Forensic Science International: Genetics 53 (2021) 102507

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Evaluation of supervised machine-learning methods for predicting appearance traits from DNA

Maria-Alexandra Katsara [a], Wojciech Branicki [b], Susan Walsh [c], Manfred Kayser [d], Michael Nothnagel [a,e,*], on behalf of the VISAGE Consortium

[a] Cologne Center for Genomics, University of Cologne, Cologne, Germany
[b] Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland
[c] Department of Biology, Indiana University Purdue University Indianapolis (IUPUI), Indianapolis, IN, USA
[d] Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands
[e] Faculty of Medicine and the Cologne University Hospital, Cologne, Germany

ARTICLE INFO

Keywords:
Externally visible characteristics
Predictive DNA analysis
Appearance prediction
Genetic prediction
DNA phenotyping
Forensic DNA phenotyping
Machine learning
Classifiers

ABSTRACT

The prediction of human externally visible characteristics (EVCs) based solely on DNA information has become an established approach in forensic and anthropological genetics in recent years. While for a large set of EVCs, predictive models have already been established using multinomial logistic regression (MLR), the prediction performances of other possible classification methods have not been thoroughly investigated thus far. Motivated by the question to identify a potential classifier that outperforms these specific trait models, we conducted a systematic comparison between the widely used MLR and three popular machine learning (ML) classifiers, namely support vector machines (SVM), random forest (RF) and artificial neural networks (ANN), that have shown good performance outside EVC prediction. As examples, we used eye, hair and skin color categories as phenotypes and genotypes based on the previously established IrisPlex, HIrisPlex, and HIrisPlex-S DNA markers. We compared and assessed the performances of each of the four methods, complemented by detailed hyper-parameter tuning that was applied to some of the methods in order to maximize their performance. Overall, we observed that all four classification methods showed rather similar performance, with no method being substantially superior to the others for any of the traits, although performances varied slightly across the different traits and more so across the trait categories. Hence, based on our findings, none of the ML methods applied here provide any advantage on appearance prediction, at least when it comes to the categorical pigmentation traits and the selected DNA markers used here.

1. Introduction

In recent years, Forensic DNA Phenotyping (FDP), used to predict Externally Visible Characteristics (EVCs) of unknown crime scene sample donors or unknown deceased persons directly from DNA, has become a suitable addition to the forensic genetics toolbox. In criminal cases where suspects are unknown to the investigating authorities and therefore cannot be identified by comparative forensic DNA profiling, FDP can be used to generate investigative leads to help find unknown suspected perpetrators, and can also help in missing person identification when known relatives or ante mortem samples are not available [1–3]. By using FDP outcomes, police investigators can narrow down a large number of potential suspects, as is the case without known suspects, and they can subsequently proceed to generate standard forensic STR profiles for a reduced set of individuals that visually share such EVC FDP predicted outcomes.

As a prerequisite for developing FDP markers, in the past decade many studies have identified genetic markers involved in pigmentation traits [4–11]. Moreover, other studies have used them for developing lab tools and statistical tools for predicting eye, hair and skin color through DNA markers [12–20]. Most widely used predictive marker sets, lab tools and statistical models include in the IrisPlex system [13,17,21] for eye color prediction, the HIrisPlex system [20] for hair (and eye) color prediction, and the HIrisPlex-S system [19] for skin (and hair and eye)

* Correspondence to: Cologne Center for Genomics, Department of Statistical Genetics and Bioinformatics, University of Cologne, Weyertal 115b, 50931 Cologne, Germany.
E-mail address: michael.nothnagel@uni-koeln.de (M. Nothnagel).

https://doi.org/10.1016/j.fsigen.2021.102507
Received 6 November 2020; Received in revised form 26 February 2021; Accepted 17 March 2021
Available online 23 March 2021

color prediction. The aforementioned statistical models are based on multinomial logistic regression (MLR) using established genetic marker panels, resulting in posterior probabilities for each trait category i.e., three eye color, four hair color, and five skin color categories [19], and are publically available for use via https://hirisplex.erasmusmc.nl/. Almost all previously established pigmentation prediction models were based on MLR. Some exceptions include fuzzy logic, artificial neural networks and classification trees used by Liu et al. [13] for eye colour prediction modelling and Snipper [14], which is a Bayesian classifier that provides the prediction results as likelihood ratios. Further exceptions include the iterative naïve Bayesian approach from Maroñas and Söchtig [22,23] for skin and hair color respectively, and classification trees and partition modeling applied by Allwood et al. [24] (see [25] for a further review).

Currently, machine learning (ML) has become a powerful and widely used method for solving classification and clustering problems. It is a field in data analytics that focuses on the development of mathematical models that have the ability to recognize patterns in the datasets and use this information to predict future events. In parts inspired by the human brain, these algorithms can be trained on the data (training data) [26]. The training data is actually a set of examples which are used in order to fit, or estimate, the parameters of the model. The use of these algorithms is motivated by problems with large numbers of classes, linear and non-linear boundaries between them and can be implemented for different applications in versatile areas such as such as those observed in medicine, education, robotics and many others [27–29]. These boundaries refer to the decision boundaries, a hyper-surface that separates the vector space in mutually exclusive sets, one for each class. They can be either straight lines or non-linear curves. Some indicative examples of ML algorithms are linear and logistic regression [30], decision trees, random forests (RF) [31], k-nearest neighbors (k-NN) [32], support-vector machines (SVM) [33] and artificial neural networks (ANN) [34]. Despite the fact that these methods have huge potential in different fields, and an ability to handle various types of data, selecting a ML algorithm for specific data sets (problems) as well as their optimal hyperparameters to gain maximal performance can be challenging. A comparative analysis is often necessary in order to arrive at a method that provides the best prediction accuracy for the data set used.

In the context of forensic sciences, various classifiers have been used and compared for different purposes, such as the inference of biogeographic ancestry from DNA, file type detection - the identification of evidential files that criminals hide in order to mislead police authorities, glass identification etc. [35–40]. To the best of our knowledge, a systematic quantitative comparative performance analysis of different classification methods for DNA-based prediction of appearance traits has not been conducted thus far, except for some Naïve Bayes approaches [14,16].

In this study, we focused on the evaluation and comparison of three different popular ML approaches, namely SVM, RF and ANN, and compared them with MLR, for the set of EVCs most widely used in FDP, namely categorical eye, hair and skin color and by using the previously established DNA predictors from the IrisPlex, HIrisPlex, and HIrisPlex-S systems. These ML methods have gained a lot of importance in many different application areas and, despite their higher computational cost, are well-known for their often very good prediction performances; however, within the context of FDP, they have barely been used. The main motivation of this work is to assess whether any of these ML approaches has a higher prediction performance compared with the standard MLR that is currently widely applied in the context of EVC prediction, as one may expect from the experience in other areas. In this study, all methods are applied to two different datasets, namely one containing samples from different continental ancestries and one including only the European samples thereof, and results are compared. For all four methods, we assess the standard performance for each trait category and overall, for each trait, with the aim to investigate whether ML methods are superior, or not, over conventional MLR for DNA-based appearance prediction using pigmentation traits as examples.

2. Materials and methods

2.1. Data sets

For the present study, part of the previously used datasets for the establishment of IrisPlex model for eye color [17], the HIrisPlex model for hair color [20], and the HIrisPlex-S model for skin color [19] were applied for the prediction of those EVCs. More specifically, we used phenotype and genotype datasets from 1095 samples for eye, 1702 for hair, and 1318 for skin color prediction (*complete dataset;* **CD**), originating from Europeans, Americans, South and East Asians, African, Middle Eastern and few admixed samples. Furthermore, we used the *European subset* (**ES**) of this collection in order to restrict the analysis to a more homogenous population, comprising 821 samples for eye, 1429 for hair, and 980 for skin color prediction and originating from Ireland, Poland, Russia, Germany and Spain. These datasets were randomly split into 80% for model training and 20% for model evaluation (Table 1) for all four methods (see below).

Samples from which these data were previously obtained had been collected for the purpose of appearance genetic research under written informed consent, and sample collections were approved by the Ethics Committee of the Jagiellonian University (KBET/17/B/2005), the Commission on Bioethics of the Regional Board of Medical Doctors in Krakow (48 KBL/OIL/2008), the Clinical Research Ethics Committee of the Cork Teaching Hospitals (ref ECM 4 (dd) 11/01/11) and by the Indiana University Ethical Institutional Review Board (#1409306349).

For all available datasets considered here, we used the same eye, hair, and skin color categorization as previously applied and already established. These well-defined broad categories have been used in a number of studies before and could be considered to be close to a standard for trait categories for the time being. Furthermore, such broad categorization, with its clear distinction between a few trait categories, may better serve their application in police investigations rather than some sort of continuous scales that are more difficult to be distinguished. Furthermore, such categories are likely to be closer genetically, likely negatively affecting the respective prediction model's performance due to a larger genetic overlap between categories, rendering the model less able to distinguish between categories. For these reasons and as previously described in detail [17,19,20], eye colour was classified into three categories (blue, intermediate, brown) and hair colour into four categories (red, blond, brown, black), while skin colour was classified into five categories (very pale, pale, intermediate, dark, dark to black), following previously established categories. Since the European subset did not comprise samples with dark or dark to black skin colour, analyses in this subset were based on three categories only (very pale, pale, intermediate). The 41 HIrisPlex-S DNA markers were previously described by Chaitanya et al. [19]. In brief, for eye colour, hair colour, and skin colour, we applied the 6 SNPs from the previously established IrisPlex model for eye color prediction [17]; the 22 SNPs used for hair color prediction from the previously reported HIrisPlex model [20], and the 36 SNPs applied for the skin color prediction from the previously described HIrisPlex-S model [19], respectively.

Table 1
EVC-specific data sets used for prediction model training and testing for all four classification methods.

Appearance trait	N Training set (80%)	N Test set (20%)	Data references
Eye color	876 (656)	219 (165)	[17,19,20]
Hair color	1361 (1143)	341 (286)	
Skin color	1054 (784)	264 (196)	

Given are the numbers for the complete dataset (**CD**) and, in paratheses, for the European subset (**ES**).

2.2. Appearance trait categories

Trait categories were coded as *categorical* variables and ascendingly named as '1', '2', '3' etc. up to the corresponding number of categories for each trait:

- Eye color: Blue (1), Intermediate (2), Brown (3)
- Hair color: Blond (1), Brown (2), Red (3), Black (4)
- Skin color: Very Pale (1), Pale (2), Intermediate (3), Dark (4), Dark to Black (5); the latter two were considered only for the complete dataset

Total samples of each color category for each trait are described in detail in Supplementary Table S1. The genetic markers included in the model were converted from their initial form of the bases adenine (A), cytosine (C), guanine (G) and thymine (T) and coded *numerically* as 0, 1, 2 where 0 indicates homozygosity of the major allele, 1 heterozygosity and 2 homozygosity of the minor allele. For example, for an autosomal marker with major allele C and minor allele T, an individual's genotype CC, CT and TT would be converted to 0, 1 and 2, respectively. In all models no interaction terms were taken into account, thus only the additive effects of the corresponding genetic markers were included, similar to the previously established models [17,19,20]. Given the simple nature of our data and their final coding form as described above, we did not pursue feature engineering, such as considering squared variables or their products, since this would most likely not strongly affect our final outcomes. All data sets were previously quality controlled [17,19,20], including deviations from Hardy-Weinberg equilibrium, excessive heterozygosity, low minor allele frequencies, genetic outlier detection using principal-components analysis etc., and could therefore be directly used for prediction modelling. Samples with missing genotype data were excluded from our analysis.

2.3. Statistical analysis

The analysis was conducted in R version 3.4.3 [41] and 'RStudio' version 3.5.1 [42] using the packages 'nnet' [43], 'caret' [44], 'e1071' [45] and 'randomForest' [46]. Samples with missing genotype information were excluded.

2.4. Classification algorithms and hyperparameter tuning

We conducted a comparative statistical analysis in order to obtain the efficacy and classification accuracy of four different classification methods, namely Multinomial Logistic Regression (MLR), Support Vector Machines (SVM), Random Forest (RF) and Artificial Neural Networks (ANN). Tuned hyperparameters play an important role in obtaining the optimal performance and accuracy results when using SVM, RF and ANN. Each classifier requires different tuning steps and hyperparameters that need tuning and tuned values depend each time on the training dataset. For each classifier, we tested a series of values for the tuning process with the optimal hyperparameters determined based on the lower out-of-bag (OOB) prediction error. OOB is an estimation that measures the prediction error of each method. The classified results based on the optimal set of hyperparameters were used afterwards for the comparison of all classifiers. In order to assess the accuracy of classification performances, we report metrics such as sensitivity, specificity, positive predictive value, negative predictive value, area under curve, confusion matrix and overall accuracy were reported.

2.5. Multinomial logistic regression (MLR)

The MLR approach is a ML classification method that is used to predict a nominal dependent variable based on multiple independent variables. The independent variables can be either continuous or dichotomous. It is a simple extension of the binary logistic regression that allows the dependent variable to have more than two categories. Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation in order to evaluate the probability of each category. The model can be defined as follows for the 3-class traits [30]:

$$ln\left(\frac{p_2}{p_1}\right) = \alpha_2 + \sum_{j=1}^{k} \beta_2 (p_2)_j x_j \tag{1}$$

$$ln\left(\frac{p_3}{p_1}\right) = \alpha_3 + \sum_{j=1}^{k} \beta_3 (p_3)_j x_j \tag{2}$$

Where $\alpha_i, \beta_i (i = 2, 3)$ are the regression coefficients and $p_i (i = 1, 2, 3)$ are denoting the probabilities for each individual sample to belong to a certain category. The latter can be calculated as follows:

$$p_2 = \frac{\exp\left(a_2 + \sum_{j=1}^{k} \beta_2 (p_2)_j x_j\right)}{1 + exp\left(a_2 + \sum_{j=1}^{k} \beta_2 (p_2)_j x_j\right) + exp\left(a_3 + \sum_{j=1}^{k} \beta_3 (p_3)_j x_j\right)} \tag{3}$$

$$p_3 = \frac{\exp\left(a_3 + \sum_{j=1}^{k} \beta_3 (p_3)_j x_j\right)}{1 + exp\left(a_3 + \sum_{j=1}^{k} \beta_3 (p_3)_j x_j\right) + exp\left(a_2 + \sum_{j=1}^{k} \beta_2 (p_2)_j x_j\right)} \tag{4}$$

$$p_1 = 1 - p_2 - p_3 \tag{5}$$

where x_j is the number of minor (less frequent) allele of the j^{th} SNP and j is an indicator for the number of the genetic markers included for trait prediction. For this method no parameter tuning was done. Individuals were classified to the colour category with the maximum probability p_i without any threshold values to be taken into account.

2.6. Support vector machines (SVM)

SVM [33] is a machine learning approach which finds the optimal hyperplane that separates the different classes with the maximum margin, i.e. the maximum distance between the data points that belong to the different categories. It can solve linear or non-linear problems regarding the kernel function used each time [47]. In our case, we applied the Gaussian radial basis function (RBF) which is a widely used kernel appropriate for non-linear classification. It can be defined as follows:

$$\mathrm{K}(X_1, X_2) = \exp(-\gamma \|X_1 - X_2\|^2) \tag{6}$$

where $\|X_1 - X_2\|$ is the Euclidean distance between the data points X_1, X_2. There are two parameters that need to be tuned when using SVM classifier with RBF kernel: the parameters of cost (C) and the kernel width parameter (γ). The parameter C determines the influence of the misclassification on the objective function and γ the shape and the smoothing of the optimal hyperplane obtained. These two parameters can significantly affect the performance of an SVM model. More specifically, large C values may lead to over-fitting models while large γ could affect the shape of the hyperplane which, as a result, can affect the classification outcomes. In order to find the optimal parameters for both CD and ES, we applied the grid-searching process between ten values of γ (2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}) and ten values of C (2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}, 2^{5}, 2^{6}, 2^{7}). This procedure was applied for all three traits tested and the optimal values were chosen according to the lowest OOB error (Supplementary Figs. S1 and S4).

2.7. Random forest (RF)

The RF [31] is a ML method for classification and regression tasks. It operates by constructing multiple decision trees during training and, in order to classify a new instance, each decision tree provides a classification for input data. The majority-vote classification is then chosen as the prediction. In its implementation we chose to tune two hyperparameters: the number of trees (*ntree*) and the number of features at each split (*mtry*). Several studies have already been published that focus on the appropriate number of trees for which one could obtain optimal results from the RF model. However, different opinions have been voiced during these studies. One typical example is the study of Liaw and Wiener [46] which states that larger numbers of trees provide more stable results of variable importance. On the other hand, studies such as those by Latinne et al. [48], and Hernandez-Lobato [49], found that smaller numbers of trees can also be sufficient. The study of Oshiro et al. [50] comprehensively addressed this question by applying the RF model to 29 different data sets and comparing their Area Under Curve (AUC) values. The main conclusion of this study was that the performance of an RF model does not necessarily improve when number of trees is increased, suggesting that a range between 64 and 128 trees can provide satisfactory results.

For optimal tree number (*ntree*), we checked and compared the OOB error rate for a range of 1–1000 trees and chose, separately for each trait, that number which resulted in the lowest OOB error rate. In Supplementary Figs. S2 and S5 the best values for each trait for both CD and ES are presented. For optimal *mtry* hyperparameter values, we used the default of the integer-rounded value of $\sqrt{p}$, where p denotes the number of variables in the model, i.e. the number of genetic markers. The corresponding mtry values for the two datasets for eye, hair and skin color therefore equaled 2, 4 and 6, respectively.

2.8. Artificial neural networks (ANN)

ANN [34,51] is a family of approaches for classification and clustering that was inspired by the human brain in order to recognize patterns in data sets. Its history starts from the early 1940s where McCulloch and Pitts [52] wrote a paper on the functionality of human brain neurons and modeled a simple neural network by using electrical circuits. Later on 1949 Donald Hebb [53] introduced the fundamental idea of learning by supporting that neural pathways are strengthened every time that are used (Hebbian learning). In the 1950s when computers became more advanced, many ANN approaches were developed and simulated. Some examples were the approach of Farley and Clark [54], who simulated the aforementioned Hebbian Network and also the approach of Rosenblatt [55], who created the perceptron, an algorithm for pattern recognition. The interest of ANN continued also in the 1970 s where Werbos [56] introduced the backpropagation algorithm that enabled the training of multi-layer networks. More recent approaches have already been established, and successfully addressed the previous challenges of deep neural networks [57–59].

The ANN consists of connected units, or nodes, called artificial neurons and these connections, just as the functionality of the human brain, can transmit signals or activate other neurons [60]. Most ANN are organized in layers and neurons, and the input data are "moving" through them only in the forward direction until some final output is obtained. Each node has its own weight which is continuously adjusted during the training procedure until data with same labels consistently yield similar output.

A number of parameters need to be tuned in order to obtain the maximum performance of the ANN model. Here, we started by tuning the number of hidden layers. At first, we looked at a range of values, starting from 1 till 10 for the hidden layers. We obtained no significant differences in the model performance for eye color prediction when we increased the number of layers. For hair and skin color prediction, we noticed some deterioration in the model performance as we increased the number of layers. Therefore, for all three traits considered here we trained our models using only one hidden layer and used the logistic function as the activation function. Other parameters that required tuning were the layer size, referring to the number of units in the hidden layer, and the decay value, acting as a regularization parameter to avoid over-fitting. Supplementary Figs. S3 and S6 give the optimal values for CD and ES respectively, according to the lowest OOB error, chosen for each of the traits.

2.9. Accuracy assessment and comparisons

In order to compare the performance of the different classifiers we presented the model measurements evaluated on the corresponding test datasets. More specifically, for each model we calculated the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under curve (AUC), confusion matrix and overall accuracy. Sensitivity (true positive rate) measures the proportion of the actual positive samples that are correctly identified by the model while specificity (true negative rate) refers to the proportion of the actual negative samples that were correctly identified. In addition, PPV denotes the proportion of the correct classifications among all predictions of the trait category tested each time, and NPV refers to the proportion of the correct classifications among all predictions other than the trait category of interest. AUC is a performance measure of a classification model across all possible classification thresholds while the confusion matrix describes the performance of a classification model on the test dataset for which the true values are known. Ultimately, the overall accuracy refers to the proportion of all samples that were classified correctly.

3. Results

3.1. Parameter tuning

For three out of the four methods applied, namely SVM, RF and ANN, we proceeded into parameter tuning for each of the two datasets and for the three traits (i.e. eye, hair and skin color) in order to obtain the optimal performance of the classifiers. The best parameters were chosen according to the lowest out-of-bag (OOB) error. For SVM, the parameters that needed to be tuned were γ and C. We found out that the optimal value for γ was 0.03125 for all three traits and for both CD and ES. The optimal C in the CD was equal to 2 for eye and skin color and equal to 16 for hair color (Supplementary Fig. S1). For the ES, optimal value of C was equal to 1 for eye and skin color and equal to 8 for hair color

Table 2
Overall accuracy of the EVC predictions by the four classifiers.

		MLR	SVM	RF	ANN
Eye Color	CD	0.79 (0.73–0.84)	0.78 (0.72–0.83)	0.78 (0.71–0.83)	0.79 (0.73–0.84)
	ES	0.69 (0.61–0.76)	0.68 (0.60–0.75)	0.67 (0.59–0.74)	0.69 (0.61–0.76)
Hair Color	CD	0.60 (0.55–0.65)	0.57 (0.50–0.60)	0.55 (0.49–0.60)	0.58 (0.49–0.60)
	ES	0.59 (0.54–0.65)	0.55 (0.49–0.61)	0.53 (0.47–0.59)	0.56 (0.50–0.61)
Skin Color	CD	0.63 (0.57–0.69)	0.60 (0.53–0.65)	0.59 (0.52–0.64)	0.56 (0.49–0.66)
	ES	0.65 (0.58–0.72)	0.65 (0.58–0.71)	0.66 (0.59–0.72)	0.57 (0.50–0.64)

MLR: multinomial logistic regression; **SVM**: support-vector machine; **RF**: random forest; **ANN**: artificial neural network. **CD**: complete dataset; **ES**: European subset.

Table 3
Predictive measurements for eye color for the four classifiers.

		MLR			SVM			RF			ANN		
Category		1	2	3	1	2	3	1	2	3	1	2	3
Sensitivity	CD	0.93 (0.87–0.97)	0.18 (0.09–0.32)	0.91 (0.82–0.95)	0.93 (0.87–0.97)	0.13 (0.05–0.26)	0.91 (0.82–0.95)	0.92 (0.86–0.96)	0.15 (0.05–0.26)	0.91 (0.82–0.95)	0.93 (0.86–0.96)	0.20 (0.12–0.38)	0.91 (0.82–0.95)
	ES	0.84 (0.75–0.91)	0.23 (0.13–0.38)	0.82 (0.67–0.90)	0.83 (0.73–0.90)	0.23 (0.13–0.38)	0.80 (0.66–0.89)	0.83 (0.73–0.90)	0.26 (0.15–0.41)	0.73 (0.60–0.84)	0.82 (0.72–0.89)	0.26 (0.15–0.41)	0.84 (0.71–0.91)
Specificity	CD	0.72 (0.63–0.80)	0.97 (0.94–0.99)	0.93 (0.88–0.96)	0.74 (0.64–0.80)	0.98 (0.95–0.99)	0.89 (0.84–0.94)	0.74 (0.64–0.80)	0.98 (0.94–0.99)	0.90 (0.84–0.94)	0.72 (0.65–0.81)	0.97 (0.94–0.99)	0.94 (0.87–0.96)
	ES	0.68 (0.58–0.77)	0.92 (0.86–0.96)	0.89 (0.82–0.93)	0.66 (0.56–0.75)	0.92 (0.86–0.96)	0.89 (0.82–0.93)	0.64 (0.53–0.73)	0.89 (0.82–0.93)	0.92 (0.86–0.96)	0.67 (0.57–0.76)	0.94 (0.89–0.97)	0.87 (0.80–0.92)
PPV	CD	0.75 (0.67–0.82)	0.58 (0.32–0.81)	0.87 (0.78–0.93)	0.76 (0.68–0.82)	0.63 (0.31–0.86)	0.81 (0.89–0.95)	0.76 (0.67–0.82)	0.60 (0.24–0.76)	0.82 (0.73–0.90)	0.75 (0.68–0.83)	0.62 (0.39–0.84)	0.88 (0.78–0.92)
	ES	0.70 (0.60–0.78)	0.47 (0.27–0.68)	0.75 (0.62–0.85)	0.68 (0.58–0.77)	0.47 (0.27–0.68)	0.75 (0.62–0.85)	0.67 (0.57–0.75)	0.42 (0.24–0.61)	0.80 (0.66–0.89)	0.68 (0.58–0.77)	0.59 (0.36–0.78)	0.73 (0.60–0.83)
NPV	CD	0.92 (0.85–0.96)	0.84 (0.78–0.88)	0.95 (0.90–0.98)	0.92 (0.85–0.96)	0.83 (0.78–0.88)	0.95 (0.90–0.97)	0.91 (0.84–0.96)	0.84 (0.78–0.88)	0.95 (0.90–0.98)	0.92 (0.84–0.96)	0.84 (0.79–0.89)	0.95 (0.90–0.98)
	ES	0.83 (0.73–0.90)	0.79 (0.72–0.85)	0.92 (0.85–0.96)	0.82 (0.71–0.89)	0.79 (0.72–0.85)	0.91 (0.84–0.95)	0.81 (0.70–0.89)	0.79 (0.72–0.85)	0.89 (0.82–0.93)	0.80 (0.70–0.88)	0.80 (0.73–0.86)	0.93 (0.86–0.96)

Eye color categories: **1**: Blue; **2**: Intermediate; **3**: Brown. **MLR**: multinomial logistic regression; **SVM**: support-vector machine; **RF**: random forest; **ANN**: artificial neural network. **PPV**: Positive predictive value; **NPV**: negative predictive value. **CD**: complete dataset; **ES**: European subset.

Table 4
Predictive measurements for hair color for the four classifiers.

		MLR				SVM				RF				ANN			
Category		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Sensitivity	CD	0.70 (0.62–0.77)	0.59 (0.50–0.67)	0.66 (0.47–0.80)	0.20 (0.11–0.38)	0.66 (0.58–0.73)	0.65 (0.57–0.73)	0.21 (0.10–0.38)	0	0.67 (0.59–0.74)	0.56 (0.45–0.62)	0.28 (0.15–0.46)	0.26 (0.14–0.42)	0.69 (0.59–0.74)	0.46 (0.38–0.55)	0.58 (0.41–0.74)	0.31 (0.19–0.48)
	ES	0.81 (0.73–0.87)	0.43 (0.35–0.52)	0.69 (0.50–0.83)	0.26 (0.13–0.47)	0.88 (0.81–0.93)	0.40 (0.32–0.49)	0.23 (0.11–0.42)	0	0.78 (0.70–0.85)	0.44 (0.35–0.53)	0.31 (0.17–0.50)	0	0.72 (0.63–0.79)	0.46 (0.38–0.55)	0.62 (0.43–0.78)	0.17 (0.07–0.37)
Specificity	CD	0.70 (0.63–0.76)	0.68 (0.63–0.76)	0.98 (0.96–0.99)	0.98 (0.96–0.99)	0.68 (0.60–0.73)	0.58 (0.51–0.64)	1	1	0.62 (0.55–0.69)	0.67 (0.60–0.73)	0.99 (0.98–0.99)	0.97 (0.94–0.98)	0.67 (0.60–0.73)	0.67 (0.61–0.74)	0.99 (0.97–0.99)	0.93 (0.90–0.95)
	ES	0.57 (0.50–0.64)	0.82 (0.76–0.87)	0.98 (0.95–0.99)	0.97 (0.94–0.98)	0.41 (0.34–0.49)	0.83 (0.77–0.88)	0.99 (0.98–0.99)	1	0.48 (0.41–0.56)	0.75 (0.68–0.81)	0.99 (0.97–0.99)	0.99	0.59 (0.52–0.67)	0.73 (0.65–0.79)	0.99 (0.97–0.99)	0.96 (0.93–0.98)
PPV	CD	0.63 (0.55–0.70)	0.54 (0.48–0.64)	0.79 (0.60–0.91)	0.58 (0.32–0.81)	0.60 (0.51–0.66)	0.50 (0.43–0.58)	1	NA	0.56 (0.49–0.63)	0.52 (0.43–0.59)	0.80 (0.49–0.94)	0.47 (0.27–0.68)	0.60 (0.52–0.67)	0.48 (0.40–0.57)	0.85 (0.64–0.95)	0.34 (0.20–0.52)
	ES	0.56 (0.49–0.64)	0.64 (0.53–0.74)	0.75 (0.55–0.88)	0.43 (0.22–0.67)	0.50 (0.44–0.57)	0.64 (0.52–0.73)	0.86 (0.49–0.97)	NA	0.51 (0.44–0.58)	0.56 (0.46–0.66)	0.80 (0.49–0.94)	0	0.55 (0.47–0.62)	0.55 (0.46–0.65)	0.84 (0.62–0.94)	0.29 (0.12–0.55)
NPV	CD	0.76 (0.70–0.82)	0.72 (0.65–0.78)	0.96 (0.94–0.98)	0.91 (0.87–0.95)	0.74 (0.66–0.79)	0.72 (0.65–0.79)	0.93 (0.90–0.95)	0.90	0.72 (0.65–0.79)	0.70 (0.62–0.75)	0.94 (0.90–0.96)	0.92 (0.88–0.94)	0.74 (0.67–0.80)	0.66 (0.60–0.72)	0.96 (0.93–0.98)	0.92 (0.89–0.94)
	ES	0.82 (0.74–0.88)	0.66 (0.60–0.72)	0.97 (0.94–0.98)	0.94 (0.90–0.96)	0.83 (0.74–0.90)	0.66 (0.59–0.72)	0.93 (0.89–0.95)	0.92	0.77 (0.68–0.84)	0.65 (0.58–0.71)	0.93 (0.90–0.96)	0.92	0.75 (0.67–0.82)	0.65 (0.58–0.71)	0.96 (0.93–0.97)	0.93 (0.93–0.95)

Hair color categories: **1**: Blond; **2**: Brown; **3**: Red; **4**: Black. **MLR**: multinomial logistic regression; **SVM**: support-vector machine; **RF**: random forest; **ANN**: artificial neural network. **PPV**: Positive predictive value; **NPV**: negative predictive value. **CD**: complete dataset; **ES**: European subset.

(Supplementary Fig. S4). For RF, we needed to tune the number of trees (ntree) and the optimal values for each of the traits tested. We obtained 141 trees for eye color, 713 for hair color, and 589 for skin color for CD, respectively (Supplementary Fig. S2). For the ES we obtained 349 trees for eye color, 319 for hair color and 572 for skin color (Supplementary Fig. S5). Regarding ANN, the parameters that needed to be tuned were the layer size and the regularization parameter of decay for avoiding over-fitting. For the size, we obtained optimal values of 2 for eye color, 6 for hair color, and 3 for skin color for the CD, while for the ES we obtained optimal values of 7 for eye and hair color and 1 for skin color (Supplementary Figs. S3 and S6). For the decay in the CD, the optimal values were equal to 0.5 for hair and skin color, while for eye color it was 0.4 (Supplementary Fig. S3). For the ES we obtained the optimal values for decay equal to 0.5 for eye and hair color and 0.1 for skin color (Supplementary Fig. S6).

3.2. Overall prediction accuracy

As shown in Table 2, in terms of overall accuracy, the four classification methods performed equally well in predicting each of the three considered EVCs. For eye color and the CD, we found that MLR and ANN were able to predict the trait with an overall accuracy of 0.79, while SVM and RF performed almost at the same level with 0.78. Similarly, for the ES the highest performance was obtained with MLR and ANN (0.69), followed by SVM and RF with overall accuracy values of 0.68 and 0.67, respectively. For hair and skin color, the discrepancies among the classifiers were higher compared to eye color for both datasets. More specifically, in the CD the highest overall accuracy for hair color was obtained with MLR (0.60), while SVM and ANN performed almost equally well with accuracies of 0.57 and 0.58, respectively. The RF classifier, however, appeared to have a slightly inferior performance compared to the other classifiers, reaching the lowest overall accuracy of all classifiers at 0.55 for hair color. Similarly, for the ES the MLR had the highest performance of 0.59, followed by ANN and SVM which accuracies were equal to 0.56 and 0.55, respectively. The RF classifier appeared to have a deteriorated performance compared to the other three classifiers. Similar behavior was observed also for skin color prediction in the CD, where the MLR classifier yielded the highest performance with an accuracy of 0.63 compared to the other methods. The SVM classifier yielded an overall accuracy equal to 0.60, while RF and ANN yielded the lowest performances of 0.59 and 0.56, respectively. For the ES both MLR and SVM raised the accuracy to 0.65 for skin color, while the ANN had the lowest accuracy performance of 0.57.

3.3. Predictive measurements

Similar to the results of the overall accuracies, the prediction accuracy measurements for eye color presented very little to no differences between the four methods regarding blue and brown eye color, while a few deviations between the methods were seen for intermediate eye color (Table 3). For example, the sensitivity of the intermediate eye color prediction for the CD equaled 0.20 for ANN but dropped to 0.18, 0.13 and 0.15 for MLR, SVM and RF, respectively. Another example is the PPV of the intermediate eye color prediction, which obtained its highest value of 0.63 for SVM, while it dropped to 0.58 for MLR. For the ES the PPV value of intermediate eye color was raised to 0.59 for ANN while for RF it dropped to 0.42. The confusion matrices for eye color showed, for both CD and ES, small deviations among the four classifiers. Blue and brown eye colors appeared to be better predicted by the model in comparison with the intermediate eye color (Supplementary Tables S2 and S3). AUC values were at similar levels, especially for SVM, RF and ANN, while MLR slightly outperformed (Supplementary Tables S4 and S5).

For *hair color*, we also observed rather similar prediction performances for all four methods, although more pronounced differences were seen for some trait categories (Table 4) compared to eye color (Table 3). In particular, the sensitivity of Red hair color prediction in the CD reached its highest value with MLR (0.66), followed by ANN (0.58), while its value was almost halved to 0.28 for RF, and for SVM it reached 0.21 (Table 4). The sensitivity of Black hair color prediction completely dropped to zero for SVM, while its highest value was equal to 0.31 for ANN. Another example was the PPV for Black hair color, where we obtained the highest values with MLR and RF (0.58 and 0.47, respectively), while it dropped to 0.34 for ANN. We observed a similar behavior to the CD in the ES for the sensitivity of red hair color prediction where its highest values were yielded by MLR and ANN (0.69 and 0.62, respectively), while for RF and SVM the value was halved to 0.31 and 0.23, respectively. Sensitivity of black hair color dropped to zero for SVM and RF, while its highest value was obtained with MLR (0.26). PPV for black hair color reached its highest value with MLR, while it dropped to zero for RF. The confusion matrices for hair color showed similar patterns for CD and ES where the categories with fewer samples in the datasets, such as red and black hair color categories, showed higher deviations compared to blond and brown hair color (Supplementary Tables S6 and S7). AUC values for MLR outperformed for most category comparisons compared to the other ML classifiers (Supplementary Tables S2 and S3).

For *skin color*, as with hair color, we also observed uneven differences between classifiers for some predictive measurements and trait categories (Table 5). For example, in the complete dataset the sensitivity of the Very Pale skin color category prediction was 0.11 for both MLR and SVM but zero when RF and ANN were applied. Similar diminution was also observed for the sensitivity and the PPV of RF in predicting Dark skin color. RF was the only classification method where these values equaled zero (Table 5). Higher discrepancies were also observed for the specificity of pale skin color where its highest values were obtained for both MLR and RF (0.60); with SVM was applied the value dropped to 0.40. Sensitivity of dark to black category dropped to 0.66 for ANN, while for SVM and RF it reached the highest value of 0.96. In the ES, the sensitivity of very pale skin color reached the highest value of 0.25 with MLR, while for the rest of the classifiers it was almost equal to zero. The specificity of pale skin color yielded its highest value of 0.65 with MLR

Table 5
Predictive measurements for skin color for the four classifiers.

		MLR					SVM			
Category		1	2	3	4	5	1	2	3	4
Sensitivity	CD	0.11 (0.02–0.44)	0.76 (0.68–0.83)	0.47 (0.38–0.57)	0.75 (0.30–0.95)	0.88 (0.69–0.96)	0.11	0.83 (0.75–0.89)	0.31 (0.23–0.41)	0.25 (0.05–0.70)
	ES	0.25 (0.09–0.53)	0.70 (0.61–0.78)	0.65 (0.54–0.75)			0	0.76 (0.68–0.83)	0.58 (0.47–0.69)	
Specificity	CD	0.99 (0.97–0.99)	0.60 (0.52–0.68)	0.80 (0.73–0.86)	0.98 (0.96–0.99)	1.00 (0.98–1.00)	0.99	0.46 (0.38–0.54)	0.85 (0.78–0.89)	0.98 (0.96–0.99)
	ES	0.97 (0.93–0.98)	0.65 (0.55–0.74)	0.74 (0.65–0.81)			1	0.56 (0.45–0.66)	0.80 (0.73–0.85)	
PPV	CD	0.25 (0.05–0.70)	0.61 (0.53–0.69)	0.62 (0.51–0.72)	0.38 (0.14–0.69)	1.00 (0.85–1.00)	0.25	0.56 (0.48–0.63)	0.59 (0.46–0.70)	0.25 (0.05–0.70)
	ES	0.33 (0.12–0.65)	0.72 (0.63–0.80)	0.60 (0.49–0.70)			NA	0.69 (0.60–0.76)	0.58 (0.47–0.69)	
NPV	CD	0.97 (0.94–0.98)	0.76 (0.67–0.83)	0.69 (0.62–0.75)	0.99 (0.97–0.99)	0.99 (0.96–0.99)	0.96	0.76 (0.67–0.84)	0.64 (0.57–0.70)	0.99 (0.97–0.99)
	ES	0.95 (0.91–0.97)	0.63 (0.53–0.72)	0.78 (0.69–0.84)			0.94	0.65 (0.54–0.75)	0.80 (0.73–0.85)	

Skin color categories: **1**: Very pale; **2**: Pale; **3**: Intermediate; **4**: Dark; **5**: Dark to Black. **MLR**: multinomial logistic regression; **SVM**: support-vector machine; **RF**: random forest; **ANN**: artificial neural network. **PPV**: Positive predictive value; **NPV**: negative predictive value. **CD**: complete dataset; **ES**: European subset.

SVM	RF					ANN				
5	1	2	3	4	5	1	2	3	4	5
0.96 (0.80–0.99)	0.00 0	0.76 (0.68–0.83) 0.81 (0.73–0.87)	0.19 (0.12–0.27) 0.54 (0.43–0.65)	0.00	0.96 (0.80–0.99)	0.00 0.08 (0.01–0.35)	0.61 (0.52–0.70) 0.68 (0.59–0.76)	0.50 (0.42–0.60) 0.49 (0.38–0.60)	0.50 (0.15–0.85)	0.66 (0.47–0.82)
0.99 (0.97–0.99)	1.00 1	0.60 (0.52–0.68) 0.49 (0.39–0.59)	0.92 (0.87–0.96) 0.81 (0.73–0.87)	0.99	0.99 (0.96–0.99)	0.99 0.97 (0.94–0.99)	0.58 (0.50–0.66) 0.50 (0.40–0.60)	0.65 (0.58–0.72) 0.70 (0.62–0.78)	0.99 (0.97–0.99)	0.99 (0.97–0.99)
0,95 (0.80–0.99)	NA NA	0.61 (0.53,0.69) 0.67 (0.59–0.74)	0.63 (0.45–0.77) 0.63 (0.51–0.74)	0.00	0.88 (0.71–0.96)	0.00 0.17 (0.03–0.56)	0.55 (0.46–0.63) 0.64 (0.55–0.72)	0.50 (0.41–0.60) 0.50 (0.39–0.61)	0.40 (0.12–0.77)	0.94 (0.73–0.99)
0.99 (0.97–0.99)	0.97 0.94	0.76 (0.67–0.83) 0.67 (0.54–0.77)	0.62 (0.56–0.68) 0.74 (0.66–0.81)	0.98	0.99 (0.97–0.99)	0.97 0.94 (0.90–0.97)	0.65 (0.56–0.73) 0.55 (0.44–0.66)	0.66 (0.58–0.73) 0.69 (0.61–0.77)	0.99 (0.97–0.99)	0.97 (0.94–0.98)

but dropped to 0.40 for RF. For most of the other skin color categories and predictive measurements, the four classification methods performed almost equally (Table 5). In the confusion matrices for skin color, the categories with the highest number of samples, namely Pale and Intermediate categories, were better predicted in comparison to the other categories (Supplementary Tables S8 and S9). Also and similar to eye and hair color prediction, the AUC values for MLR mostly outperformed the other classifiers (Supplementary Tables S2 and S3).

4. Discussion

In the present study, we compared four different ML classification methods, namely MLR, as widely used for EVC prediction from DNA in general, and pigmentation prediction in particular, in addition to SVM, RF and ANN with respect to their ability to predict various eye, hair and skin color categories based on the previously established IrisPlex, HIrisPlex, and HIrisPlex-S DNA markers. Since these ML methods have been barely applied for EVC prediction so far and are well-known for their often very good prediction performance in other application fields, the basic motivation for this study was to investigate and to identify, for each of the tested EVCs, the optimal classifier yielding the highest performance and assess whether any of them outperforms the standard MLR approach. In order to obtain the maximum performance of the SVM, RF and ANN methods, we first needed to perform hyperparameter tuning. Parameters such as cost and gamma for SVM, ntree for RF and size and decay for ANN were tuned and their optimal values were chosen according to the lowest OOB error (Supplementary Figs. S1–S6).

Our results showed that when it comes to overall accuracy, all four classifiers performed almost equally well for all pigmentation traits tested, with almost no variation across the classifiers for eye color and slight variation for hair and skin color. Thus, none of the other ML methods outperformed the conventional method of MLR in predicting eye, hair and skin color based on the IrisPlex, HIrisPlex, and HIrisPlex-S DNA markers, respectively. When looking at the full suite of prediction measurements per each of the three pigmentation traits, we noted slight differences between some classifiers for several trait categories, somewhat more for hair and skin color than for eye color. However, these differences do not allow a conclusion that any of the three ML classifiers perform superior over MLR, which is supported by our conclusion derived from the overall accuracy results. This pattern was also observed when we compared the prediction performances between the two datasets, CD and ES, where highest deviations were observed for hair and skin color compared to eye color. This was to be expected since European samples represent the major part of the CD, implying that our model was trained mostly on European samples and therefore, when we compare the performance of the CD-derived model with the one trained on the ES, we do not expect to see high differences in the overall performance.

For eye color and for both datasets, we saw a small but noticeable deviation between the four classification methods for the intermediate eye color category, while for blue and brown eye color categories, all four methods performed almost identically. As obtained with all four methods, prediction accuracies were high for blue and brown eye color, but low for intermediate eye color. This finding is in line with previous results obtained mostly based on MLR [13,17,19–21,24]. As emphasized in all previous IrisPlex publications [17,19,20], the six IrisPlex DNA markers used here are very suitable for predicting blue and brown eye color, while their ability to predict non-blue and non-brown eye colors, which are all grouped into the intermediate eye color category, is limited. Currently, it is proposed that the limitation to predict intermediate eye color with all four classification methods is more likely explained by missing DNA predictors as opposed to the modeling type. Similarly, it may be caused by phenotype definition, as the intermediate eye color category can be expected to be more heterogeneous than the blue and brown eye colour categories that both reflect the two extremes of the eye colour phenotype distribution. A large-scale genome-wide association study (GWAS) on eye color is currently underway, aiming to increase the number of independently eye color associated DNA variants. Thus, their future use in prediction modelling of categorical eye color will help ascertain if it is the number of DNA predictors that underlies the currently limited prediction accuracy of intermediate eye color, which based on our current findings appears to be independent of the classification method used.

Regarding hair color, the prediction performances among the four classifiers were also quite similar for the two datasets; however, the deviations were higher compared to eye color, while skin color was the trait with the highest deviations among the model measurements for some categories. This could possibly be explained by the fact that these traits and especially hair and skin color are adaptive traits that can be affected by some external or environmental factors that are not included in the genetic prediction models and consequently can affect the prediction outcomes of the different methods at various extents. In other words, each classification method has probably a different level of sensitivity in detecting such external factors, which possibly leads to higher deviations between the results. Another explanation could be the much larger number of predictors included in the hair and skin color model compared to the few markers in the eye color model, giving the ML models more freedom to pick up local patterns in the parameter space, although such patterns may represent random events that deteriorate the performance of such approaches.

The non-substantial differences obtained in the overall accuracies of the four classifiers could be explained by the fact that we only look at the additive effects of the genetic markers and not at potential interaction effects. This may be due to the underlying genetic mechanisms but may be equally well explained by the way those genetic markers included in the established MLR models were identified in the first place. The latter has been usually done in GWASs, which mostly focus on additive independent marker contributions to the traits. Possible incorporation of interactive effects could add some additional information that might affect the prediction performances of each classifier and probably distinguish some prediction methods that are more sensitive to the addition of interactive effects. Previous studies have already identified and incorporated SNP-SNP interactions in MLR-based modelling for eye color prediction [18,61]. However, the previously noted predictive effects of SNP-SNP interactions were small, maybe because of the use of MLR, which requires active intervention by the analyst to consider

two-way or higher-order interaction effects, whereas other ML methods often do this automatically. In our case, since with the currently available DNA predictors the interaction effects were small and no substantial differences were obtained among the four classifiers, we would not recommend interaction effects at this stage. Future ML-based pigmentation prediction studies using elongated lists of DNA predictors that already are available from large-scale GWASs for hair [62] and skin color [63] and will soon be for eye color shall consider these interaction effects which might improve the overall prediction performance.

Another possible explanation for the non-substantial differences between the four classification methods could be the data sizes used for each trait and the number of samples for each trait category. Since ML methods are computational methods that 'learn' directly from the data, the amount of the datasets used for model training can affect the model performance. When increasing the datasets, more information regarding the patterns of each group is incorporated into the model and therefore allows the observations to be separated into the different classes more accurately. This is due to them being based on data patterns and not on weak correlations that can occur in small datasets. Thus, we could expect that this may have affected, to some extent, the prediction performances of the methods applied due to the use of these currently available datasets that may not represent all combination patterns of alleles. This can be confirmed to some extent by our case where we noticed that prediction performance was higher when using the complete dataset in comparison with the European subset which appears to have a slightly deteriorated performance, especially for eye and hair color prediction. Larger datasets in general are often necessary and interesting to be considered for future pigmentation prediction studies, in order to release the full potential of these differing ML approaches.

In addition to the above, another possible approach for future studies would be the combined analysis and prediction of visible traits in order to see whether one could gain additional information that helps improving the current prediction accuracies. While this is out of scope of the current study, future investigation on the topic would be worthwhile in order to assess possible benefits of such an approach. A recent study by Chen et al. [64] focused on the impact of correlations of pigmentation phenotypes on the genetic EVC prediction. This study provided valuable insights; however, it highlights the importance of further research that might help in the improvement of the current prediction accuracies.

In summary, our results did not show substantial differences between the four ML-based methods tested to predict appearance prediction, in particular eye, hair, and skin color using the previously established IrisPlex, HIrisPlex, and HIrisPlex-S DNA markers, respectively. Given this outcome and because of the easier interpretation and often substantially lower computational costs of MLR with respect to the modelled function compared to other ML approaches, we suggest, at least for now, the use of the MLR as the most appropriate method for predicting appearance traits from DNA, especially with regards the three pigmentation traits used here. MLR describes a simple relationship between the inputs and the outputs, which makes the outcomes of the predictions more interpretable compared to ML methods. Contributions and feature interactions can also be easily represented by the coefficients in the MLR but require active pursuit of such interactions by the analyst, while the inner workings of SVM, RF and ANN are harder to understand and interpret, although they do offer more automated consideration of interaction terms. The latter three ML methods also do not provide a direct estimate of the importance of each feature for the model's prediction performance, although secondary, resampling-based approaches exist that may provide such an assessment. Thus, for ML methods it is harder to understand the interaction between the different features in the model.

Notably, our findings and conclusions obtained are based on a relatively small number of established DNA predictors and we did not consider interactions between them. In general, ML approaches are expected to show their full potential when larger sets of genetic markers are included in the model since they will likely seize better the patterns of the data that subsequently could lead to better prediction performance. Therefore, once more appearance DNA predictors and interactions between them have been established, it would be interesting to use them in a classifier method comparison as performed here, to find out, if the results we obtained here may have been affected by the type and number of DNA markers used, or the classification of the phenotype being predicted. However, for the time being, and with the established pigmentation DNA predictors currently available, MLR remains the preferred classification method of choice for predicting categorical pigmentation traits from DNA.

Funding

This study received support from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 740580 within the framework of the *Visible Attributes through Genomics* (VISAGE) Project and Consortium. The IUPUI US site was supported in part by the US National Institute of Justice (NIJ) under grant number 2014-DN-BX-K031 and 2018-DU-BX-0219. None of the funding organizations had any influence on the design, conduct, or conclusions of the study.

Conflict of interest

The authors declare that they have no competing interests.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2021.102507.

Appendix B. Centres and investigators of the VISible Attributes through GEnomics (VISAGE) Consortium

Website: http://www.visage-h2020.eu/

- **Erasmus University Medical Center Rotterdam (Netherlands):** Manfred Kayser, Vivian Kalamara, Arwin Ralf, Athina Vidaki
- **Jagiellonian University (Poland):** Wojciech Branicki, Ewelina Pośpiech, Aleksandra Pisarek
- **Universidade de Santiago de Compostela (Spain) :** Ángel Carracedo, Maria Victoria Lareu, Christopher Phillips, Ana Freire-Aradas, Ana Mosquera-Miguel, María de la Puente
- **Medizinische Universität Innsbruck (Austria)**: Walther Parson, Catarina Xavier, Antonia Heidegger, Harald Niederstätter
- **Universität zu Köln (Germany):** Michael Nothnagel, Maria-Alexandra Katsara, Tarek Khellaf
- **King's College London (United Kingdom):** Barbara Prainsack, Gabrielle Samuel
- **Klinikum der Universität zu Köln (Germany):** Peter M. Schneider, Theresa E. Gross, Jan Fleckhaus
- **Bundeskriminalamt (Germany):** Ingo Bastisch, Nathalie Schury, Jens Teodoridis, Martina Unterländer
- **Institut National De Police Scientifique (France):** François-Xavier Laurent, Caroline Bouakaze, Yann Chantrel, Anna Delest, Clémence Hollard, Ayhan Ulus, Julien Vannier
- **Netherlands Forensic Institute (Netherlands):** Titia Sijen, Kris van der Gaag, Marina Ventayol-Garcia
- **National Forensic Centre, Swedish Police Authority (Sweden):** Johannes Hedman, Klara Junker, Maja Sidstedt
- **Metropolitan Police Service, London (United Kingdom):** Shazia Khan, Carole E. Ames, Andrew Revoir
- **Centralne Laboratorium Kryminalistyczne Policji (Poland):** Magdalena Spólnicka, Ewa Kartasińska, Anna Woźniak

References

[1] M. Kayser, P. de Knijf, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (3) (2011) 179–192.
[2] M. Kayser, Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes, Forensic Sci. Int. Genet. 18 (2015) 33–48.
[3] M. Kayser, P.M. Schneider, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations, Forensic Sci. Int. Genet. 3 (3) (2009) 154–161.
[4] F. Liu, et al., Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up, Hum. Genet. 134 (8) (2015) 823–835.
[5] S.I. Candille, et al., Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four European populations, PLoS One 7 (2012) 10.
[6] M.R. Gerstenblith, J. Shi, M.T. Landi, Genome-wide association studies of pigmentation and skin cancer: a review and meta-analysis, Pigment Cell Melanoma Res. 23 (5) (2010) 587–606.
[7] P. Sulem, et al., Two newly identified genetic determinants of pigmentation in Europeans, Nat. Genet. 40 (2008) 835–837.
[8] P. Sulem, et al., Genetic determinants of hair, eye and skin pigmentation in Europeans, Nat. Genet. 39 (2007) 1443–1452.
[9] J. Han, et al., A genome-wide association study identifies novel alleles associated with hair color and skin, PLoS Genet. 4 (2008) 5.
[10] L. Rawofi, et al., Genome-wide association study of pigmentary traits (skin and iris color) in individuals of East Asian Ancestry, PeerJ 2 (2017) 5.
[11] R.P. Stokowski, et al., A genomewide association study of skin pigmentation in a South Asian population, Am. J. Hum. Genet. 81 (6) (2007) 1119–1132.
[12] J. Alghamadi, et al., Eye color prediction using single nucleotide polymorphisms in Saudi population, Saudi J. Biol. Sci. 26 (7) (2019) 1607–1612.
[13] F. Liu, et al., Eye color and the prediction of complex phenotypes from genotypes, Curr. Biol. 19 (5) (2009) R192–R193.
[14] Y. Ruiz, et al., Further development of forensic eye color predictive tests, Forensic Sci. Int. Genet. 7 (1) (2013) 28–40.
[15] W. Branicki, et al., Model-based prediction of human hair color using DNA variants, Hum. Genet. 129 (4) (2011) 443–454.
[16] S. Walsh, et al., Global skin colour prediction from DNA, Hum. Genet. 136 (7) (2017) 847–863.
[17] S. Walsh, et al., IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, Forensic Sci. Int. Genet. 5 (3) (2010) 170–180.
[18] E. Pospiech, et al., The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction, Forensic Sci. Int. Genet. 11 (2014) 64–72.
[19] L. Chaitanya, et al., The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: introduction and forensic developmental validation, Forensic Sci. Int. Genet. 35 (2018) 123–135.
[20] S. Walsh, et al., The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, Forensic Sci. Int. Genet. 7 (1) (2013) 98–115.
[21] S. Walsh, et al., DNA-based eye colour prediction across Europe with the IrisPlex system, Forensic Sci. Int. Genet. 6 (3) (2012) 330–340.
[22] O. Maronas, et al., Development of a forensic skin colour predictive test, Forensic Sci. Int. Genet. (2014).
[23] J. Söchtig, et al., Exploration of SNP variants affecting hair colour prediction in Europeans, Int. J. Leg. Med. 129 (5) (2015) 963–975.
[24] S. A.J, S. Harbison, SNP model development for the prediction of eye colour in New Zealand, Forensic Sci. Int. Genet. 7 (4) (2013) 444–452.
[25] M.A. Katsara, M. Nothnagel, True colors: a literature review on the spatial distribution of eye and hair pigmentation, Forensic Sci. Int. Genet. 39 (2019) 109–118.
[26] E. Alpaydin, Introduction to Machine Learning, MIT Press, 2004.
[27] S.B. Kotsiantis, Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades, Artif. Intell. Rev. 37 (2012) 331–344.
[28] J.A.M. Sidey-Gibbons, C.J. Sidey-Gibbons, Machine learning in medicine: a practical introduction, BMC Med. Res. Methodol. (2019).
[29] J. Kreuziger, Application of machine learning to robotics - an analysis. In Proceedings of the Second International Conference on Automation, Robotics, and Computer Vision (ICARCV '92), (1992).
[30] D.W. Hosmer, S. Lemeshow. Applied Logistic Regression, Second ed., John Wileys & sons, Inc, Canada, 2000.
[31] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[32] A. Mucherino, P.J. Papajorgji, P.M. Pardalos, k-nearest neighbor classification. Data Mining in Agriculture, Springer, New York, NY, 2009.
[33] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, Berlin, Heidelberg, 1995.
[34] B.D. Ripley, Neural networks and related methods for classification, J. R. Stat. Soc. Ser. B (Methodol.) 56 (3) (1994) 409–456.
[35] S. Goswami, E.J. Wegman, Comparison of different classification methods on glass identification for forensic research, J. Stat. Sci. Appl. 4 (2016) 65–84.
[36] G. Zadora, Glass analysis for forensic purposes—a comparison of classification methods, J. Chemom. 54 (1) (2007) 49–59.
[37] E.Y.Y. Cheung, M.E. Gahan, D. McNevin, Prediction of biogeographical ancestry from genotype: a comparison of classifiers, Int. J. Leg. Med. 131 (4) (2017) 901–912.
[38] K. Karampidis, E. Kavallieratou, G. Papadourakis, Comparison of classification algorithms for file type detection a digital forensics perspective, POLIBITS 56 (2017) 15–20.
[39] O.M. Hurtado, et al., Comparing machine learning classifiers and linear/logistic regression to explore the relationship between hand dimensions and demographic characteristics, PLoS One 11 (2016) 11.
[40] T.T. Toma, J.M. Dawson, D.A. Adjeroh, Human ancestry indentification under resource constraints – what can one chromosome tell us about human biogeographical ancestry? BMC Med. Genom. 11 (2018) 5.
[41] R.C. Team, R: a language and environment for statistical computing, R Found. Stat. Comput. (2017).
[42] R. Team, RStudio: integrated development environment for R, (2016). Available from: ⟨http://www.rstudio.com/⟩.
[43] W.N. Venables, B.D. Ripley. Modern Applied Statistics with S, Fourth ed., Springer, New York, 2002.
[44] M. Kuhn, Caret: classification and regression training, (2020).
[45] D. Meyer, et al., e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, (2019).
[46] A. Liaw, M. Wiener, Classification and regression by random forest, R News 2 (3) (2002) 18–22.
[47] V. Kecman. Support Vector Machines – An Introduction, Springer, Berlin, Heidelberg, 2005.
[48] P. Latinne, D. Olivier, C. Decaestecker, Limiting the number of trees in random forests, Lect. Notes Comput. Sci. 2096 (2001) 178–187.
[49] D. Hernandez-Lobato, G. Martinez-Munoz, A. Suarez, How large should ensembles of classifiers be? Pattern Recognit. 46 (5) (2013) 1323–1336.
[50] T.M. Oshiro, P.S. Perez, J.A. Baranauskas, How many trees in a random forest? Lect. Notes Comput. Sci. (2012) 154–168.
[51] G.G. Daniel, in: A.L.C. Runehov (Ed.), Artificial Neural Network, Springer, Dordrecht, 2013.
[52] W. McCulloch, W. Pitts, A logical calculus of ideas immanent in nervous activity, Bull. Math. Biophys. 5 (4) (1943) 115–133.
[53] D.O. Hebb. The Organization of Behavior, Wiley, New York, 1949, p. 437.
[54] B. Farley, W. Clark, Simulation of self-organizing systems by digital computer, Trans. IRE Prof. Group Inf. Theory 4 (4) (1954) 76–84.
[55] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (6) (1958) 386–408.
[56] P.J. Werbos, Beyond regression: new tools for prediction and analysis in the behavioral sciences, (1975).
[57] J. Schmidhuber, Learning complex, extended sequences using the principle of history compression, Neural Comput. 4 (1992) 234–242.
[58] D. Scherer, A.C. Müller, S. Behnke, Evaluation of pooling operations in convolutional architectures for object recognition, in: Proceedings of the 20th International Conference Artificial Neural Networks (ICANN), (2010) p. 92–101.
[59] A.Y. Ng, et al., Building high-level features using large scale unsupervised learning, (2012).
[60] D. Kriesel, A brief introduction to neural networks, (2007) p. 286. Available at ⟨http://www.dkriesel.com⟩.
[61] E. Pospiech, et al., Gene–gene interactions contribute to eye colour variation in humans, J. Hum. Genet. 56 (2011) 447–455.
[62] P.G. Hysi, et al., Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability, Nat. Genet. 50 (2018) 652–656.
[63] A. Visconti, et al., Genome-wide association study in 176,678 Europeans reveals genetic loci for tanning response to sun exposure, Nat. Commun. 9 (1) (2018) 1684.
[64] Y. Chen, et al., The impact of correlations between pigmentation phenotypes and underlying genotypes on genetic prediction of pigmentation traits, Forensic Sci. Int. Genet. 50 (2021), 102395.