Predictive Modeling of Biogeographical Ancestry using a novel SNP panel and Supervised Learning approaches

Cosimo Grazzini[a,b], Giorgia Spera[a,b], Stefania Morelli[c], Daniele Castellana[a,b], Giulia Cosenza[c], Michela Baccini[a,b], Giulia Cereda[a,b], Elena Pilli[c]

[a]Department of Statistics, Computer Science, Applications "Giuseppe Parenti" (DiSIA), University of Florence, Viale Giovanni Battista Morgagni 59/65, Florence, 50134, Italy.
[b]Florence Center for Data Science, University of Florence, Viale Giovanni Battista Morgagni 59, Florence, 50134, Italy.
[c]IRIS (Infrastruttura per la Ricerca e l'identificazione degli Scheletri senza nome) Department of Biology, University of Florence, Via del Proconsolo, 12 Florence, 50122, Italy.

Authors' emails
Cosimo Grazzini cosimo.grazzini@unifi.it
Giorgia Spera giorgia.spera@unifi.it
Stefania Morelli stefania.morelli@unifi.it
Daniele Castellana daniele.castellana@unifi.it
Giulia Cosenza giulia.cosenza1303@gmail.com
Michela Baccini michela.baccini@unifi.it
Giulia Cereda giulia.cereda@unifi.it
Elena Pilli elena.pilli@unifi.it

Corresponding author
Giulia Cereda giulia.cereda@unifi.it

Abstract

Inferring an individual’s BioGeographical Ancestry (BGA) through DNA analysis is a valuable tool in various fields such as forensic science, especially when traditional methods fail to identify suspects or victims. Advances in Next-Generation Sequencing (NGS) have revolutionized genomic data acquisition, enabling the development of comprehensive Single Nucleotide Polymorphism (SNP) panels for ancestry inference. This study assessed the effectiveness of a novel panel containing 3,234 SNPs at both inter-continental and a more detailed BGA level, using various supervised Machine Learning (ML) models, including Categorical Naive Bayes, Penalized Multinomial Logistic Regression, Linear Support Vector Machines, Random Forest, and tree-based Gradient Boosting. A nested cross-validation approach was employed for model tuning and evaluation, with balanced accuracy as the main performance metric to address class imbalance. At the inter-continental level, all ML models demonstrated high balanced accuracy, confirming their reliability for BGA inference. However, performance declined at the more detailed continental level, likely due to a combination of factors including increased class imbalance, reduced sample sizes for certain populations, and the inherent complexity of distinguishing genetically and geographically proximate groups. Nonetheless, promising results were observed for South Asians, Northeast Asians, Europeans, and West Africans classes. In contrast, performance was notably lower for underrepresented classes such as Inner Asians. Misclassification patterns at both levels appeared to reflect known geographical and historical relationships, although further analysis revealed that these were often concentrated in underrepresented or genetically complex groups. These findings highlighted the potential of this SNP panel and ML approaches as valuable tools for forensic investigations.

KEYWORDS: Biogeographical ancestry; SNP panel; NGS technology; Forensic samples; Supervised ML models; Class imbalance.

1. Introduction

BGA refers to the ethnic background of a trace or an individual/skeleton, encompassing both biological and cultural components. The ability to infer individual BGA from DNA analysis has become increasingly important across various fields, including population studies, medicine, epidemiology, and forensic science (Bergström et al., 2020). In forensic contexts, when Short Tandem Repeat (STR) typing fails to identify perpetrators or victims of unsolved cases —due to limited investigative leads or the absence of reference profiles— BGA inference can provide crucial additional genetic information, offering investigative clues, narrowing the suspect pool, and guiding the direction of the investigation (Jin et al., 2020). The advancements in DNA sequencing technology, known collectively as NGS or massively parallel sequencing (MPS), have led to a significant increase in available genomic data by enabling the collection of vast amounts of information previously unattainable using traditional molecular biology methods. This technology has also found increasing applications in forensics, and tools for BGA analysis have been developed thanks to the ability to multiplex a greater number of markers. SNPs are particularly suitable for this purpose due to their stability, broad genomic distribution, and population-specific allele frequencies. While existing SNP panels have demonstrated good performance in distinguishing individuals at the continental level (Al-Asfi et al., 2018; Alladio et al., 2022; Bulbul & Filoglu, 2018; Eduardoff et al., 2016; Guo et al., 2020; Jäger et al., 2017; Jia et al., 2014; Jiang et al., 2018; Kersbergen et al., 2009; Phillips et al., 2019, 2014; Rogalla et al., 2015; Setser et al., 2020), classification at a more detailed BGA level remains challenging. This is often due to the limited number of markers and the difficulty in distinguishing genetically similar populations. Moreover, the scarcity of samples from certain populations contributes to poorly characterized reference databases, leading to class imbalance issues (He & Garcia, 2009). Therefore, the possibility of considerably increasing the number of markers investigated simultaneously is leading the forensic expert to create innovative panels with a larger number of SNPs, which, however, require the adoption of appropriate computer tools and/or algorithms for their processing and interpretation. ML methods are particularly well-suited for this task, as they can uncover complex patterns in high-dimensional data and build predictive models for BGA inference (Kloska et al., 2023; Zou et al., 2019). In recent years, ML has made significant progress across various disciplines, and although still in its infancy, its integration into forensic science is on the rise (Barash et al., 2024; Qu et al., 2019). Attempts to develop ML bioinformatics tools to infer BGA at a more detailed level in geographically limited regions have yielded promising results (Allocco et al., 2007; J.-Q. Gu et al., 2022; Guillot et al., 2016; Hwa et al., 2019; Kloska et al., 2023). Few studies have attempted to infer BGA on a global scale using ML approaches. Two main ML-based strategies have been proposed for global BGA inference: 1. a two-stage approach, where inter-continental classification is followed by detailed inference within each predicted inter-continental class (Alladio et al., 2022; Hajiloo et al., 2013), and 2. a single-step approach that directly predicts the detailed ancestry without relying on intermediate hierarchical steps (Battey et al., 2020; Toma et al., 2018). Pilli et al. (Pilli et al., 2023) recently obtained interesting results by applying powerful multivariate techniques such as Partial Least Squares Discriminant Analysis (PLS-DA) and variable selection techniques (Backward Variable Elimination, Genetic Algorithm, and Regularized Elimination Procedure), to infer BGA and select the best SNPs at both continental and a more detailed BGA levels, outperforming traditional approaches like PCA and STRUCTURE.

The contribution of this study is twofold. First, an innovative BGA panel (Pilli et al., 2023) was evaluated using a wide range of supervised ML models, selected for their suitability in handling high-dimensional categorical data. The analysis was conducted at both inter-continental and detailed-continental levels (as described in the Materials and Methods section), adopting a single-step classification strategy on a global scale. This approach was chosen to explore the behavior of common ML models in a complex, multi-class setting and to assess their performance without relying on intermediate classification stages. Although hierarchical classification can leverage structured relationships between broader and more specific categories, it also introduces risks such as error propagation, class-membership inconsistencies, and increased model complexity —especially in the absence of re-weighting or retraining mechanisms (Rezende et al., 2022; Silla & Freitas, 2011). Moreover, Toma et al. (2018) reported comparable results between hierarchical and single-step strategies, supporting the feasibility of the latter. Second, the performance of classical ML models — both base learners (e.g., Categorical Naive Bayes, Penalized Multinomial Logistic Regression, and Linear Support Vector Machines) and ensemble methods (e.g., Random Forest and Gradient Boosting)— was assessed in the presence of strong class imbalance and missing genotype data. The evaluation followed a rigorous nested cross-validation strategy, using balanced accuracy as the primary metric, and included an in-depth analysis of confusion matrices to identify systematic misclassification patterns.

2. Materials and Methods

2.1 SNP Panel and Reference Population Dataset

The Ancestry Informative SNPs (AISNPs) included in this study were selected following the methodology proposed by Pilli et al. (Pilli et al., 2023), resulting in a final panel of 3,234 markers. SNP filtering from the Variant Call Format (VCF) files was performed using BCFtools (http://github.com/samtools/bcftools) (Danecek et al., 2021). Each individual's genotype was encoded as a pair of alleles for each locus in the panel.

Reference population data were collected from three major population projects: the 1,000 Genomes Project (N=2,504) (Auton et al., 2015), the Simons Genome Diversity Project (SGDP, N=279) (https://www.simonsfoundation.org/simons-genome-diversity-project/), and the Human Genome Diversity Project (HGDP, N=929) (Bergström et al., 2020). After removing overlapping data and adding 34 new individuals from other studies (Lorente-Galdos et al., 2019; Serra-Vidal et al., 2019), the final dataset comprised 3,591 individuals from 117 populations. These were grouped into 19 macro-populations based on geographic and literature-based criteria (Bryc et al., 2010; Henn et al., 2012; Hodgson et al., 2014; Mulindwa et al., 2020; Pagani et al., 2012; Resutik et al., 2023; Schlebusch et al., 2012) (Figure 1). Due to the absence of individuals in the South-East Africans group, this class was excluded. The resulting labels were denoted as "detailed-continental" BGA labels, whereas a set of broader labels, more commonly used in literature, was referred to as "inter-continental" BGA labels. The distributions of individuals at inter- and detailed-continental BGA levels in the dataset are reported in Table 1 and Figure 1, respectively, highlighting a notable class imbalance.

Table 1 Distribution of individuals per inter-continental BGA class.

Inter-continental label	**Count**
Africa	837

South Asia	693
East Asia	692
Europe	663
America	420
Middle East	172
North Asia	45
Oceania	35
Central Asia	34

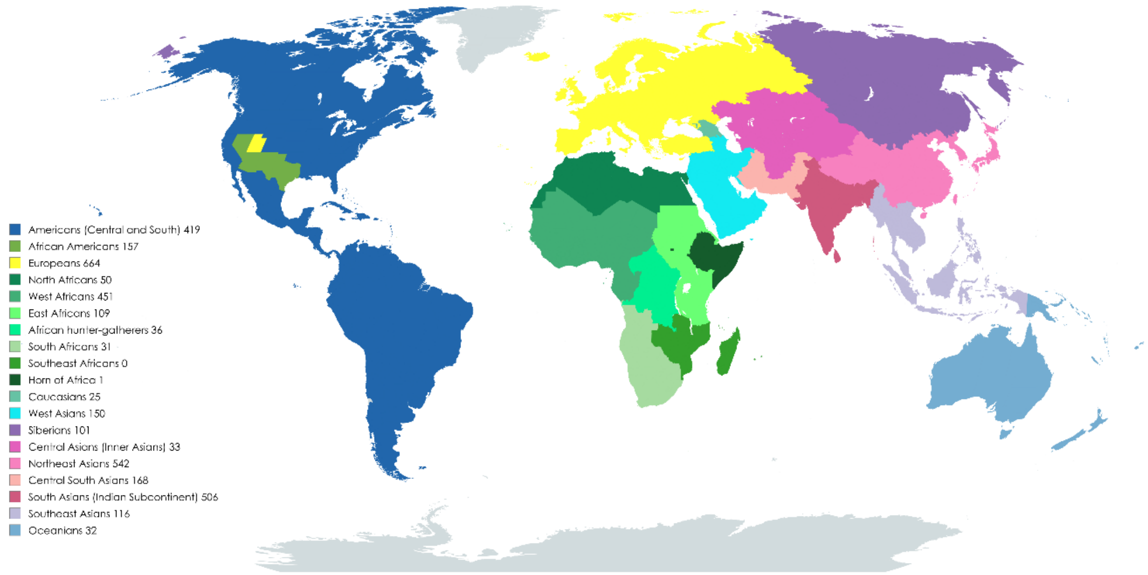

Fig. 1. Approximate geographic distribution of the 19 macro-populations used in this study, defined according to detailed-continental BGA labels. Colored regions represent genetically inferred ancestry groups and do not necessarily reflect current geographic residence. The number next to each BGA label indicates the number of individuals assigned to that group.

2.2 Pre-Processing, Supervised Machine Learning Models, and Validation Scheme

To evaluate the discriminative power of the SNP panel and the behavior of classical ML models in the context of imbalanced data, three base learner methods —Categorical Naive Bayes (NB), penalized multinomial Logistic Regression (LR), and Linear Support Vector Machines (SVMs)— alongside two ensemble methods —Random Forest (RF) and Gradient Boosting (GB) — were tested. Additionally, a balanced version of Random Forest (bRF) optimized for class imbalance was implemented. A detailed explanation of the adopted machine learning methods and their implementation is available in Supplementary Materials (File S1).

The adopted pipeline consisted of a preprocessing phase followed by the implementation of selected classification algorithms. During pre-processing, missing genotype values were imputed by replacing them with the most frequent genotype observed for each marker. The categorical genotype data were then transformed into a binary matrix using one-hot encoding, where each unique genotype category observed for a SNP was represented by a vector of 0s and 1s, indicating the absence or presence of that category. This encoding strategy enables ML models to effectively process categorical genotype information. To address class imbalance, an oversampling procedure was applied following the approach described by Menardi & Torelli (Menardi & Torelli, 2014) to prevent poor performance on less-represented classes. This involved resampling individuals from the original dataset until all classes contained an equal number of samples. The oversampling strategy was applied to all classification models except for the bRF, which intrinsically addresses class imbalance. The adoption of such resampling techniques is a widespread practice in ML when models assume balanced class distributions, as in the present study.

Model tuning and assessment were conducted using a stratified 5x5 nested Cross-Validation (CV) procedure (Figure 2 and File S1). Specifically, the described model pipeline was applied within each fold of the CV to prevent information leakage (Kaufman et al., 2012). The inner loop of the CV was used to identify the optimal set of hyperparameters for each ML model, while the outer loop evaluated model performance based on the best hyperparameter configuration found. The hyperparameters for each method, along with their respective ranges of variation, are reported in Table S1.

All methods and models were implemented in Python (version 3.11.9) (Van Rossum & Drake, 2009) using NumPy (Harris et al., 2020), pandas (McKinney, 2010), and scikit-learn (Pedregosa et al., 2011). Visualization was made in R using the package circlize (Z. Gu et al., 2014). A GitHub repository with the complete implementation of the adopted pipeline, model tuning and assessment is publicly available[1].

[1] https://github.com/danielecastellana22/BGA-prediction-with-ML

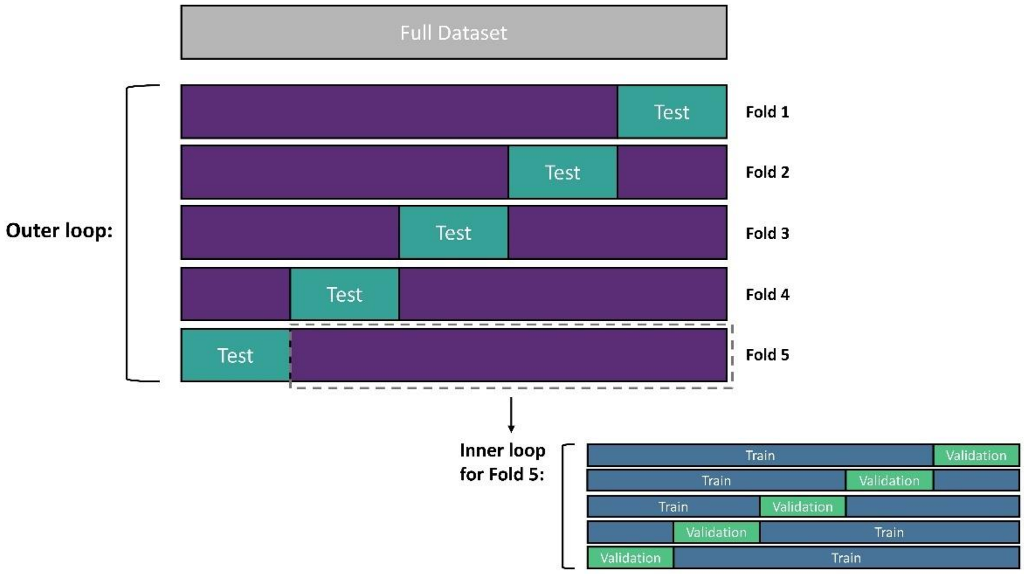

Fig. 2. Schematic representation of a nested cross-validation procedure with five outer folds. Each fold in the outer loop consists of a test set (in teal) and a training set (in purple). For each outer training set, a five-fold inner cross-validation is performed. The inner loop partitions the outer training data into training and validation sets, shown in blue and green, respectively (as illustrated for Fold 5).

2.3 Evaluation Methods and Statistical Tests

Model performance was assessed using multi-class classification metrics (Grandini et al., 2020; Opitz & Burst, 2019). Given the strong class imbalance, evaluating classifiers can be challenging (Fernández et al., 2018; He & Garcia, 2009; Khan et al., 2024). In addition to accuracy —reported for comparison with previous studies— balanced accuracy was used as a more informative metric, as it gives equal weight to all classes regardless of their frequencies. This makes it particularly suitable for imbalanced datasets. Table 2 provides a detailed description of the adopted multi-class metrics, where N is the total number of predictions, K the number of classes, TP_k the number of correctly classified individuals in class k_{th}, FP_k the number of individuals incorrectly assigned to class k_{th}, and FN_k the number of individuals from class k_{th} misclassified into another class.

To assess computational efficiency, the elapsed time was also recorded, measured in seconds from start (t_0) to end (t_1) of execution. All the ML algorithms were executed on an Intel® Xeon® Platinum 8260 CPU@2.40GHz to ensure consistency of timing.

To test the null hypothesis that there is no statistically significant difference in performance among ML models, a non-parametric Friedman test (Friedman, 1940), based on their ranking among the five outer test folds of the nested CV, was conducted with a significance level α = 0.05. A post hoc Nemenyi test (Nemenyi, 1963) was applied to identify which pairs of models differed significantly. This test performs all pairwise (one-vs-one) comparisons between models, with a significance level α = 0.05.

To check the stability of per-class balanced accuracy within each outer CV fold, the Wilson confidence interval was adopted. It provides better coverage, especially for the classes with the smallest sample sizes (Orawo, 2021).

Table 2 Adopted performance metrics.

Metric	Formula	Interpretation	Range
Accuracy	$\frac{1}{N}\sum_{k}^{K} TP_k$	Proportion of correct predictions over the total number of predictions.	$[0,1]$
Balanced Accuracy (Macro-Recall)	$\frac{1}{K}\sum_{k}^{K}\frac{TP_k}{TP_k + FN_k}$ $= \frac{1}{K}\sum_{k}^{K} Recall_k$	Class average of the proportion of correct predictions among the numerosity of each class. For a balanced dataset, it approximates the accuracy.	$[0,1]$
Macro-Precision	$\frac{1}{K}\sum_{k}^{K}\frac{TP_k}{TP_k + FP_k}$ $= \frac{1}{K}\sum_{k}^{K} Precision_k$	Class average of the proportion of correct predictions among all predictions made for each class.	$[0,1]$
Macro-F1 score	$\frac{1}{K}\sum_{k}^{K} 2\cdot\frac{Precision_k \cdot Recall_k}{Precision_k + Recall_k}$ $= \frac{1}{K}\sum_{k}^{K} F1_k$	Class average of harmonic means between Precision and Recall computed for each class.	$[0,1]$
Elapsed Time	$t_1 - t_0$	Time in seconds needed to run the method with the best configuration of hyperparameters.	$[0,+\infty[$

3. Results and Discussion

3.1 Analysis of Missing Values and Dataset Refinement

Prior to CV implementation, a preliminary analysis of missing data was conducted. The dataset comprised 3,591 individuals and 3,234 SNP markers, with a total of 73,256 missing values, corresponding to 0.63% of the dataset. Most individuals exhibited minimal missingness: 95.66% had ≤ 2% missing data, and five individuals had complete genotypes. Notably, 70.37% of individuals fell within the (0.0%-0.5%] missing data percentage range.

Among those with more than 2% missing data, 4.18% had between 2.0% and 2.5%, and five individuals (0.14% of the total) had between 2.5% and 3.0%. Only one individual had the highest proportion of missing data, equal to 3.12%.

One individual from the “Horn of Africa” class was excluded due to being the sole representative of that group.

Regarding the distribution of the missing values per marker, 97.77% of SNPs had ≤ 3.5% missing data, and 290 markers (8.97% of the total) were fully observed. Concerning the few markers with a percentage of missing data greater than 3.5%, 31 markers had a percentage of missing data within (3.5%-10%] range, and 41 genetic markers had more than 10% missing values, representing just 1.27% of the total.

To maintain the informativeness of the panel and minimize the impact of imputation techniques on performance, a total of 72 SNPs with > 3.5% missing values were removed, resulting in a final dataset of 3,590 individuals and 3,162 SNPs, with only 0,14% missing data remaining (16,140 values).

3.2 Ancestry Inference at Inter-Continental Level

The classification performance of the ML models at the inter-continental level is summarized in Figure 3, based on the test sets of the outer folds of the nested CV. Mean metrics and their standard deviations were computed using the best hyperparameter configuration selected in the inner loop based on mean balanced accuracy. The corresponding mean performance metrics on the training and validation sets of the inner folds are reported in Tables S2 and S3, respectively.

All ML methods, except RF, achieved a mean balanced accuracy ≥0.79. SVMs showed the highest performance (0.84±0.02), followed by LR, bRF, NB, and GB, with comparable results (0.82±0.03, 0.82±0.04, 0.81±0.02, and 0.79±0.02, respectively). RF showed the lowest performance (0.65±0.02), likely due to the inefficacy of the external oversampling strategy in correcting class imbalance, as previously noted by Chen & Breiman (Chen & Breiman, 2004). This supports the use of bRF as a more suitable alternative for imbalanced datasets. The Friedman test confirmed that the ML models were statistically significantly different in terms of mean balanced accuracy (p-value <0.005). Further comparisons using the post hoc Nemenyi test revealed that both bRF and SVMs performed significantly better than RF (p-values <0.05 and <0.005, respectively). This indicates that RF with simple oversampling may not be as effective as bRF or SVMs for this specific task. In terms of mean accuracy, all models achieved results ≥0.90, in line with previous studies (Alladio et al., 2022; Pilli et al., 2023; Toma et al., 2018). However, this metric alone may reflect performance on underrepresented classes. For instance, although RF and bRF had similar mean accuracy, their mean balanced accuracy differed significantly, highlighting the importance of using metrics sensitive to class imbalance. Regarding mean macro-precision and mean macro-F1 score, SVMs and GB appear to slightly outperform the other approaches, suggesting better control over false positives and more balanced performance across classes. In contrast, RF showed the lowest mean macro-precision and mean macro-F1 score (<0.72 and <0.70, respectively), suggesting reduced effectiveness in accurately predicting minority classes despite its high mean accuracy.

In terms of computational efficiency, NB was the fastest algorithm (0.81±0.01 sec.), as expected due to its simplicity, followed by SVMs (7.18±1.52 sec.). LR had the longest mean elapsed time (2,053.53±3,065.46 sec.), likely due to slow convergence in certain outer folds, particularly associated with weaker regularization on the parameters' norm, which can lead to ill-conditioned problems (see Table S4). Among ensemble methods, bRF (13.78±5.84 sec.) was significantly faster than RF (106.05±2.52 sec.), further supporting its practical advantage in imbalanced classification tasks. Instead, GB exhibited a moderate mean elapsed time compared to them (73.44±59.21 sec.).

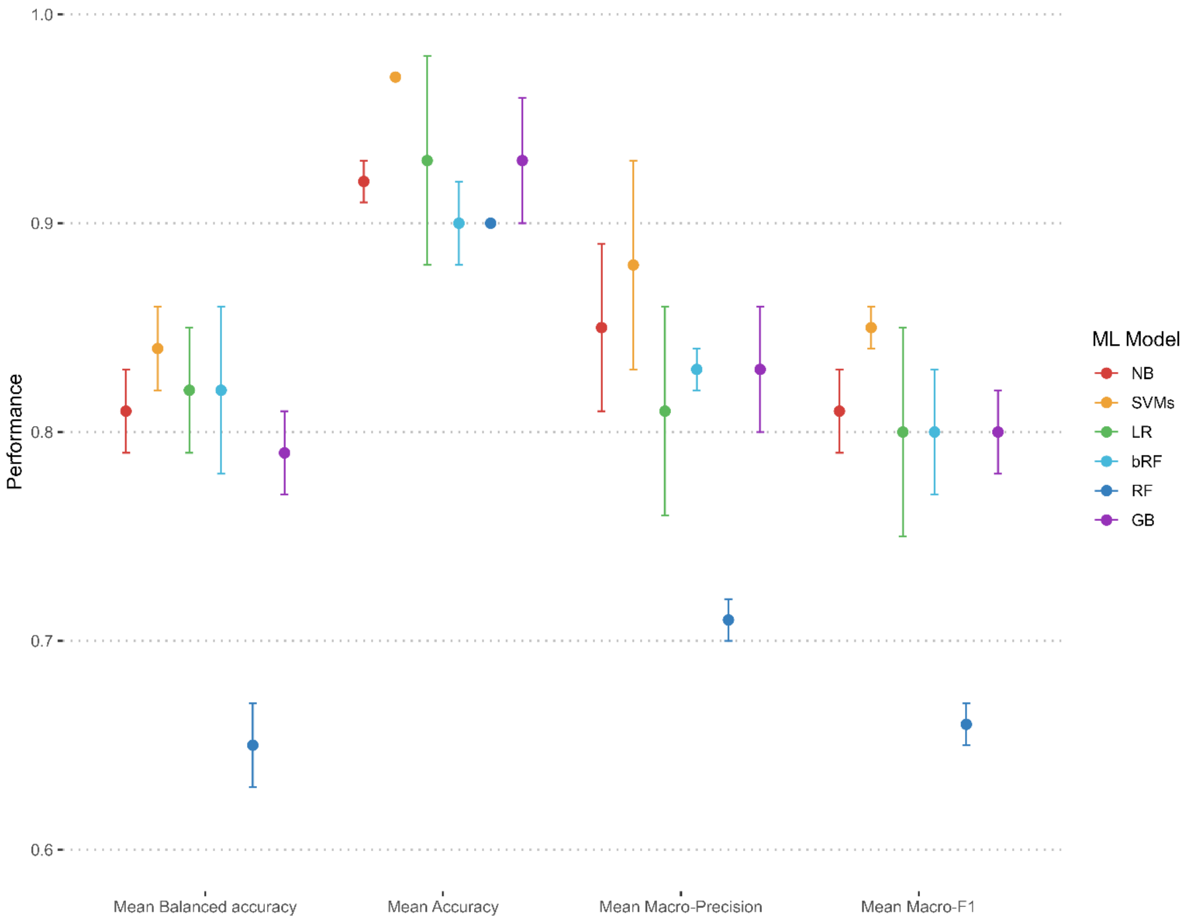

Fig. 3. Comparison of the average performance of six ML models at inter-continental BGA level evaluated across four metrics: mean balanced accuracy, mean accuracy, mean macro-precision, and mean macro-F1 score. The models include NB, SVMs, LR, bRF, RF, and GB. Each point represents the mean value obtained from the five test sets of the outer cross-validation, with error bars indicating the standard deviation.

The mean balanced accuracy achieved by the six ML models across nine inter-continental labels is presented in Figure 4. The most represented classes —Africa, South Asia, East Asia, and Europe— were classified with high scores and low variability, particularly by SVMs. In contrast, Central Asia was the most challenging class with models performing poorly (≤0.44).

For Oceania, all models performed similarly well (≥0.86). Among base learners, SVMs delivered the best performance across most classes, followed by LR, which shows slightly more fluctuation. NB showed lower results overall, especially for America (0.66±0.04). Among ensemble methods, bRF achieved generally good performance but struggled with America and North Asia (0.72±0.04 and 0.73±0.17, respectively). RF was highly unstable: it performed well in large classes (East Asia and Europe), but failed completely in Middle East and North Asia, highlighting its sensitivity to class imbalance despite the simple oversampling. GB showed high performance but encountered problems in classifying North Asia (0.62±0.13).

The stability analysis of the per-class balanced accuracy of the methods across test folds of the outer loop is reported in Supplementary Figure S1. Classes with larger sample sizes and greater genetic homogeneity —such as Europe, Africa, and South Asia— exhibited narrower Wilson confidence intervals, indicating more reliable and consistent performance. In contrast, underrepresented classes like North Asia, Oceania, and Central Asia showed wider intervals, reflecting increased uncertainty and variability in classification outcomes. These findings highlighted the influence of sample size and population structure on model robustness and suggested caution when interpreting results for less-represented groups.

In summary, linear SVMs demonstrated the most consistent and accurate classification across both well- and underrepresented classes, highlighting their suitability for ancestry inference in forensic and population genetics applications. Overall, ML models obtained high performances but encountered some problems classifying individuals for more admixed or underrepresented groups. The results highlight the importance of carefully controlling the impact of oversampling strategies on model performance.

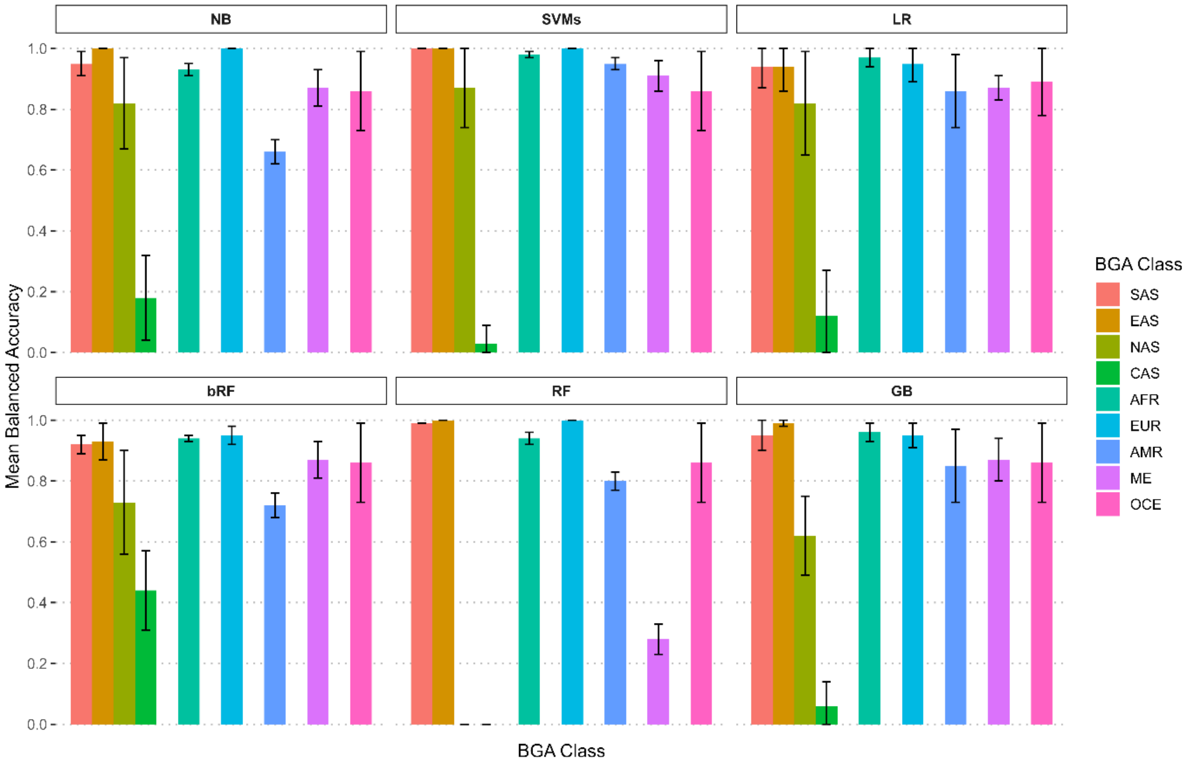

Fig. 4. Class-wise performance of six ML models in predicting inter-continental BGA classes, measured using mean balanced accuracy. The BGA classes include South Asia (SAS), East Asia (EAS), North Asia (NAS), Central Asia (CAS), Africa (AFR), Europe (EUR), America (AMR), Middle East (ME), and Oceania (OCE). Each panel corresponds to a different model: NB, SVMs, LR, bRF, RF, and GB. Bars represent the mean balanced accuracy across the five test sets of the outer cross-validation, with error bars indicating the standard deviation. Bars are visually grouped by continent where applicable, and are ordered in decreasing sample size both within and across continents/regions.

3.3 Ancestry Inference at Detailed-Continental Level

The test results from the outer CV loop at the detailed-continental classification level are presented in Figure 5. Mean metrics and their standard deviations were computed using the best hyperparameter configuration selected in the inner loop based on mean balanced accuracy. The mean performance metrics for the training and validation sets are reported in Tables S5 and S6, respectively.

As expected, classification at this level proved more challenging due to the increased number of BGA classes, reduced sample sizes per class, and lower genetic representativeness of some groups. Additionally, complexity arose from the genetic similarity between geographically close populations, historical admixture, and label ambiguity. Despite these challenges, most models achieved satisfactory results. In terms of mean balanced accuracy, LR achieved the highest performance (0.77±0.02), followed by SMVs (0.75±0.01) and NB (0.75±0.02), suggesting that base learners are well-suited for detailed-scale ancestry inference. In contrast, RF showed the lowest result (0.45±0.06), likely due to its limitation in handling imbalanced data. The performance gap between RF and bRF was particularly notable, suggesting the effectiveness of internal balancing mechanisms (Chen & Breiman, 2004). The Friedman test revealed a statistically significant difference in mean balanced accuracy across the ML models (p-value <0.001), confirming that not all models performed equally well. Post hoc Nemenyi tests revealed that the RF performed statistically significantly worse than three base learners: LR, NB, and SVMs (p-values <0.005, <0.05, and <0.05, respectively). Additionally, LR outperformed GB with statistical significance (p-value <0.05). These results highlighted that, when using a simple oversampling strategy, simpler models may deliver more reliable performance than ensemble methods. In terms of mean accuracy, all models outperformed RF (0.76±0.03) with SVMs reaching the highest value (0.90±0.01), consistent with previous findings (Toma et al., 2018). Regarding mean macro-precision and mean macro-F1 score, SVMs and LR again outperformed other models, indicating better class-wise performance. NB also showed competitive results, while RF exhibited the lowest and most variable performance across all metrics.

In terms of computational efficiency, NB was the fastest model (1.66±0.10 sec.), followed by SVMs (14.88±0.43 sec.) and bRF (42.27±15.45 sec.). LR and GB showed intermediate mean elapsed time (51.37±2.15 sec. and 81.28±6.25 sec., respectively), while RF was the slowest (165.76±36.88 sec.), further limiting its applicability in this context.

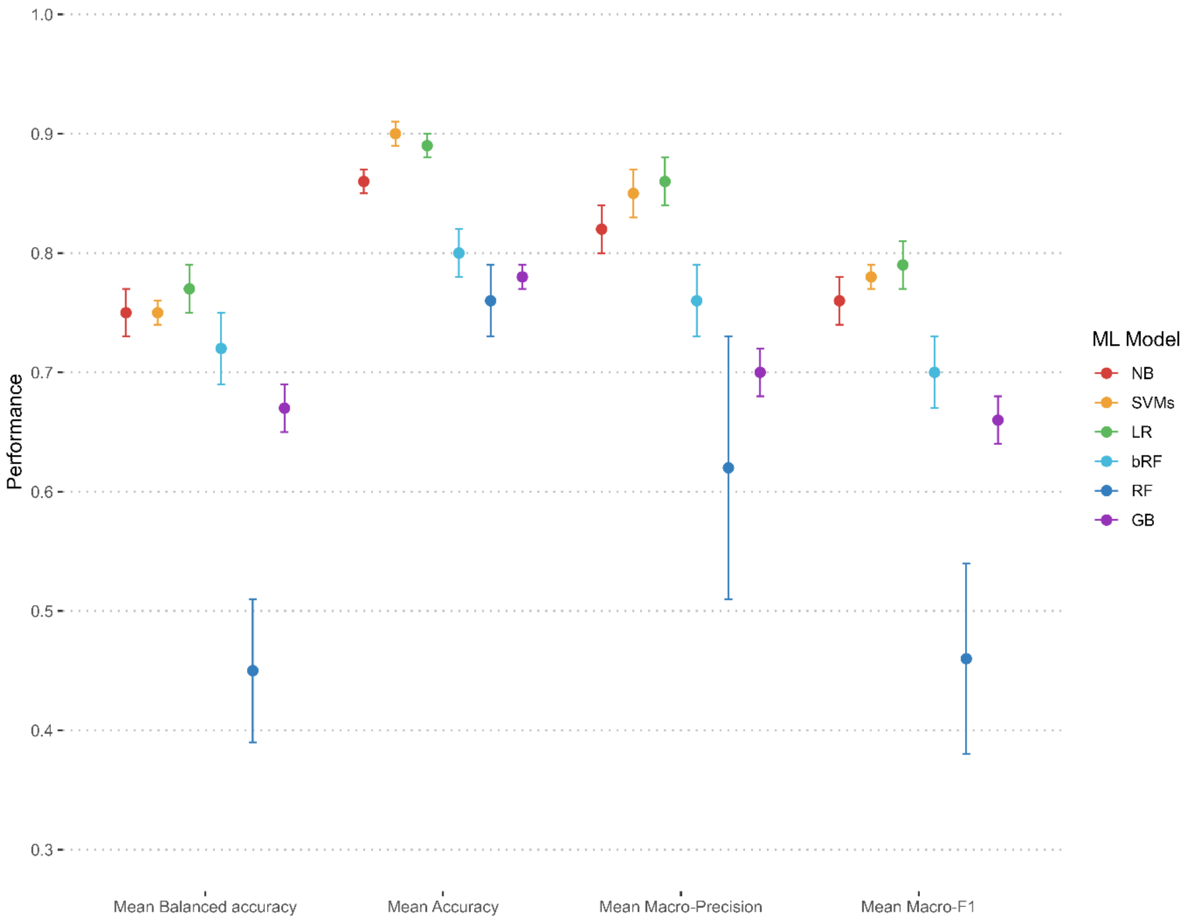

Fig. 5. Comparison of the average performance of six ML models at detailed-continental BGA level evaluated across four metrics: mean balanced accuracy, mean accuracy, mean macro-precision, and mean macro-F1 score. The models include NB, SVMs, LR, bRF, RF, and GB. Each point represents the mean value obtained from the five test sets of the outer cross-validation, with error bars indicating the standard deviation.

The performance of the models in terms of mean balanced accuracy across the seventeen detailed-continental BGA labels is shown in Figure 6. Classification results varied substantially across population groups and algorithms, reflecting differences in sample size, genetic distinctiveness, and population complexity. Excellent classification performance was observed for Europeans, West Africans, Indian Subcontinent, and African hunter-gatherers, with several models —particularly SVMs and LR— achieving near-perfect mean balanced accuracy. These results highlighted the models' ability to distinguish well-represented and genetically distinct populations. Notably, Oceanians and African hunter-gatherers —underrepresented classes— were classified with high mean balanced accuracy across models. The high performance in Oceanians (≥0.94) may be explained by the genetic homogeneity shaped by strong founder effects and island-specific genetic drift in allele frequencies (Choin et al., 2021; Kimura et al., 2008). Similarly, the performance in African hunter-gatherers likely reflects their distinct genetic profiles, shaped by long-term isolation and deep ancestral lineages. These populations often retain high levels of genetic diversity and unique allele frequencies, making them more distinguishable from other groups in genomic analyses (Schlebusch et al., 2012; Skoglund et al., 2017). In contrast, Caucasians, South Africans, and Inner Asians were generally poorly classified (admixed and underrepresented populations). Among base learners, SVMs and LR performed well overall. However, SVMs struggled with Inner Asians (0.40±0.08) and Southeast Asians (0.47±0.14). NB showed comparable performance, excelling in Central South Asians and Inner Asians (0.93±0.03 and 0.77±0.14, respectively). Among ensemble methods, bRF showed moderate overall results with some problems in classifying East Africans (0.43±0.14) and Central South Asians (0.46±0.07), but stood out in underrepresented classes, achieving high performance in North Africans (0.90±0.11) and Caucasians (0.84±0.15). GB yielded moderate results across several classes but lacked consistent strength, while RF was the least consistent: well in large, well-defined classes, and worst in most underrepresented ones.

The stability analysis of the per-class balanced accuracy of the methods for each test fold of the outer loop is reported in Supplementary Figure S2. All methods showed similar confidence interval patterns, with minor differences reflecting their inherent variability. Underrepresented classes —such as North Africans, African hunter-gatherers, Inner Asians, Oceanians, South Africans, and Caucasians— exhibited wider Wilson confidence intervals, indicating greater uncertainty and potential limitations in generalizing the results. In contrast, more frequent classes displayed narrower intervals, highlighting the stability and reliability of their predictions.

In summary, penalized multinomial LR, linear SVMs, and categorical NB demonstrated strong and stable performance across most classes. bRF showed particular promise in handling underrepresented or challenging groups, such as Caucasians, where it outperformed all other models. RF was the least effective model, often failing in minority classes, likely due to its sensitivity to class imbalance. These results underscore the challenges of inferring ancestry into regions characterized by limited sample sizes, low genetic representativeness, and complex demographic histories shaped by migration and admixture —factors that likely contribute to the variability in model performance observed both within and across continents. Additionally, the geographical and cultural ambiguity of certain regions may further complicate classification.

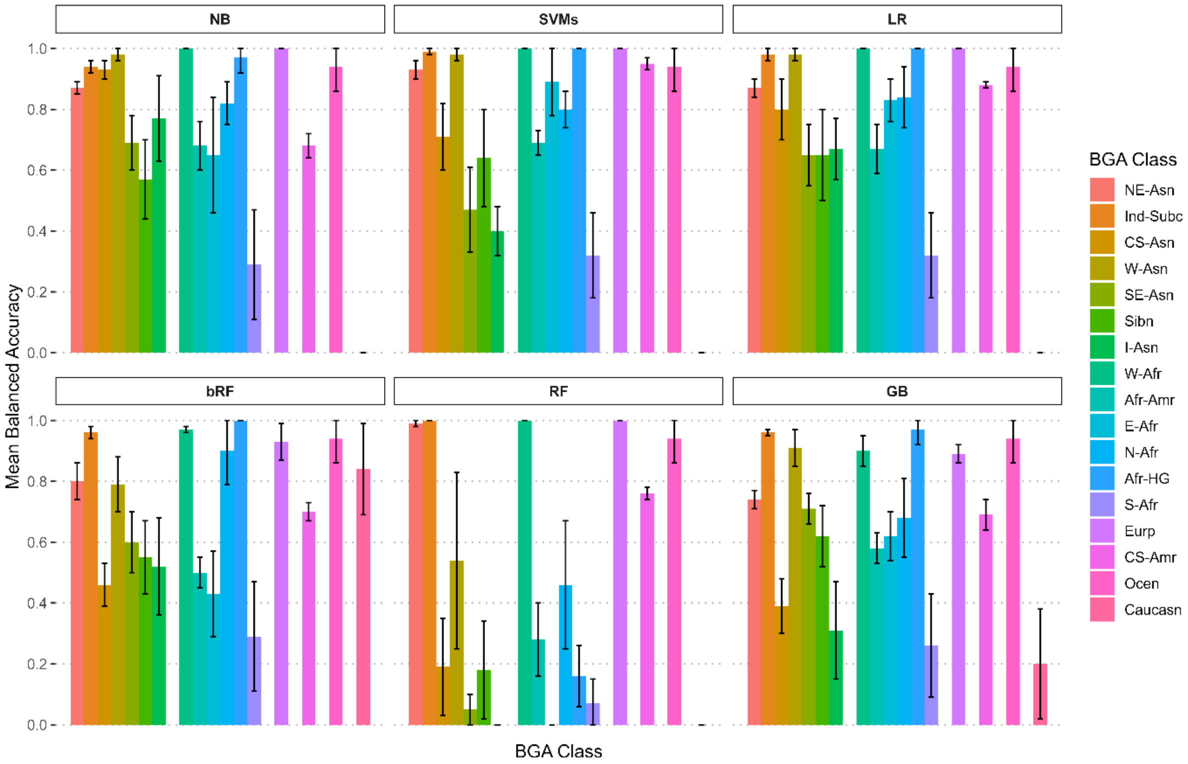

Fig. 6. Class-wise performance of six ML models in predicting detailed-continental BGA classes, measured using mean balanced accuracy. BGA classes include Northeast Asians (NE-Asn), South Asians/Indian Subcontinent (Ind-Subc), Central South Asians (CS-Asn), West Asians (W-Asn), Southeast Asians (SE-Asn), Siberians (Sibn), Central Asians/Inner Asians (I-Asn), West Africans (W-Afr), African Americans (Afr-Amr), East Africans (E-Afr), North Africans (N-Afr), African hunter-gatherers (Afr-HG), South Africans (S-Afr), Europeans (Eurp), (Central and South) Americans (CS-Amr), Oceanians (Ocen), and Caucasians (Caucasn). Each panel corresponds to a different model: NB, SVMs, LR, bRF, RF, and GB. Bars represent the mean balanced accuracy across the five test sets of the outer cross-validation, with error bars indicating the standard deviation. Bars are visually grouped by continent where applicable, and are ordered in decreasing sample size both within and across continents/regions.

3.4 Comparison between Inter- and Detailed-Continental Levels Ancestry Inference

As expected, the mean performance of all ML models decreased at the detailed-continental level compared to the inter-continental BGA level, while the mean elapsed time increased (Figures 3 and 5).

The decline was less pronounced in accuracy than in the other metrics, highlighting its limitations in representing the performance of minority classes. Comparing Figures 4 and 6, differences between ensemble methods and base learners were minimal at the inter-continental level, excluding RF— although base learners slightly outperformed ensemble methods. At the detailed-continental level, these differences became more evident, particularly due to RF’s poor performance on underrepresented classes.

In terms of computational time, classification at the inter-continental BGA level was generally faster, likely due to the lower detailed data and more uniform class distribution. In contrast, the detailed-continental level involved more underrepresented classes, increasing task complexity.

All ML models, except LR, required more time at the detailed-continental level:

- SVMs took approximately twice as long.
- bRF more than tripled its runtime.

Despite these differences, all methods (except LR) completed within a reasonable time frame (under 2 minutes and 46 seconds).

In conclusion, these findings supported the use of simpler models in complex classification tasks such as ancestry inference, at both broad and detailed BGA levels —particularly when combined with appropriate preprocessing and validation strategies. These results are consistent with findings by Kloska et al. (Kloska et al., 2023), who also reported the effectiveness of linear SVMs in ancestry classification. At the same time, the findings underscored the importance of addressing class imbalance and missing data, encouraging further research into more robust and adaptive modeling pipelines.

3.5 Insights on BGA Label Prediction Inaccuracies

As shown in Figures 4 and 6, classification performance varied significantly across BGA labels. Inaccurate predictions were examined in detail using the ML models' average confusion matrices (Tables S7 and S8) to uncover valuable insights, characteristics, and potential misclassification patterns both at inter- and detailed-continental levels. Since similar conclusions can be drawn from all ML methods, the average confusion matrix of the balanced random forest at both inter- and detailed-continental levels was visualized using chord diagrams (Figures 3 and 4), which illustrate correct classifications and misclassifications as arcs. Chords connecting an actual BGA class to the corresponding predicted class represent correctly classified individuals. In contrast, chords linking actual classes to different predicted classes indicate misclassifications. Specifically, incoming chords to a predicted class represent false positives (individuals incorrectly assigned to that class), while outgoing chords from an actual label represent false negatives (individuals belonging to that class but predicted as another). The width of each chord is proportional to the number of shared individuals between actual and classified classes. The initial width reflects the proportion relative to the actual class size, while the final width corresponds to the proportion relative to the predicted class size. For improved readability, Figures 3 and 4 are split into two panels: (a) highlights correct classifications, and (b) emphasizes misclassification patterns. Additionally, a colored bar at the base of the outgoing arc indicates the arc's destination.

At the inter-continental level (Figure 7a), thick arcs highlighted the model's overall good performance. However, several misclassification patterns emerged (Figure 7b). The Middle East class exhibited a high number of false positives —as indicated by the diverse colored arcs entering its classified label—, with individuals from Europe, America, Africa, and South Asia often misclassified into this group. Additionally, some individuals from the actual Middle East class tended to be classified into these classes, as indicated by outgoing arcs from the Middle East label —though these are less visible due to the considerable number of individuals in predicted classes. This pattern likely reflects the region's geographic position at the crossroads of continents and its long history of gene flow and migration (Abou Tayoun & Rehm, 2020; Pereira et al., 2019). Furthermore, substantial post-World War II migration into the Middle East may have contributed to increased genetic admixture in the population.

Similar patterns were observed among the four Asian classes (North Asia, East Asia, Central Asia, and South Asia), particularly for Central Asia, which showed both high false negatives and false positives. This can be explained by region's complex genetic landscape, shaped by extensive historical gene flow and its role as a crossroads of human migration across Asia (Abdulla et al., 2009). Misclassifications between America and Europe were also noted, possibly due to historical migration events (Homburger et al., 2015).

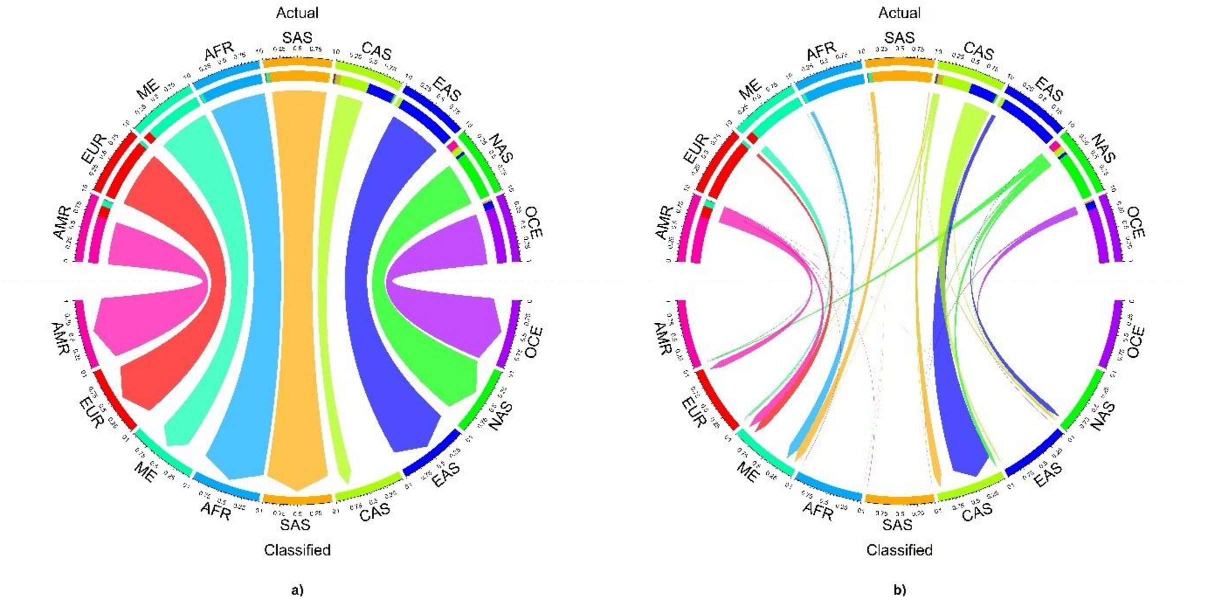

Fig. 7. Chord diagrams illustrate the classification performance of the bRF at inter-continental BGA level, based on the mean confusion matrix from the outer CV loop. The top half of each diagram represents actual BGA classes, while the bottom half shows predicted classes. The BGA classes include America (AMR), Europe (EUR), Middle East (ME), Africa (AFR), South Asia (SAS), Central Asia (CAS), East Asia (EAS), North Asia (NAS), Oceania (OCE). Chord thickness reflects the proportion of samples classified into each group. Chords between closely related populations are displayed in proximity. Correct classifications are shown as chords connecting matching segments, while misclassifications appear as chords linking different predicted and actual BGA classes. Diagram (a) on the left shows correct classifications, diagram (b) on the right shows misclassifications.

At the detailed-continental level, similar to Figure 7a, Figure 8a shows thick arcs between actual and the corresponding classified BGA classes, highlighting the moderate average classification performance of bRF.

Misclassification patterns (Figure 8b) were consistent with those observed at inter-continental level. Errors were most frequent between geographically or genetically close classes, such as among African and Asian subgroups, particularly the Northeast Asians class, which receives the thickest outgoing arc from the actual Siberians class.

Notably, the Caucasians class showed the most diverse set of false positives, including individuals from Europe, Central and South America, West Asia, Central South Asia, and Inner Asia. This reflects the historical and anthropological ambiguity of the term "Caucasian" (Bhopal & Donaldson, 1998; Hall, 2004).

Another notable pattern involved the Central and South Americans class, which showed high heterogeneity due to admixture from European colonization and the transatlantic slave trade (Homburger et al., 2015).

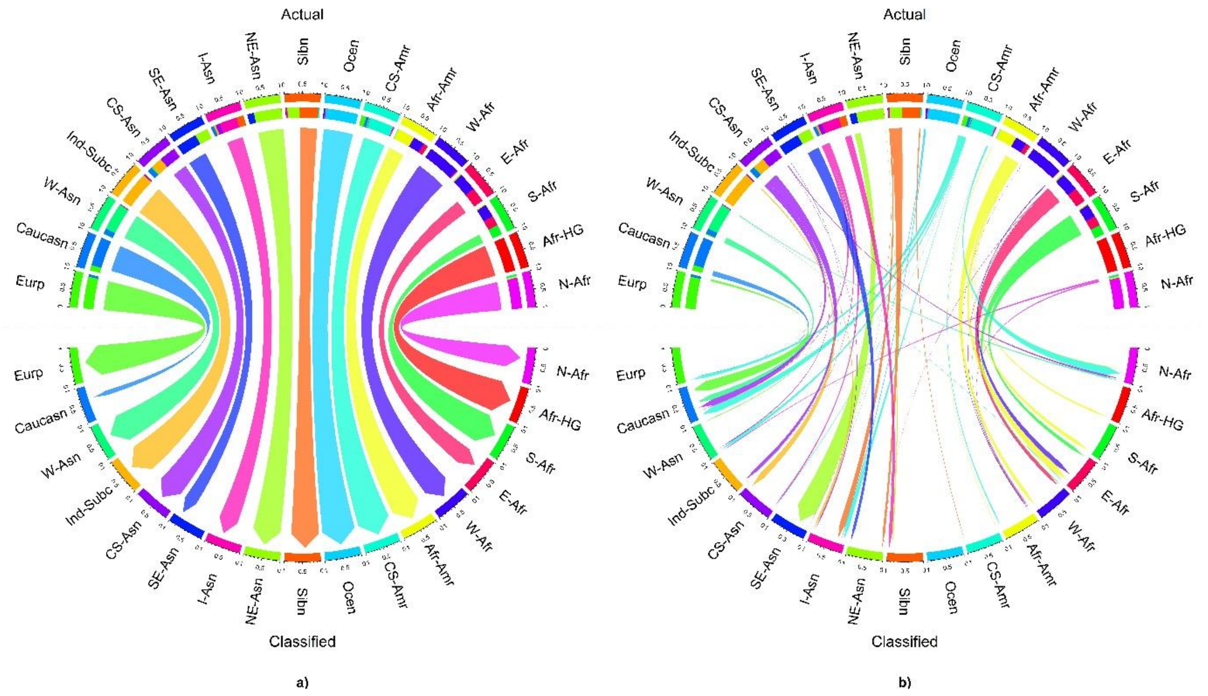

Fig. 8. Chord diagrams illustrate the classification performance of the bRF at detailed-continental BGA level, based on the mean confusion matrix over the outer CV loop. The top half of each diagram represents actual BGA classes, while the bottom half shows predicted classes. BGA classes include Europeans (Eurp), Caucasians (Caucasn), West Asians (W-Asn), South Asians/Indian Subcontinent (Ind-Subc), Central South Asians (CS-Asn), Southeast Asians (SE-Asn), Central Asians/Inner Asians (I-Asn), Northeast Asians (NE-Asn), Siberians (Sibn), Oceanians (Ocen), (Central and South) Americans (CS-Amr), African Americans (Afr-Amr), West Africans (W-Afr), East Africans (E-Afr), South Africans (S-Afr), African hunter-gatherers (Afr-HG), North Africans (N-Afr). Chord thickness reflects the proportion of samples classified into each group. Chords between closely related populations are displayed in proximity. Correct classifications appear as chords connecting matching segments, while misclassifications are shown as chords linking different predicted and actual BGA classes. Diagram (a) on the left shows correct classifications, diagram (b) on the right shows misclassifications.

3.5.1 Pattern Inaccuracies and BGA Uncertainty

The observed misclassification patterns may not solely reflect model limitations but also various sources of uncertainty inherent in ancestry inference. These uncertainties can be broadly categorized into two main domains: those related to BGA labeling and those stemming from the underlying genetic data and population structure.

Label-related uncertainties include:

- Label uncertainty, which arises from limited genealogical information. In this study, BGA labels were assigned based on declared ancestry up to two generations, which may not fully capture an individual's complete genetic background.
- Ambiguity in population definitions, where geographic or cultural groupings used for labeling may not align with actual genetic clusters.
- The assumption of single-class memberships, which simplifies individuals into one BGA category, despite the possibility of admixed ancestry (Royal et al., 2010). This is particularly problematic in forensic contexts, where the unknown trace may correspond to a genetically homogeneous individual.

Genetic and data-related uncertainties include:

- Low genetic representativeness, where certain populations are poorly represented in the reference dataset, limiting the model's ability to generalize across the full spectrum of genetic diversity.
- Sampling bias, which occurs when some populations are overrepresented while others are underrepresented, skewing model training and evaluation.
- Clinal genetic variation, where genetic differences change gradually across geography, making it difficult to define discrete population boundaries.
- Temporal dynamics, as population structures evolve over time, and historical or outdated reference data may not reflect current genetic patterns.
- Technical and methodological limitations, such as differences in genotyping platforms, which may detect different sets of genetic variants; variations in SNP coverage, which affect the resolution and comparability of datasets; and inconsistencies in data preprocessing steps, which can introduce noise and affect classification accuracy.

In conclusion, this analysis underscores the need for improved data quality and more nuanced labeling strategies. Increasing the number of sequenced individuals and refining population labels could help reduce uncertainty and improve model performance (Peterson et al., 2019).

4. Ethical considerations in ancestry prediction

The inference of BGA from genetic data is a scientifically robust and ethically appropriate tool, particularly in forensic contexts, where it can provide crucial investigative leads for the identification of unknown individuals or missing persons. Far from being ethically questionable, this approach can support humanitarian efforts and contribute meaningfully to justice and public safety. However, as with any powerful analytical method, it is important to ensure that its application is guided by principles of scientific rigor, transparency, and fairness. The following considerations aim not to question the legitimacy of BGA inference, but to promote its responsible and informed use:

1. Data Representation and Bias

The performance of ancestry prediction models depends on the quality and representativeness of the reference datasets. Underrepresentation of certain populations — such as Inner Asians, South Africans, or Caucasians in our study— can lead to reduced performance and potential misclassification. This highlights the need to expand and diversify genomic databases, not a flaw in the method itself.

2. Communication of Uncertainty

BGA inference provides probabilistic insights, not definitive assignments. Especially in forensic settings, it is essential that results are interpreted within the appropriate context and communicated with clarity regarding their limitations, particularly in cases involving admixed individuals or populations with complex genetic histories.

3. Preventing Misuse and Misinterpretation

While BGA inference is designed to assist investigations, there is a risk that results could be misused or overinterpreted —e.g., by attributing undue weight to ancestry predictions in legal or social contexts. Clear guidelines and interdisciplinary collaboration between forensic DNA experts and statisticians/computer scientists are essential to prevent misuse and ensure responsible application.

4. Respect for Individual Identity and Diversity

Genetic ancestry does not correspond directly to cultural, national, or personal identity. It is therefore essential to avoid conflating genetic inference with socially constructed categories such as ethnicity, and to acknowledge the complex, multidimensional nature of human identity.

In conclusion, BGA inference is a valuable and ethically appropriate tool when used with transparency, caution, and respect for its limitations. Its value in forensic science can offer meaningful support in investigative scenarios, and ongoing efforts to improve data quality, model interpretability, and population representation will further enhance its reliability and societal benefit.

5. Conclusions

Inferring an individual’s BGA has significant implications across various fields, including medicine, anthropology, archaeology, genetics, immigration, and forensic science. In forensic context, where investigative leads may be limited, BGA inference through DNA analysis can provide valuable insights, contributing to the identification of unknown perpetrators or missing persons. To ensure accurate BGA inference at both inter- and detailed-continental levels, this study evaluated the performance of a novel panel of 3,234 SNPs —originally proposed by Pilli et al. (Pilli et al. 2023)— using a range of supervised ML methods: categorical naive Bayes, penalized multinomial logistic regression, linear support vector machines, random forest, and tree-based gradient boosting. Special attention was given to selecting appropriate algorithms and evaluation metrics, particularly balanced accuracy, to mitigate the effect of class imbalance. A nested cross-validation framework was employed to prevent information leakage and ensure robust model assessment, in line with best practices in genomic data analysis (Whalen et al., 2022). The results demonstrated promising classification performance at the inter-continental level across all ML methods, with the exception of the Central Asia class, which consistently proved challenging. Among the models, SVMs achieved the highest balanced accuracy across most BGA classes, particularly for South

Asia, East Asia, and Europe. Other models, such as NB, LR, and bRF, also performed well, albeit with greater variability.

Although lower compared to inter-continental BGA level results, particularly good values at detailed-continental level were also obtained, except for some classes. High balanced accuracy values were observed for classes such as Europeans, West Africans, Indian Subcontinent, African hunter-gatherers, and Oceanians. However, several classes — particularly Inner Asians, Caucasians, and South Africans— remained difficult to classify accurately, likely due to limited sample sizes and reduced genetic representativeness.

Notably, the Caucasians exhibited consistently low performance across all ML models, except for bRF, which achieved a high balanced accuracy. These findings underscore the potential of classical ML models to perform well in BGA inference, even at a more detailed level of population resolution, when coupled with a well-designed SNP panel.

At the same time, they also highlighted the critical importance of addressing class imbalance and improving the representativeness of reference datasets to ensure reliable and unbiased results. In particular, caution is warranted when interpreting predictions for classes with lower confidence, often due to high genetic variability and limited sample sizes. Increasing the number of samples —particularly for these underrepresented classes— would enhance model generalizability and reduce bias. Furthermore, the study draws attention to the intrinsic complexity of BGA analysis, which is compounded by high-dimensional uncertainty. The difficulty of reconstructing ancestral information from past generations, combined with the possibility of individuals having multiple ancestral origins, challenges the validity of assigning a single BGA label. This complexity is reflected in the observed misclassification patterns, particularly among geographically or historically connected populations.

In conclusion, the proposed SNP panel, combined with supervised ML approaches, represents a valuable tool for forensic applications, particularly in cases where conventional investigative leads are lacking. Future research should explore more advanced imputation methods and refined strategies to address class imbalance. Integrating transfer learning techniques — leveraging data from genetically or geographically related populations— could further enhance model generalization and selection. Additionally, the adoption of explainable AI methods may offer deeper insight into model behavior and decision-making processes. Overall, this approach holds great potential for advancing ancestry inference in both forensic and population genetics contexts.

Authors' contribution

Cosimo Grazzini: Conceptualization, Investigation, Data Curation, Formal Analysis, Methodology, Software, Validation, Visualization, Writing -original draft, Writing -review & editing.

Giorgia Spera: Investigation, Resources, Data Curation, Visualization, Writing -original draft, Writing -review & editing.

Giulia Cereda: Conceptualization, Methodology, Validation, Writing -review & editing, supervision, Funding Acquisition, Project Administration.

Daniele Castellana: Conceptualization, Methodology, Software, Data Curation, Writing - review & editing, Supervision, Funding Acquisition.
Michela Baccini: Validation, Writing -review & editing.
Stefania Morelli: Investigation, Data curation, Resources, Validation of genetic data in collaboration with Giulia Cosenza.
Elena Pilli: Conceptualization, Supervision, Funding Acquisition, Project Administration, Writing -original draft, Writing -review & editing.
All authors read and approved of the final manuscript.

Acknowledgments

This project was supported by the European Union - Next GenerationEU, UNIFI - Competitive biennial projects for Fixed-Term Researchers (RTD) - 2023-2024 - Call. CUP B55F21007810001, through the project entitled "Machine learning ed origine biogeografica: il profilo biologico degli scheletri senza nome".

References

Abdulla, M. A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S. K., Calacal, G. C., Chaurasia, A., Chen, C.-H., Chen, J., Chen, Y.-T., Chu, J., Cutiongco-de la Paz, E. M. C., De Ungria, M. C. A., Delfin, F. C., Edo, J., Fuchareon, S., Ghang, H., Gojobori, T., Han, J., … Zilfalil, B. A. (2009). Mapping human genetic diversity in Asia. *Science (New York, N.Y.)*, *326*(5959), 1541–1545. https://doi.org/10.1126/science.1177074

Abou Tayoun, A. N., & Rehm, H. L. (2020). Genetic variation in the Middle East—an opportunity to advance the human genetics field. *Genome Medicine*, *12*(1), 116. https://doi.org/10.1186/s13073-020-00821-7

Al-Asfi, M., McNevin, D., Mehta, B., Power, D., Gahan, M. E., & Daniel, R. (2018). Assessment of the Precision ID Ancestry panel. *International Journal of Legal Medicine*, *132*(6), 1581–1594. https://doi.org/10.1007/s00414-018-1785-9

Alladio, E., Poggiali, B., Cosenza, G., & Pilli, E. (2022). Multivariate statistical approach and machine learning for the evaluation of biogeographical ancestry inference in the forensic field. *Scientific Reports 2022 12:1*, *12*(1), 1–17. https://doi.org/10.1038/s41598-022-12903-0

Allocco, D. J., Song, Q., Gibbons, G. H., Ramoni, M. F., & Kohane, I. S. (2007). Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms. *BMC Genomics*, *8*, 68. https://doi.org/10.1186/1471-2164-8-68

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/NATURE15393

Barash, M., McNevin, D., Fedorenko, V., & Giverts, P. (2024). Machine learning applications in forensic DNA profiling: A critical review. *Forensic Science International. Genetics*, *69*, 102994. https://doi.org/10.1016/j.fsigen.2023.102994

Battey, C. J., Ralph, P. L., & Kern, A. D. (2020). Predicting geographic location from genetic variation with deep neural networks. *ELife*, *9*. https://doi.org/10.7554/eLife.54507

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J. F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., … Tyler-Smith, C. (2020). *Insights into human genetic variation and population history from 929 diverse genomes*. *367*(6484). https://pubmed.ncbi.nlm.nih.gov/32193295/

Bhopal, R., & Donaldson, L. (1998). White, European, Western, Caucasian, or what? Inappropriate labeling in research on race, ethnicity, and health. *American Journal of Public Health*, *88*(9), 1303–1307. https://doi.org/10.2105/ajph.88.9.1303

Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S. A., & Bustamante, C. D. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(2), 786–791. https://doi.org/10.1073/pnas.0909559107

Bulbul, O., & Filoglu, G. (2018). Development of a SNP panel for predicting biogeographical ancestry and phenotype using massively parallel sequencing. *Electrophoresis*, *39*(21), 2743–2751. https://doi.org/10.1002/ELPS.201800243

Chen, C., & Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data. *University of California, Berkeley*.

Choin, J., Mendoza-Revilla, J., Arauna, L. R., Cuadros-Espinoza, S., Cassar, O., Larena, M., Ko, A. M.-S., Harmant, C., Laurent, R., Verdu, P., Laval, G., Boland, A., Olaso, R., Deleuze, J.-F., Valentin, F., Ko, Y.-C., Jakobsson, M., Gessain, A., Excoffier, L., … Quintana-Murci, L. (2021). Genomic insights into population history and biological adaptation in Oceania. *Nature*, *592*(7855), 583–589. https://doi.org/10.1038/s41586-021-03236-5

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Eduardoff, M., Gross, T. E., Santos, C., De La Puente, M., Ballard, D., Strobl, C., Børsting, C., Morling, N., Fusco, L., Hussing, C., Egyed, B., Souto, L., Uacyisrael, J., Syndercombe Court, D., Carracedo, Lareu, M. V., Schneider, P. M., Parson, W., Phillips, C., … Phillips, C. (2016). Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM^{TM}. *Forensic*

Science International: Genetics, *23*, 178–189. https://doi.org/10.1016/j.fsigen.2016.04.008

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer.

Friedman, M. (1940). A Comparison of Alternative Tests of Significance for the Problem of \$m\$ Rankings. *Annals of Mathematical Statistics*, *11*, 86–92. https://api.semanticscholar.org/CorpusID:121778036

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for Multi-Class Classification: an Overview. *ArXiv*, *abs/2008.0*. https://api.semanticscholar.org/CorpusID:221112671

Gu, J.-Q., Zhao, H., Guo, X.-Y., Sun, H.-Y., Xu, J.-Y., & Wei, Y.-L. (2022). A high-performance SNP panel developed by machine-learning approaches for characterizing genetic differences of Southern and Northern Han Chinese, Korean, and Japanese individuals. *Electrophoresis*, *43*(11), 1183–1192. https://doi.org/10.1002/elps.202100184

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize implements and enhances circular visualization in R. *Bioinformatics*, *30*(19), 2811–2812. https://doi.org/10.1093/bioinformatics/btu393

Guillot, G., Jónsson, H., Hinge, A., Manchih, N., & Orlando, L. (2016). Accurate continuous geographic assignment from low- to high-density SNP data. *Bioinformatics (Oxford, England)*, *32*(7), 1106–1108. https://doi.org/10.1093/bioinformatics/btv703

Guo, Y. X., Jin, X. Y., Xia, Z. Y., Chen, C., Cui, W., & Zhu, B. F. (2020). A small NGS-SNP panel of ancestry inference designed to distinguish African, European, East, and South Asian populations. *Electrophoresis*, *41*(9), 649–656. https://doi.org/10.1002/ELPS.201900231

Hajiloo, M., Sapkota, Y., Mackey, J. R., Robson, P., Greiner, R., & Damaraju, S. (2013). ETHNOPRED: a novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction. *BMC Bioinformatics*, *14*, 61. https://doi.org/10.1186/1471-2105-14-61

Hall, J. (2004). The unexamined "Caucasian." *Nature Genetics*, *36*(6), 541. https://doi.org/10.1038/ng0604-541

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Henn, B. M., Botigué, L. R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J. K., Fadhlaoui-Zid, K., Zalloua, P. A., Moreno-Estrada, A., Bertranpetit, J., Bustamante, C. D., & Comas, D. (2012). Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genetics*, *8*(1), e1002397. https://doi.org/10.1371/journal.pgen.1002397

Hodgson, J. A., Mulligan, C. J., Al-Meeri, A., & Raaum, R. L. (2014). Early back-to-Africa migration into the Horn of Africa. *PLoS Genetics*, *10*(6), e1004393. https://doi.org/10.1371/journal.pgen.1004393

Homburger, J. R., Moreno-Estrada, A., Gignoux, C. R., Nelson, D., Sanchez, E., Ortiz-Tello, P., Pons-Estel, B. A., Acevedo-Vasquez, E., Miranda, P., Langefeld, C. D., Gravel, S., Alarcón-Riquelme, M. E., & Bustamante, C. D. (2015). Genomic Insights into the Ancestry and Demographic History of South America. *PLoS Genetics*, *11*(12), e1005602. https://doi.org/10.1371/journal.pgen.1005602

Hwa, H.-L., Wu, M.-Y., Lin, C.-P., Hsieh, W. H., Yin, H.-I., Lee, T.-T., & Lee, J. C.-I. (2019). A single nucleotide polymorphism panel for individual identification and ancestry assignment in Caucasians and four East and Southeast Asian populations using a machine learning classifier. *Forensic Science, Medicine, and Pathology*, *15*(1), 67–74. https://doi.org/10.1007/s12024-018-0071-y

Jäger, A. C., Alvarez, M. L., Davis, C. P., Guzmán, E., Han, Y., Way, L., Walichiewicz, P., Silva, D., Pham, N., Caves, G., Bruand, J., Schlesinger, F., Pond, S. J. K., Varlaro, J., Stephens, K. M., & Holt, C. L. (2017). Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. *Forensic Science International: Genetics*, *28*, 52–70. https://doi.org/10.1016/j.fsigen.2017.01.011

Jia, J., Wei, Y. L., Qin, C. J., Hu, L., Wan, L. H., & Li, C. X. (2014). Developing a novel panel of genome-wide ancestry informative markers for bio-geographical ancestry estimates. *Forensic Science International. Genetics*, *8*(1), 187–194. https://doi.org/10.1016/J.FSIGEN.2013.09.004

Jiang, L., Wei, Y. L., Zhao, L., Li, N., Liu, T., Liu, H. B., Ren, L. J., Li, J. L., Hao, H. F., Li, Q., & Li, C. X. (2018). Global analysis of population stratification using a smart panel of 27 continental ancestry-informative SNPs. *Forensic Science International. Genetics*, *35*, e10–e12. https://doi.org/10.1016/J.FSIGEN.2018.05.006

Jin, X.-Y., Guo, Y.-X., Chen, C., Cui, W., Liu, Y.-F., Tai, Y.-C., & Zhu, B.-F. (2020). Ancestry Prediction Comparisons of Different AISNPs for Five Continental Populations and Population Structure Dissection of the Xinjiang Hui Group via a Self-Developed Panel. *Genes*, *11*(5). https://doi.org/10.3390/genes11050505

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Trans. Knowl. Discov. Data, 6(4). https://doi.org/10.1145/2382577.2382579

Kersbergen, P., van Duijn, K., Kloosterman, A. D., den Dunnen, J. T., Kayser, M., & de Knijff, P. (2009). Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. *BMC Genetics*, *10*(1), 69. https://doi.org/10.1186/1471-2156-10-69

Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, *244*, 122778. https://doi.org/https://doi.org/10.1016/j.eswa.2023.122778

Kimura, R., Ohashi, J., Matsumura, Y., Nakazawa, M., Inaoka, T., Ohtsuka, R., Osawa, M., & Tokunaga, K. (2008). Gene Flow and Natural Selection in Oceanic Human Populations Inferred from Genome-Wide SNP Typing. *Molecular Biology and Evolution*, *25*(8), 1750–1761. https://doi.org/10.1093/molbev/msn128

Kloska, A., Giełczyk, A., Grzybowski, T., Płoski, R., Kloska, S. M., Marciniak, T., Pałczyński, K., Rogalla-Ładniak, U., Malyarchuk, B. A., Derenko, M. V, Kovačević-Grujičić, N., Stevanović, M., Drakulić, D., Davidović, S., Spólnicka, M., Zubańska, M., & Woźniak, M. (2023). A Machine-Learning-Based Approach to Prediction of Biogeographic Ancestry within Europe. *International Journal of Molecular Sciences*, *24*(20). https://doi.org/10.3390/ijms242015095

Lorente-Galdos, B., Lao, O., Serra-Vidal, G., Santpere, G., Kuderna, L. F. K., Arauna, L. R., Fadhlaoui-Zid, K., Pimenoff, V. N., Soodyall, H., Zalloua, P., Marques-Bonet, T., & Comas, D. (2019). Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. *Genome Biology*, *20*(1), 77. https://doi.org/10.1186/s13059-019-1684-5

McKinney, W. (2010). {D}ata {S}tructures for {S}tatistical {C}omputing in {P}ython. In S. van der Walt & J. Millman (Eds.), *{P}roceedings of the 9th {P}ython in {S}cience {C}onference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92–122. https://doi.org/10.1007/s10618-012-0295-5

Mulindwa, J., Noyes, H., Ilboudo, H., Pagani, L., Nyangiri, O., Kimuda, M. P., Ahouty, B., Asina, O. F., Ofon, E., Kamoto, K., Kabore, J. W., Koffi, M., Ngoyi, D. M., Simo, G., Chisi, J., Sidibe, I., Enyaru, J., Simuunza, M., Alibu, P., … Matovu, E. (2020). High Levels of Genetic Diversity within Nilo-Saharan Populations: Implications for Human Adaptation. *American Journal of Human Genetics*, *107*(3), 473–486. https://doi.org/10.1016/j.ajhg.2020.07.007

Nemenyi, P. B. (1963). *Distribution-free Multiple Comparisons.* Princeton University.

Opitz, J., & Burst, S. (2019). Macro F1 and Macro F1. *ArXiv*, *abs/1911.0*. https://api.semanticscholar.org/CorpusID:207847781

Orawo, L. A. (2021). Confidence Intervals for the Binomial Proportion: A Comparison of Four Methods. *Open Journal of Statistics*, *11*(5). https://doi.org/10.4236/ojs.2021.115047

Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub, Q., Mehdi, S. Q., Thomas, M. G., Luiselli, D., Bekele, E., Bradman, N., Balding, D. J., & Tyler-Smith, C. (2012). Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *American Journal of Human Genetics*, *91*(1), 83–96. https://doi.org/10.1016/j.ajhg.2012.05.015

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Pereira, V., Freire-Aradas, A., Ballard, D., Børsting, C., Diez, V., Pruszkowska-Przybylska, P., Ribeiro, J., Achakzai, N. M., Aliferi, A., Bulbul, O., Carceles, M. D. P., Triki-Fendri, S., Rebai, A., Court, D. S., Morling, N., Lareu, M. V., Carracedo, & Phillips, C. (2019). Development and validation of the EUROFORGEN NAME (North African and Middle Eastern) ancestry panel. *Forensic Science International: Genetics*, *42*, 260–267. https://doi.org/10.1016/J.FSIGEN.2019.06.010

Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C.-Y., Popejoy, A. B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R. J., Brick, L., Carey, C. E., Martin, A. R., Meyers, J. L., Su, J., Chen, J., Edwards, A. C., Kalungi, A., Koen, N., Majara, L., … Duncan, L. E. (2019). Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*, *179*(3), 589–603. https://doi.org/10.1016/j.cell.2019.08.051

Phillips, C., McNevin, D., Kidd, K. K., Lagacé, R., Wootton, S., de la Puente, M., Freire-Aradas, A., Mosquera-Miguel, A., Eduardoff, M., Gross, T., Dagostino, L., Power, D., Olson, S., Hashiyada, M., Oz, C., Parson, W., Schneider, P. M., Lareu, M. V., & Daniel, R. (2019). MAPlex - A massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations. *Forensic Science International: Genetics*, *42*, 213–226. https://doi.org/10.1016/j.fsigen.2019.06.022

Phillips, C., Parson, W., Lundsberg, B., Santos, C., Freire-Aradas, A., Torres, M., Eduardoff, M., Børsting, C., Johansen, P., Fondevila, M., Morling, N., Schneider, P., Carracedo, A., & Lareu, M. V. (2014). Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. *Forensic Science International. Genetics*, *11*, 13–25. https://doi.org/10.1016/j.fsigen.2014.02.012

Pilli, E., Morelli, S., Poggiali, B., & Alladio, E. (2023). Biogeographical ancestry, variable selection, and PLS-DA method: a new panel to assess ancestry in forensic samples via MPS technology. *Forensic Science International: Genetics*, *62*. https://doi.org/10.1016/J.FSIGEN.2022.102806

Qu, Y., Tran, D., & Ma, W. (2019). Deep Learning Approach to Biogeographical Ancestry Inference. *Procedia Computer Science*, *159*, 552–561. https://doi.org/10.1016/J.PROCS.2019.09.210

Resutik, P., Aeschbacher, S., Krützen, M., Kratzer, A., Haas, C., Phillips, C., & Arora, N. (2023). Comparative evaluation of the MAPlex, Precision ID Ancestry Panel, and VISAGE Basic Tool for biogeographical ancestry inference. *Forensic Science International. Genetics*, *64*, 102850. https://doi.org/10.1016/j.fsigen.2023.102850

Rezende, P. M., Xavier, J. S., Ascher, D. B., Fernandes, G. R., & Pires, D. E. V. (2022). Evaluating hierarchical machine learning approaches to classify biological databases. *Briefings in Bioinformatics*, *23*(4), bbac216. https://doi.org/10.1093/bib/bbac216

Rogalla, U., Rychlicka, E., Derenko, M. V., Malyarchuk, B. A., & Grzybowski, T. (2015). Simple and cost-effective 14-loci SNP assay designed for differentiation of European, East Asian and African samples. *Forensic Science International. Genetics*, *14*, 42–49. https://doi.org/10.1016/J.FSIGEN.2014.09.009

Royal, C. D., Novembre, J., Fullerton, S. M., Goldstein, D. B., Long, J. C., Bamshad, M. J., & Clark, A. G. (2010). Inferring Genetic Ancestry: Opportunities, Challenges, and Implications. In *American Journal of Human Genetics* (Vol. 86, Issue 5, pp. 661–673). Cell Press. https://doi.org/10.1016/j.ajhg.2010.03.011

Schlebusch, C. M., Skoglund, P., Sjödin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G. B., Soodyall, H., & Jakobsson, M. (2012). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science (New York, N.Y.)*, *338*(6105), 374–379. https://doi.org/10.1126/science.1227721

Serra-Vidal, G., Lucas-Sanchez, M., Fadhlaoui-Zid, K., Bekada, A., Zalloua, P., & Comas, D. (2019). Heterogeneity in Palaeolithic Population Continuity and Neolithic Expansion in North Africa. *Current Biology : CB*, *29*(22), 3953-3959.e4. https://doi.org/10.1016/j.cub.2019.09.050

Setser, C. H., Planz, J. V., Barber, R. C., Phillips, N. R., Chakraborty, R., & Cross, D. S. (2020). Differentiation of Hispanic biogeographic ancestry with 80 ancestry informative markers. *Scientific Reports*, *10*(1). https://doi.org/10.1038/S41598-020-64245-4

Silla, C. N., & Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, *22*(1), 31–72.

Skoglund, P., Thompson, J. C., Prendergast, M. E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., Heinze, A., Olalde, I., Ferry, M., Harney, E., Michel, M., Stewardson, K., Cerezo-Román, J. I., Chiumia, C., Crowther, A., … Reich, D. (2017). Reconstructing Prehistoric African Population Structure. *Cell*, *171*(1), 59-71.e21. https://doi.org/10.1016/J.CELL.2017.08.049/ATTACHMENT/778640B8-6CB2-4996-85AF-84C35FD3C82A/MMC7.XLSX

Toma, T. T., Obafemi-Ajayi, T., Dawson, J. M., & Adjeroh, D. A. (2018). Random Subspace Projection for Predicting Biogeographical Ancestry. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1719–1725. https://api.semanticscholar.org/CorpusID:59233196

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2022). Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews. Genetics*, *23*(3), 169–181. https://doi.org/10.1038/s41576-021-00434-9

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, *51*(1), 12–18. https://doi.org/10.1038/s41588-018-0295-5