	0.896	0.928	0.912	111
Yes	0.822	0.755	0.787	49
macro avg	0.859	0.842	0.849	160
weighted avg	0.873	0.875	0.873	160
Accuracy	0.875			

Table 27. *Models' classification results for Light vs. Dark hair colors.*

Logistic Regression				
	Precision	Recall	F1-score	Support
Blonde to Dark Blonde	0.388	0.776	0.517	192
Brown to Dark Brown/Black	0.937	0.731	0.821	874
macro avg	0.662	0.754	0.669	1066
weighted avg	0.838	0.739	0.767	1066
Accuracy	0.739			
Random Forest				
	Precision	Recall	F1-score	Support
Blonde to Dark Blonde	0.561	0.167	0.257	192
Brown to Dark Brown/Black	0.841	0.971	0.902	874
macro avg	0.701	0.569	0.579	1066
weighted avg	0.791	0.826	0.786	1066
Accuracy	0.826			
Gradient Boosting				
	Precision	Recall	F1-score	Support
Blonde to Dark Blonde	0.547	0.302	0.389	192
Brown to Dark Brown/Black	0.860	0.945	0.901	874
macro avg	0.704	0.624	0.645	1066
weighted avg	0.804	0.829	0.809	1066
Accuracy	0.829			
Neural Network (MLP)				
	Precision	Recall	F1-score	Support
Blonde to Dark Blonde	0.424	0.391	0.407	192
Brown to Dark Brown/Black	0.868	0.883	0.876	874
macro avg	0.646	0.637	0.641	1066
weighted avg	0.788	0.795	0.791	1066
Accuracy	0.795			
Decision Tree Classifier				
	Precision	Recall	F1-score	Support
Blonde to Dark Blonde	0.383	0.365	0.373	192
Brown to Dark Brown/Black	0.862	0.871	0.866	874
macro avg	0.622	0.618	0.620	1066
weighted avg	0.776	0.780	0.777	1066
Accuracy	0.780			

Table 28. *Models' classification report for Ancestry*

Logistic Regression				
	Precision	Recall	F1-score	Support
African	0.997	0.996	0.997	1030
Central and South American	0.930	0.913	0.922	277
East Asia and China	0.978	0.972	0.975	909
Eastern European and West Eurasia	0.176	0.158	0.167	19
European	0.964	0.968	0.966	688
Middle East	0.796	0.883	0.837	163
Native Americans	0.954	0.965	0.960	86
Northeast Asia and Siberia	0.490	0.595	0.538	42
Oceania	1.000	0.967	0.983	30
South Asian	0.970	0.942	0.956	583
macro avg	0.826	0.836	0.830	3827
weighted avg	0.958	0.957	0.957	3827
Accuracy	0.957			
Random Forest				
	Precision	Recall	F1-score	Support
African	0.994	0.995	0.995	1030
Central and South American	0.953	0.661	0.780	277
East Asia and China	0.949	0.993	0.970	909
Eastern European and West Eurasia	0.000	0.000	0.000	19
European	0.780	0.997	0.875	688
Middle East	0.892	0.405	0.557	163
Native Americans	0.988	0.942	0.964	86
Northeast Asia and Siberia	1.000	0.024	0.047	42
Oceania	1.000	0.933	0.966	30
South Asian	0.930	0.937	0.933	583
macro avg	0.849	0.689	0.709	3827
weighted avg	0.923	0.920	0.908	3827
Accuracy	0.920			
Gradient Boosting				
	Precision	Recall	F1-score	Support
African	0.996	0.993	0.995	1030
Central and South American	0.922	0.852	0.886	277
East Asia and China	0.961	0.988	0.974	909
Eastern European and West Eurasia	0.000	0.000	0.000	19
European	0.909	0.972	0.940	688
Middle East	0.814	0.804	0.809	163
Native Americans	0.952	0.919	0.935	86
Northeast Asia and Siberia	0.900	0.214	0.346	42
Oceania	0.880	0.733	0.800	30
South Asian	0.948	0.967	0.958	583
macro avg	0.828	0.744	0.764	3827
weighted avg	0.944	0.949	0.944	3827
Accuracy	0.949			

Neural Network (MLP)				
	Precision	Recall	F1-score	Support
African	0.995	0.995	0.995	1030
Central and South American	0.933	0.899	0.915	277
East Asia and China	0.972	0.988	0.980	909
Eastern European and West Eurasia	0.182	0.105	0.133	19
European	0.954	0.974	0.964	688
Middle East	0.823	0.828	0.826	163
Native Americans	0.920	0.930	0.925	86
Northeast Asia and Siberia	0.760	0.452	0.567	42
Oceania	1.000	0.967	0.983	30
South Asian	0.952	0.961	0.956	583
macro avg	0.849	0.810	0.824	3827
weighted avg	0.956	0.958	0.956	3827
Accuracy	0.958			
Decision Tree Classifier				
	Precision	Recall	F1-score	Support
African	0.953	0.954	0.954	1030
Central and South American	0.592	0.523	0.556	277
East Asia and China	0.910	0.912	0.911	909
Eastern European and West Eurasia	0.000	0.000	0.000	19
European	0.743	0.808	0.774	688
Middle East	0.479	0.497	0.488	163
Native Americans	0.788	0.779	0.784	86
Northeast Asia and Siberia	0.273	0.286	0.279	42
Oceania	0.806	0.833	0.820	30
South Asian	0.740	0.702	0.720	583
macro avg	0.628	0.629	0.629	3827
weighted avg	0.809	0.812	0.810	3827
Accuracy	0.812			